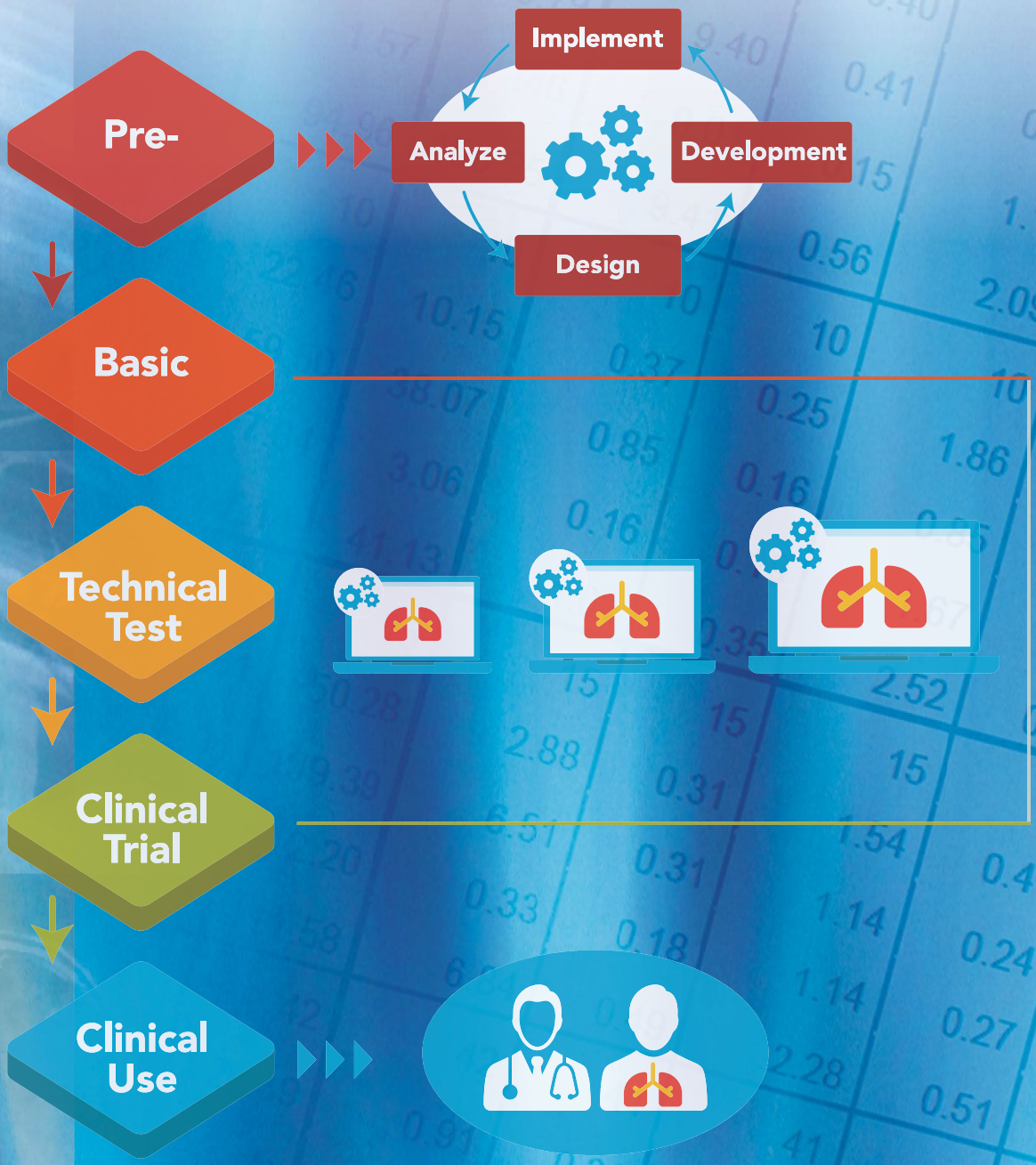
Title	The impact of arterial input function determination variations on prostate dynamic contrast-enhanced magnetic resonance imaging pharmacokinetic modeling: a multicenter data analysis challenge, part II
Authors	Huang, Wei;Chen, Yiyi;Fedorov, Andriy;Li, Xia;Jajamovich, Guido H.;Malyarenko, Dariya I.;Aryal, Madhava P.;LaViolette, Peter S.;Oborski, Matthew J.;O'Sullivan, Finbarr;Abramson, Richard G.;Jafari-Khouzani, Kourosh;Afzal, Aneela;Tudorica, Alina;Moloney, Brendan;Gupta, Sandeep N.;Besa, Cecilia;Kalpathy-Cramer, Jayashree;Mountz, James M.;Laymon, Charles M.;Muzi, Mark;Schmainda, Kathleen;Cao, Yue;Chenevert, Thomas L.;Taouli, Bachir;Yankeelov, Thomas E.;Fennessy, Fiona;Li, Xin
Publication date	2019-03
Original Citation	Huang, W., Chen, Y., Fedorov, A., Li, X., Jajamovich, G. H., Malyarenko, D. I., Aryal, M. P., LaViolette, P. S., Oborski, M. J., O'Sullivan, F., Abramson, R. G., Jafari-Khouzani, K., Afzal, A., Tudorica, A., Moloney, B., Gupta, S. N., Besa, C., Kalpathy-Cramer, J., Mountz, J. M., Laymon, C. M., Muzi, M., Kinahan, P. E., Schmainda, K., Cao, Y., Chenevert, T. L., Taouli, B., Yankeelov, T. E., Fennessy, F. and Li, X. (2019) 'The Impact of Arterial Input Function Determination Variations on Prostate Dynamic Contrast-Enhanced Magnetic Resonance Imaging Pharmacokinetic Modeling: A Multicenter Data Analysis Challenge, Part II', Tomography (Ann Arbor, Mich.), 5(1), pp. 99-109. (10pp.) DOI: 10.18383/j.tom.2018.00027
Type of publication	Article (peer-reviewed)
Link to publisher's version	10.18383/j.tom.2018.00027
Rights	©2019 The Authors. Published by Grapho Publications, LLC This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/). - http://creativecommons.org/licenses/by-nc-nd/4.0/
Download date	2024-09-21 19:12:37

TOMOGRAPHY®

WWW.TOMOGRAPHY.ORG | VOLUME 5 NUMBER 1 | MARCH 2019



The Quantitative Imaging Network Issue

Benchmarking for Tool Development

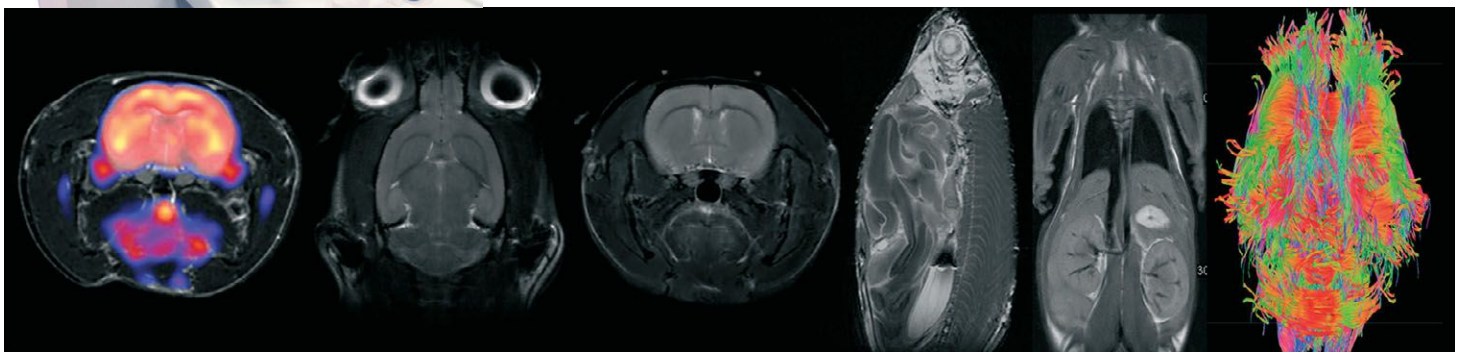
MR SOLUTIONS

Imaging INNOVATION

The Future of Preclinical Imaging

The next generation of superconducting 7T, 4.7T and 3T cryogen free MRI.

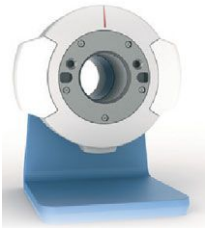
- Up to 12 hour power protection
- No quenching due to power loss
- No cooling water required
- 1500 mT/m gradients



Multi-modality Imaging

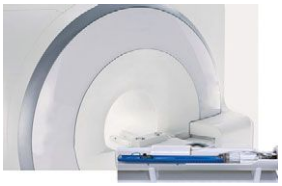
PET/MR
Simultaneous & sequential acquisition of PET & MR
MRS-PET

SPECT/MR
Sequential & simultaneous acquisition of SPECT & MR
MRS-SPECT



Upgrades
Bring your MRI back to life
MRS Upgrade

Clinical Kit
Clinical to preclinical conversion kit
C2P



Confocal
In vivo optical imaging
MRS CellLIVE

For more information contact us at:
information@mrsolutions.com
www.mrsolutions.com

TOMOGRAPHY®
AN INTERNATIONAL JOURNAL FOR IMAGING RESEARCH

Tomography® is published quarterly. Tomography publishes basic (technical and pre-clinical) and clinical scientific articles which involve the advancement of imaging technologies. Tomography encompasses studies that use single or multiple imaging modalities including for example CT, US, PET, SPECT, MR and hyperpolarization technologies, as well as optical modalities (i.e. bioluminescence, photoacoustic, endomicroscopy, fiber optic imaging and optical computed tomography) in basic sciences, engineering, preclinical and clinical medicine. Tomography also welcomes studies involving exploration and refinement of contrast mechanisms and image-derived metrics within and across modalities toward the development of novel imaging probes for image-based feedback and intervention. The use of imaging in biology and medicine provides unparalleled opportunities to noninvasively interrogate tissues to obtain real-time dynamic and quantitative information required for diagnosis and response to interventions and to follow evolving pathological conditions. As multi-modal studies and the complexities of imaging technologies themselves are ever increasing to provide advanced information to scientists and clinicians, Tomography provides a unique publication venue allowing investigators the opportunity to more precisely communicate integrated findings related to the diverse and heterogeneous features associated with underlying anatomical, physiological, functional, metabolic and molecular genetic activities of normal and diseased tissue. Thus Tomography publishes peer-reviewed articles which involve the broad use of imaging of any tissue and disease type including both preclinical and clinical investigations. In addition, hardware/software along with chemical and molecular probe advances are welcome as they are deemed to significantly contribute towards the long-term goal of improving the overall impact of imaging on scientific and clinical discovery. Tomography provides a comprehensive venue for integration of imaging modalities to address biologically important and clinically relevant questions in order to facilitate more rapid and seamless scientific and clinical advancements. Analysis of simultaneous multidimensional (e.g. content, space and time) and multivariate computational data extraction of features which support unique identification of different pathologic tissue phenotypes for advancing imaging is welcomed by Tomography. Types of articles include regular Research Articles, Brief Reports, Reviews, Image Reports and Consensus Papers sponsored from national meetings or government funding agencies.

This journal is available on-line and can be accessed at www.Tomography.org. The journal is an Open Access journal and can be freely downloaded for personal use.

Publication Information: Tomography (ISSN: 2379-1381) is published quarterly by Grapho Publications, 3000 Green Road, #131281, Ann Arbor, MI 48105, USA. Periodicals postage paid at Ann Arbor, MI and additional mailing offices.

USA POSTMASTER: Send address changes to Tomography, Grapho Publications, Customer Service Department, 3000 Green Road, #131281, Ann Arbor, MI 48105, USA.

Advertising Information: Advertising and classified advertising orders and inquiries can be sent to: e-mail: info@tomography.org.

Guide for Authors: For a full and complete Guide for Authors, please go to: http://www.tomography.org.
Author Inquiries: For inquiries relating to the submission of articles (including electronic submission) please visit this journal's homepage at http://www.tomography.org.
Contact details for questions arising after acceptance of an article, including those relating to proofs, will be provided by the publisher.

™ The paper meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).

Reprints: To order reprints for educational, commercial, or promotional use, contact Cadmus Reprints, P.O. Box 822942, Philadelphia, PA 19182-2942; E-mail: cjsreprints@cadmus.com; Phone: 866-487-5625; FAX: 877-705-1373

Funding Body Agreements and Policies: Grapho Publications has established agreements and policies to allow authors whose articles appear in journals published by Grapho Publications, to comply with potential manuscript archiving requirements as specified as conditions of their grant awards. Tomography is an open access journal which complies with NIH funding guidelines and all articles can be archived by authors in PubMed Central http://www.nihms.nih.gov or other international data bases to be in compliance with granting agency requirements.

Orders, Claims, and Journal Inquiries: Please contact the Grapho Publications Customer Service Department, Grapho Publications, 3000 Green Road, #131281, Ann Arbor, MI 48105, USA; phone: (+1) 734-221-0126 e-mail: info@tomography.org.

TOMOGRAPHY.ORG
VOLUME 5 NUMBER 1
MARCH 2019

TOMOGRAPHY®

AN INTERNATIONAL JOURNAL FOR IMAGING RESEARCH

(Contents continue)

COVER ILLUSTRATION

This issue celebrates the 10th year of operation of the Quantitative Imaging Network (QIN) of the National Cancer Institute (NCI). Articles reflect the progress of research teams within the network to develop and validate clinical decision support software tools to measure or predict treatment response of cancers. A variety of tests, including special challenges and tool benchmarking, have been instituted within the network to prepare the quantitative imaging tools for service in clinical trials. This issue provides an overview of active QIN team projects in their development and validation of clinical decision support QI tools in support of clinical trials.

TOMOGRAPHY.ORG
VOLUME 5 NUMBER 1
MARCH 2019

TOMOGRAPHY®
AN INTERNATIONAL JOURNAL FOR IMAGING RESEARCH

(Contents continue)

TOMOGRAPHY.ORG
VOLUME 5 NUMBER 1
MARCH 2019

TOMOGRAPHY®
AN INTERNATIONAL JOURNAL FOR IMAGING RESEARCH

(Contents continue)

TOMOGRAPHY.ORG
VOLUME 5 NUMBER 1
MARCH 2019

TOMOGRAPHY®
AN INTERNATIONAL JOURNAL FOR IMAGING RESEARCH

The Analysis & Collaboration Platform for Data-Intensive Research

Make your data & algorithms accessible, shareable, & reproducible.

INCREASED UTILIZATION

Easy to implement on-premise or in the cloud. Easy to use for all researchers.

FAST LANE TO DISCOVERY

Free up your time to do research, not IT.

FUNDED TO PUBLISHED

Simplify your processes from receiving funding to publishing your reproducibile data.

FLYWHEEL

Request a demo at flywheel.io

@Flywheel_io

The Quantitative Imaging Network: A Decade of Achievement

Thomas E. Yankeelov

Guest Editor, Special QIN Issue of *Tomography*, The University of Texas at Austin, TX

This issue of *Tomography* is a collection of articles derived from over 20 research teams which comprise the Quantitative Imaging Network (QIN) of the National Institutes of Health (NIH).

A primary motivation for establishing the QIN program was the acknowledgement of the lack of validated and reproducible tools appropriate for performing quantitative analysis of medical imaging data to support prediction of clinical tumor responses and outcomes. This consortium, which currently consists of 21 individual research teams, has come together to form a cohesive and productive collaborative effort that supports development, optimization, and validation of quantitative imaging methods and associated software tools. The advances that have been developed as part of these efforts form the infrastructure for which practicing oncologists and radiologists can derive and utilize quantitative imaging metrics in decision support for improvement of individual patient care as well as for clinical trial assessment of novel therapeutics. This issue of *Tomography* celebrates the 10th-year anniversary of QIN advances following its inception in 2008. During this time, the NCI program staff has supported QIN efforts through program announcements, providing opportunities for the network to grow. Furthermore, investigators and governments from the international community have taken interest in this important effort. Today, the QIN includes teams from 11 different countries; in addition to laboratories from the United States, 2 teams were added to the QIN as Full Members through support from the Canadian Government, and teams from 9 other countries joined as Associate Members. Additional funding announcements have emerged to advance research objectives that include PAR-18-248 (a UG3/UH3 mechanism). This announcement is focused on supporting the development and adaptation/implementation of quantitative imaging methods, protocols, and/or software tools based on existing commercial imaging platforms and instrumentation for application in current or planned clinical therapy trials. Moreover, an R01 mechanism, PAR-18-919, is now also available for research teams with a fully developed and optimized clinical decision tools needing clinical validation.

As evidenced from the articles presented in this special issue, translation of quantitative imaging methods and algorithms as clinical decision support tools into clinical utility has been successfully achieved across imaging modalities and instrument manufacturers. Collectively, these contributions exemplify the unified effort provided by the organizational structure of the QIN that consists of an executive committee, technical teams, working groups, and a coordinating committee.

The articles in this issue indicate the diversity and versatility of the membership and highlight their ingenuity and dedication to further improve the use of imaging in patient care. It is an honor to thank the contributors to this special issue for their excellent contributions and for their effective partnership with NIH staff to advance quantitative imaging. The impact of these efforts will be ever more apparent over time—long after the NIH QIN has formally ended funding of this vitally important program. The advances produced by the QIN teams of dedicated clinicians and scientists will directly translate into improved patient care, yielding improved clinical outcomes in the decades ahead.

QIN organizational overview.

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2019.00999

QIN Benchmarks for Clinical Translation of Quantitative Imaging Tools

Keyvan Farahani, Darrell Tata, and Robert J. Nordstrom

Cancer Imaging Program, National Cancer Institute of NIH, Bethesda, MD

Corresponding Author:
Keyvan Farahani, PhD
9609 Medical Center Drive, 4W-312, Bethesda, MD 20854;
E-mail: farahani@nih.gov

Key Words: cancer imaging, quantitative imaging, benchmarks, clinical translation, oncology
Abbreviations: National Cancer Institute (NCI), Quantitative Imaging Network (QIN), challenges and collaborative projects (CCPs), magnetic resonance imaging (MRI), positron emission tomography (PET)

ABSTRACT

The Quantitative Imaging Network of the National Cancer Institute is in its 10th year of operation, and research teams within the network are developing and validating clinical decision support software tools to measure or predict the response of cancers to various therapies. As projects progress from development activities to validation of quantitative imaging tools and methods, it is important to evaluate the performance and clinical readiness of the tools before committing them to prospective clinical trials. A variety of tests, including special challenges and tool benchmarking, have been instituted within the network to prepare the quantitative imaging tools for service in clinical trials. This article highlights the benchmarking process and provides a current evaluation of several tools in their transition from development to validation.

INTRODUCTION

A distinguishing advantage of any research network is the opportunity for the ensemble of member teams to collaborate in areas of shared interest, addressing common scientific or technological problems, to compare individual approaches, and ultimately to build consensus. As a result, the ensemble of teams in a research network is often greater than the sum of its parts. For the past 10 years, the National Cancer Institute (NCI) Quantitative Imaging Network (QIN) has provided a network environment where the development and validation of quantitative imaging (QI) analysis software tools designed to measure or predict response to cancer therapies in clinical trials have been pursued. The motivating hypothesis for the QIN has been that clinical trials in systemic or targeted chemo-, radiation-, or immunotherapies can benefit from the inclusion of QI methods in the treatment protocols. These methods involve the extraction of measurable information from medical images to assess the status or change of a disease.

To date, 36 multidisciplinary teams from academic institutions across the United States and Canada have participated in the NCI-funded research program. The current number of teams supported by the network is 20. These research teams discuss and resolve common challenges such as imaging informatics activities, clinical trial design and validation planning, and data acquisition and analysis issues, to name only a few. At the same time, each team is required to make technical and clinical progress on its individual research project.

The interest in QI as a method to gauge tumor progression or predict response to therapy predates the QIN. An early attempt at extracting numeric information from clinical images came in the form of RECIST (Response Evaluation Criteria in Solid Tumors) in 2000 (1, 2), based on earlier guidelines first published by the World Health Organization in 1981 (3). The RECIST criteria used a single straight line drawn across the widest dimension in a tumor image to provide a quantitative measure of tumor size. Size, suitability, and the number of lesions to be measured were stated in the original guidelines, and later revised in version 1.1 (4). Response criteria, measured by the change in linear dimension, were established to determine if the tumor was in complete response, partial response, or stable or progressive disease.

Although tumor shrinkage is an obvious desirable response to cancer therapy, it is not the only response that can occur, or in some cases, the response may be delayed in appearing (5). Furthermore, in a metastatic cancer setting a limited set of target lesions, as prescribed in RECIST 1.1, may not represent the overall tumor burden or response to therapy (6). These limitations restrict the usefulness of RECIST in some clinical trials. Often, immunotherapy trials, for example, show that complete response or stable response can occur after an initial increase in tumor burden (7, 8). Conventional RECIST criteria early in the therapy run the risk of labeling this initial increase as tumor progression, failing to account for the delayed onset of antitumor T cell response. Thus, a therapy under study in a clinical trial can be seen as failing. This has led to the development of iRECIST guidelines for response criteria in immunotherapy trials (9).

QI tools being developed and validated by QIN research teams measure far more than simple unidimensional tumor size, and the articles in this special issue of *Tomography* highlight a number of them. Physical attributes of tumors such as heterogeneity, diffusion and perfusion, and metabolic activity are

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2018.00045

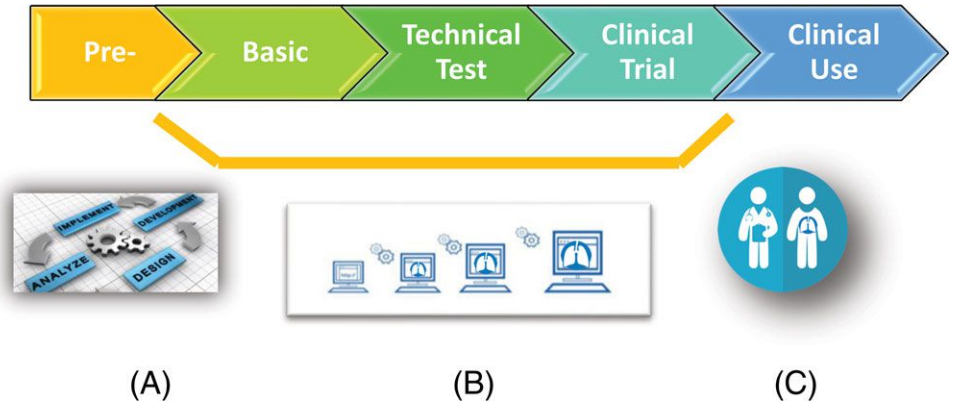


Figure 1. Quantitative Imaging Network (QIN) benchmarks, described in the text and in Figure 2, designate key milestones toward the clinical translation of quantitative imaging (QI) tools from laboratory prototype (A) to scale up and optimization (B) to clinical use (C).

being added to the more traditional size and shape measurements of QI to determine response to therapy. These attributes have been used in machine-based modeling studies driven by imaging data to characterize tumor growth (10-13). In addition, machine learning radiomics approaches for high-throughput extraction and analysis of quantitative image features are providing an even richer set of image parameters. These include intensity, texture, kurtosis, and skewness from which to extract measurement and prediction information on tumor progression (14-16).

Background

If QI is to be useful in clinical trials as a method to measure or predict response to therapy, the methods must be developed on clinically available platforms such that the final validated tools would have value in multicenter clinical trials. To this end, the NCI QIN program was initiated in 2008. The support mechanism chosen for this effort was the cooperative agreement U01 mechanism. Here, successful applicants agree to collaborations and conditions established by NCI program staff. In the case of the QIN, these conditions include participation in a network of teams, joining in monthly teleconference meetings, and collaborating in several working groups.

Applications to the QIN are subject to the NIH peer-review process conducted 3 times each year. As a result, the network teams enter the program at different times and are thus at different stages in their tool development and validation at any given point in time. This creates a need to qualify the degree of development and validation each quantitative tool has attained. Accordingly, a system of benchmarking to assess tool maturity has been implemented.

Clinical Translation

The process of translating ideas and products from laboratory demonstration to clinical utility is the exercise of transferring stated features of the idea or product into realized benefits to the user. For example, the stated feature of improved sensitivity or specificity in an imaging protocol can translate into improved personalized care in the clinic. The tool developer must be aware of the nature of the clinical need for such a tool. Likewise, the clinical user must be realistic regarding the performance characteristics needed in a clinical decision support tool.

To ensure a strong connection between developer and clinical user, each QIN team is required to have a multidisciplinary composition that brings expertise in imaging physics and radiology along with informatics, oncology, statistics, and clinical requirements to the cancer problem being addressed. This gives each team multiple perspectives on the challenges of advancing decision support tools through the development and verification stages and on to the clinical validation stage.

Translation is not a simple move from bench to bedside. It requires a constant check on progress with a compass heading set by clinical need. There must be a set of guiding milestones to point the way through the translation landscape and to measure progress along the way. This, however, can be very difficult in a network of research teams, where each team is focused on a different imaging modality or approach and cancer problem.

A guiding pathway for QIN teams in this translation process continues to be the use of benchmarks for measuring progress toward clinical utility. Even though each team is working on a different application of QI for measurement or prediction of response to cancer therapy, they all share the challenges of bringing tools and methods into clinical utility. The benchmarks offer a ubiquitous pathway for all teams to move toward clinical workflow. As such, the benchmarks measure the tasks on the development side of the translation. There is no doubt that a set of benchmarks could be established for monitoring progress on the clinical side of the translation issue, but that is not a part of the QIN mission.

Figure 1 shows a schematic pathway from initial concept and development of tools and methods for clinical decision support all the way to final clinical use. The demarcations show that the benchmark grades represent milestones in the development toward the clinical use. The details of the benchmarks and the requirements to achieve each are given in the next section.

Benchmarking

For each team, the transition from the activities of tool development to clinical performance validation is a central part of the research, but this does not occur in a sudden step. There is a period where prototype tools are tested against retrospective image data from archives such as The Cancer Imaging Archive (TCIA) (http://www.cancerimagingarchive.net/) or other data sources to objectively assess tool performance. The benchmarking initiative allows investigators the opportunity to adjust their algorithms before committing to a specific prospective clinical trial.

Another initiative embraced by the QIN team members during their period of initial verification of tool performance has been team challenges. Here, several teams with sufficiently developed tools with similar quantitative measurement functions (segmentation, volume metrics, volume transfer constant, K^{trans}, measurements, etc.) use a common data source, divided into training and test data sets, to determine and compare task-specific tool performance related to determining or predicting the therapeutic response. Within the QIN, these activities are referred to as challenges and collaborative projects (CCPs) (17) and have proven very useful in guiding the development of QI tools and analytic methods in preparation for more complete clinical validation studies. CCPs have been conducted at various points along the development pipeline, from basic concept to technical verification and preliminary clinical validation. Descriptions of CCP tasks, project design, and results have been disseminated through several peer-reviewed scientific publications (18-28).

The CCP activities highlighted the need to create a method for gauging the degree of development a tool had attained at any specific timepoint. This would help to evaluate challenge results when tools with widely different levels of development participated. To gauge the level of development for tools in the QIN, a benchmarking process was developed. A Task Force, comprising QIN members, was charged with the task of developing a system to stratify the level of progress made by teams in their efforts to develop QI tools for clinical workflow. In the context of QIN activities, a tool can be a software algorithm, a physical phantom, or a digital reference object used in the production or analysis of QI biomarkers for diagnosis and staging of cancer and for the prediction or measurement of response to therapy.

The Task Force developed QI Benchmarks as standard labels that signify the development, validation, and clinical translation of quantitative tools through a 5-tier benchmark system as shown in Figure 2 (29): pre-benchmark (level 1), basic benchmark (level 2), technical test benchmark (level 3), clinical trial benchmark (level 4), and clinical use benchmark (Level 5). In general, requirements for each benchmark designation require a peer-reviewed publication, where the scientific goals, methods, and results of the QI biomarker development or analysis are described. A benchmark is not automatically conferred on a QIN tool. The developer must make an application which includes the required information for that benchmark and conduct a discussion of the objective performance claim for the benchmark, best practices, and current limitations of the tool. In addition, it is important to note that candidates for each of the benchmarks must have fulfilled the requirements for the prior-level benchmark but not necessarily obtained it. The Coordinating Committee of QIN, consisting of the chairs of each of the Network Working Groups (30) and certain NCI program staff, reviews each benchmark application. If an application for a benchmark is rejected, the applicant will be allowed to address the concerns and resubmit the application.

The establishment of this benchmarking process will help to advance the field of QI in oncology by recognizing QI tools entering QIN (benchmark level 1), encouraging QIN investigators to participate in objective performance evaluation of their tools and methods (benchmark level 2), to streamline validation through dissemination of appropriately developed tools and methods to test sites (benchmark level 3), and to promote participation in oncology clinical trials (benchmark level 4) by providing objective evaluation of tool development to allow more accurate assessment or prediction of cancer therapies and eventual clinical use (benchmark level 5). It is anticipated that this initiative will help in proper placement of advanced tools and methods into prospective clinical trials and will streamline the process of translating such tools into the broader clinical community with adoption by industry.

RESULTS

The current catalog of QIN tools contains 67 clinical decision support tools in various stages of development. Because of the staggered entrance of teams into the network, progress in development is not uniform across the network. This has created the need for benchmarking as a measurable way to evaluate tool development status. Of the tools listed in the catalog, there are ~12 that are to the point of entering the clinical domain and qualifying for benchmark level 4 or 5.

Image segmentation of tumor from surrounding tissue is an important tool function and serves as a first step in determining treatment planning regimens in oncology and many quantitative measurements of tumor status. Several QIN teams are developing segmentation tools for various applications. One such tool developed at Columbia University (New York, NY) performs image segmentation on solid tumors and has been shown in lung, liver, and lymph nodes as a semiautomatic software tool. The segmentation of magnetic resonance imaging (MRI) and/or computed tomography (CT) images across multiple slices yields quantitative information on tumor volume (31-33) and has been used in several clinical trials. This tool can be integrated into diagnostics, radiation-treatment planning, and tumor response assessment on commercial workstations.

Volumetric measurement of breast cancer tumors using dynamic contrast-enhanced MRI has been developed by the QIN team at the University of California at San Francisco (San Francisco, CA). The tool is an image processing and analysis package based on dynamic contrast-enhanced MRI contrast kinetics and has been approved on a commercial platform. It has proven useful in clinical trials performed by several groups in the NCI clinical trials network (34, 35). In addition to the analysis of algorithm performance, the validation of a breast phantom design has been reported (36). Features of the software package include image reconstruction, image registration, segmentation, and viewer/visualization. A commercial version is being used in ~20 I-SPY clinical sites.

Auto-PERCIST (Positron Emission Tomography [PET] Response Criteria in Solid Tumors) is a software package for PET imaging and can provide clinical decision support through image segmentation, viewer/visualization, and response assessment. Similar to RECIST, the PERCIST package focuses on analyzing fludeoxyglucose-PET scans and evaluates if the study was performed properly from a technical standpoint. It establishes the appropriate threshold for the standardized uptake value corrected for lean body mass (SUL) evaluation of the lesion at baseline. Auto-PERCIST has been used to provide clinical assessment of therapy response in multicenter evalua-

Figure 2. Five levels of QI benchmark for labeling of QI products. *In addition to the requirements listed for each level, candidates for benchmarks must have fulfilled the requirements for the prior-level benchmark, but not necessarily obtained that benchmark, to be considered for the current benchmark level.

tions both here in the United States and in Korea, and a release of Auto-PERCIST for European oncology trials is planned. Although not completely developed under the QIN program, many of the features found in Auto-PERCIST were created and validated in the QIN program by teams originally at the Johns Hopkins University (Baltimore, MD) and currently at Washington University (St. Louis, MO). This tool has been used in several multicenter clinical trials, and details of its performance can be found in several publications (37-40).

Clinical support for evaluating tumor response can come in many forms. It be the algorithm, phantom, or digital reference object for direct analysis of images, and it can also be the

workspace in which the software operates. Such is the case for ePAD, a Stanford University (Palo Alto, CA) web-based image viewing and annotation platform to enable deploying QI biomarkers into clinical trial workflow (41). It supports applications such as data collection, data mining, image annotation, image metadata archiving, and response assessment. This publicly available platform predates QIN, but many of the current quantitative functionalities of ePAD have been installed and validated under QIN support.

CONCLUSIONS

The list of benchmarked tools in QIN is growing. Constant updates are being made to the catalog as new QIN teams enter the network and existing teams progress in their development and validation of their QI tools in support of clinical trials (42, 43). This issue of *Tomography* highlights several QI tools and studies in the QIN. As the network moves forward, it has begun to focus on coordinated ways to approach clinical trial groups and interested commercial parties.

ACKNOWLEDGMENTS

The QIN Challenge Task Force: Keyvan Farahani (NCI), Maryellen Giger (University of Chicago), Wei Huang (Oregon Health Sciences University), Jayashree Kalpathy-Cramer (Massachusetts General Hospital), Sandy Napel (Stanford University), Robert J. Nordstrom (NCI), Darrell Tata (NCI), and Richard Wahl (Washington University).

Disclosures: No disclosures to report.

Conflict of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubenstein L, Verweij J, Glabbeke MV, van Oosterom AT, Christian MC, Gwyther SG. New guidelines to evaluate the response to treatment in solid tumors. J Natl Cancer Inst. 2000;92:205–216.
2. Padhani AR, Ollivier L. The RECIST (Response Evaluation Criteria in Solid Tumors) criteria: implications for diagnostic radiologists. Br J Radiol. 2001;74:983–986.
3. Park JO, Lee SI, Song SY, Kim K, Kim WS, Jung CW, Park YS, Im YH, Kang WK, Lee MH, Lee KS, Park K. Measuring response in solid tumors: comparison of RECIST and WHO response criteria. Jpn J Clin Oncol. 2003 Oct;33:533–537.
4. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Ford R, Dancey J, Arbuck S, Gwyther S, Mooney M, Rubenstein L, Shankar L, Dodd L, Kaplan R, Lacombe D, Verweij J. New response evaluation criteria in solid tumors: revised RECIST guideline (version 1.1). Eur J Cancer. 2009;45:228–247.
5. Huang B. Some statistical considerations in the clinical development of cancer immunotherapies. Pharm Stat. 2018;17:49–60.
6. Khul CK, Alparslan Y, Schmoee J, Sequeria B, Keulers A, Brummendorf TH, Keil S. Validity of RECIST version 1.1 for response assessment in metastatic cancer: a prospective multireader study. Radiology. 2018. [Epub ahead of print]
7. Evangelista L, de Jong M, Del Vecchio S, Cai W. The new era of cancer immunotherapy: what can molecular imaging do to help? Clin Transl Imaging. 2017;5:299–301.
8. Carter BW, Bhosale PR, Yang WT. Immunotherapy and the role of imaging, Cancer. 2018;124:2906–2922.
9. Seymur L, Bogaerts J, Perrone A, Ford R, Schwartz LH, Mandrekar S, Lin NU, Litiere S, Dancey J, Chen A, Hodi FS, Therasse P, Hoekstra OS, Shankar LK, Wolchok JD, Ballinger M, Caramella C, de Vries EG; RECIST Work Group. iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics. Lancet Oncol. 2017;18:e143–e152.
10. Hogea C, Davatzikos C, Biros G. An image-driven parameter estimation problem for a reaction-diffusion glioma growth model with mass effects. J Math Biol. 2008;56:793–825.
11. Yankeelov TE, Atuegwu N, Hormuth D, Weis JA, Barnes Sl, Miga MI, Rericha EC, Quanranta V. Clinically relevant modeling of tumor growth and treatment response. Sci Trans Med. 2013;5:51–55.
12. Weis JA, Miga MI, Arlinghaus LR, Li X, Abramson V, Chakravarthy AB, Pendyala P, Yankeelov TE. Predicting the response of breast cancer to neoadjuvant therapy using a mechanically coupled reaction-diffusion model. Cancer Res. 2015;75:4697–707.
13. Roque T, Risser L, Kersemans V, Smart S, Allen D, Kinchesh P, Gilchrist S, Gomes AL, Schnabel JA, Chappell MA. A DECR-MRI driven 3-D reaction-diffusion model of solid tumor growth. IEEE Trans Med Imaging. 2018;37:724–732.
14. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout G, Granton P, Zegers CM, Gillies R, Boellard R, Dekker A, Aerts HJ. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48:441–446.
15. Aerts HJWL, Rios Velazques E, Leijennar RTH, Parmar C, Grossmann P, Cavalho S, Bussink J, Monshouwer R, Haibe-Kains Rietveld D, Hoeders F, Rietbergen MM, Leemans R, Dekker A, Quackenbush J, Gilles RJ, Lambin P. Decoding tumor phenotype by noninvasive imaging using quantitative radiomics approach. Nat Commun. 2014;5:4006.
16. Zhang B, Tian J, Dong D, Gu D, Dong Y, Zhang L, Lian Z, Liu J, Luo X, Pei S, Mo Z, Huang W, Ouyang F, Guo B, Liang L, Chen W, Liang C, Zhang S. Radiomics features of multiparametric MRI as novel prognostic factors in advanced nasopharyngeal carcinoma. Clin Cancer Res. 2017;23:4259–4269.
17. Farahani K, Kalpathy-Cramer J, Chenevert TL, Rubin DL, Sunderland JJ, Nordstrom RJ, Buatti J, Hylton N. Computational challenges and collaborative projects in the NCI quantitative imaging network. Tomography. 2016;2:242–249.
18. Huang W, Li X, Chen Y, Li X, Chang MC, Oborski MJ, Malyarenko DI, Muzi M, Jajamovich GH, Fedorov A, Tudorice A, Gupta SN, Laymon CM, Marro KI, Dyvorne HA, Miller JV, Barboriak DP, Chenevert TL, Yankeelov TE, Mountz JM, Kinahan PE, Kikinis R, Taouli B, Fenessay F, Kalpathy-Cramer J. Variations of dynamic contrast-enhanced magnetic resonance imaging in evaluation of breast cancer therapy response: a multicenter data analysis challenge. Transl Oncol. 2014;7:153–166.
19. Huang W, Chen Y, Fedorov A, Li X, Jajamovich GH, Malyarenko DI, Aryal MP, LaViolette PS, Oborski MJ, O'Sullivan F, Abramson RG, Jafari-Khouzani K, Afzal A, Tudorica A, Moloney B, Gupta SN, Besa C, Kalpathy-Cramer J, Mountz JM, Laymon CM, Muzi M, Schmainda K, Cao Y, Chenevert TL, Taouli B, Yankeelov TE, Fennessy F, Li X. The impact of arterial input function determination variations on prostate dynamic contrast-enhanced magnetic resonance imaging pharmacokinetic modeling: a multicenter data analysis challenge. Tomography. 2016;2:56–66.
20. Chenevert TL, Malyarenko DI, Newitt D, Jayatilake M, Tudorica A, Fedorov A, Kikinis R, Liu TT, Muzi M, Oborski MJ, Laymon CM, Li X, Yankeelov TE, Kalpathy-Cramer J, Mountz JM, Kinahan PE, Rubin DL, Fennessy F, Huang W, Hylton N, Ross BD. Errors in quantitative image analysis due to platform-dependent image-scaling. Transl Oncol. 2014;7:153–166.
21. Kalpathy-Cramer J, Zhao B, Goldgof D, Gu Y, Wang X, Yang H, Tan Y, Gillies R, Napel S. A comparison of lung nodule segmentation algorithms: methods and results from a multi-institutional study. J Digit Imaging. 2016;29:476–487.
22. Kalpathy-Cramer J, Mammonov A, Zhao B, Lu L, Cherezov D, Napel S, Echegaray S, Rubin D, McNitt-Gray M, Lo P, Sieren JC, Uthoff J, Dilger SKN, Driscoll B, Yeung I, Hadjiski L, Cha K, Balagurunathan Y, Gillies R, Goldgof D. Radiomics of lung nodules: a multi-institutional study of robustness and agreement of quantitative imaging features. Tomography. 2016;2:430–437.
23. Schmainda KM, Prah MA, Rand SD, Liu Y, Logan B, Muzi M, Rane SD, Da X, Yen YF, Kalpathy-Cramer J, Chenevert TL, Hoff B, Ross B, Cao Y, Aryal MP, Erickson B, Korfiatis P, Dondlinger T, Bell L, Hu L, Kinahan PE, Quarles CC. Multisite concordance of DSC-MRI analysis for brain tumors: results of a national cancer institute quantitative imaging network collaborative project. AJNR Am J Neuroradiol. 2018;39:1008–1016.
24. Malyarenko DI, Wilmes LJ, Arlinghuas LR, Jacobs MA, Juang W, Helmer KG, Taouli B, Yankeelov TE, Newitt D, Chenevert TL. QIN DAWG validation of gradient nonlinearity bias correction workflow for quantitative diffusion-weighted imaging in multicenter trials. Tomography. 2016;2:396–405.
25. Malyarenko D, Fedorov A, Bell L, Prah M, hectors S, Arlinghaus L, Muzi M, Solaiyappan M, Jacobs M, Fung M, Shukla-Dave A, McManus K, Boss M, Taouli B, Yankeelov TE, Quarles CC, Schmainda K, Chenevert TL, Newitt DC. Toward uniform implementation of parametric map Digital Imaging and Communication in Medicine standard in multisite quantitative diffusion imaging studies. J Med Imaging. 2018; 5:011006.

26. Newitt DC, Malyarenko D, Chenevert TL, Quarles CC, Bell L, Fedorov A, Fennessy F, Jacobs MA, Solaiyappan M, Hectors S, Taouli B, Muzi M, Kinahan PE, Schmainda KM, Prah MA, Taber EN, Kroenke C, Huang W, Arlinghaus LR, Yankeelov TE, Cao Y, Aryal M, Yen YF, Kalpathy-Cramer J, Shukla-Dave A, Fung M, Liang J, Boss M, Hylton N. Multisite concordance of apparent diffusion coefficient measurements across the NCI Quantitative Imaging Network. J Med Imaging. 2018;5:011003.
27. Beichel RR, Smith BJ, Bauer C, Ulrich EJ, Ahmadvand P, Budzevich MM, Gillies RJ, Goldgof D, Grkovski M, Hamarneh G, Huang Q, Kinahan PE, Laymon M, Mountz JP, Nehmeh S, Oborski MJ, Tan Y, Zhao B, Sunderlnd JJ, Buatti JM. Multi-site quality and variability analysis of 3D FDG PET segmentations based on phantom and clinical image data. Med Phys. 2017;44:479–496.
28. Balagurunathan Y, Beers A, Kalpathy-Cramer J, McNitt-Gray M, Hadjiski L, Zhao B, Zhu J, Yang H, Yip SSF, Aerts HJWL, Napel S, Cherezov D, Cha K, Chan HP, Flores C, Garcia A, Gillies R, Goldgof D. Semi-automated pulmonary nodule interval segmentation using NLST data. Med Phys. 2018;45:1093–1107.
29. https://imaging.cancer.gov/programs_resources/specialized_initiatives/qin/about/default.htm (accessed Nov 28, 2018).
30. The Working Groups of the Quantitative Imaging Network are: (1) Clinical Trials Design and Development, (2) Bioinformatics IT and Data Sharing, (3) Image Analysis and Performance Metrics, (4) MRI, and (5) PET-CT.
31. Tan Y, Lu L, Bonde A, Wang D, Qi J, Schwartz HL, Zhao B. Lymph node segmentation by dynamic programming and active contours. Med Phys. 2018;45:2054–2062.
32. Yan J, Schwartz LH, Zhao B. Semi-automatic segmentation of liver metastases on volumetric CT images. Med Phys. 2015;42:6283–6293.
33. Tan Y, Schwartz LH, Zhao B. Segmentation of lung tumors on CT scans using watershed and active contours. Med Phys. 2013;40:043502.
34. Hylton NM. Vascularity assessment of breast lesions with gadolinium-enhanced MR imaging. Magn Reson Imaging Clin N Am. 1999;7:411–420.
35. Partridge SC, Heumann EJ, Hylton NM. Semi-automated analysis for MRI of breast tumors. Stud Health Technol Inform. 1999;62:259–260.
36. Keenan KE, Wilmes LJ, Aliu SO, Newitt DC, Jones EF, Boss MA, Stupic KF, Russek SE, Hylton NM, Design of a breast phantom for quantitative MRI. J Magn Reson Imaging. 2016;44:610–619.
37. O JH, Lodge MA, Wahl RL. Practical PERCIST: a simplified guide to PET response criteria in solid tumors 1.0. Radiology. 2016;280:576–584.
38. Hyun OJ, Luber BS, Leal JP, Wang H, Bolejack V, Schuetze SM, Schwartz LH, Helman LJ, Reinke D, Baker LH, Wahl RL. Response to early treatment evaluated with 18F-FDG PET and PERCIST 1.0 predicts survival in patients with Ewing sarcoma family of tumors treated with a monoclonal antibody to the insulinlike growth factor 1 receptor. J Nucl Med. 2016;57:735–740.
39. O JH, Wahl RL. PERCIST in perspective. Nucl Med Mol Imaging. 2018;52:1–4.
40. Zhao XR, Zhang Y, Yu YH. Use of ^{18}F-FDG PET/CT to predict short-term outcomes early in the course of chemoradiotherapy in stage III adenocarcinoma of the lung. Oncol Lett. 2018;16:1067–1072.
41. Rubin DL, Willrett D, O'Connor MJ, Hage C, Kurtz C, Moreira DA. Automated tracking of quantitative assessments of tumor burden in clinical trials. Trans Oncol. 2014;7:23–35.
42. Yankeelov TE, Mankoff DA, Schwartz LH, Lieberman FS, Buatti JM, Mountz JM, Erickson BJ, Fennessy FM, Huang W, Kalpathy-Cramer J, Wahl RL, Linden HM, Kinahan PE, Zhao B, Hylton NM, Gillies RJ, Clarke L, Nordstrom R, Rubin DL. Quantitative imaging in cancer clinical trials. Clin Cancer Res. 2016;22:284–290.
43. Mountz JM, Yankeelov TE, Rubin DL, Buatti JM, Erikson BJ, Fennessy FM, Gillies RJ, Huang W, Jacobs MA, Kinahan PE, Laymon CM, Linden HM, Mankoff DA, Schwartz LH, Shim H, Wahl RL. Letter to cancer center directors: progress in quantitative imaging as a means to predict and/or measure tumor response in cancer therapy trials. J Clin Oncol. 2014;32:2115–2116.

Comparison of Voxel-Wise and Histogram Analyses of Glioma ADC Maps for Prediction of Early Therapeutic Change

Thomas L. Chenevert[1], Dariya I. Malyarenko[1], Craig J. Galbán[1], Diana M. Gomez-Hassan[1], Pia C. Sundgren[2], Christina I. Tsien[3], and Brian D. Ross[1]

[1]Department of Radiology, University of Michigan Medical School, Ann Arbor, MI; [2]Department of Clinical Sciences/Radiology Lund University, Lund, Sweden; and [3]Department of Radiation Oncology, Washington University School of Medicine, St. Louis, MO

Corresponding Author:
Thomas L. Chenevert, PhD
University of Michigan Hospitals, 1500 E. Medical Center Dr., UHB2 Room A209; Ann Arbor, MI 48109-5030;
E-mail: tlchenev@med.umich.edu.

Key Words: glioma therapy response, apparent diffusion coefficient, functional diffusion map, voxel-wise analysis, quantitative response metric

Abbreviations: Apparent diffusion coefficient (ADC), functional diffusion mapping (fDM), diffusion-weighted imaging (DWI), parametric response map (PRM), region of interest (ROI), volume of interest (VOI), Kaplan–Meier (KM), pretreatment (preTx), midtreatment (midTx), cumulative distribution function (CDF), T1-weighted with gadolinium enhancement (T1Gd), number of averages (NAV), Meta-image Header (MHD), area under (AUC) receiver operating curve (ROC) (AU-ROC), mean cumulative probability difference (mCPD), confidence interval (CI)

ABSTRACT

Noninvasive imaging methods are sought to objectively predict early response to therapy for high-grade glioma tumors. Quantitative metrics derived from diffusion-weighted imaging, such as apparent diffusion coefficient (ADC), have previously shown promise when used in combination with voxel-based analysis reflecting regional changes. The functional diffusion mapping (fDM) metric is hypothesized to be associated with volume of tumor exhibiting an increasing ADC owing to effective therapeutic action. In this work, the reference fDM-predicted survival (from previous study) for 3 weeks from treatment initiation (midtreatment) is compared to multiple histogram-based metrics using Kaplan–Meier estimator for 80 glioma patients stratified to responders and nonresponders based on the population median value for the given metric. The ADC histogram metric reflecting reduction in midtreatment volume of solid tumor (ADC $< 1.25 \times 10^{-3}$ mm^2/s) by >8% population-median with respect to pretreatment is found to have the same predictive power as the reference fDM of increasing midtreatment ADC volume above 4%. This study establishes the level of correlation between fDM increase and low-ADC tumor volume shrinkage for prediction of early response to radiation therapy in patients with glioma malignancies.

INTRODUCTION

Clinical oncology trials actively seek robust radiological markers of early response to cancer therapy to noninvasively guide patient treatment plans. By measuring water mobility known to be altered by tissue cellular constituents (1-3), diffusion-weighted imaging (DWI) is able to provide information on changes in tumor cellular density related to cytotoxic therapy response (4-7). Growth of viable tumor leads to increased cell density and reduced water mobility, while effective therapy decreases cell density and increases water mobility. Higher water mobility independent of therapy is also observed for necrotic tissue (8, 9). DWI measurements are typically represented as quantitative parametric diffusion maps of the apparent diffusion coefficient (ADC) based on an assumed monoexponential DWI signal decay with increasing diffusion-weighting strength (denoted by *b*-value) (5-7, 10). The therapy-related changes in the ADC maps can be quantitatively characterized spatially by the functional diffusion map (fDM) method within the general class of parametric response mapping (PRM). These approaches deal with tumor heterogeneity to display significant regional change of treatment responsive/resistant voxels, while supplying a global quantitative response metric (11-13). PRM fDM has been shown to allow earlier prediction of glioma therapy response and more accurate prediction of survival relative to conventional neuroimaging metric (12). To provide robust alternative to invasive biopsies, the predictive power of this promising method needs to be linked to changes in tumor histopathological properties.

The fDM method (13) generally requires robust spatial registration of tumor volumes between longitudinal scans, which is potentially dependent on specific registration algorithm parameters and thus may be prone to introducing additional repeatability errors due to variation in image registration workflow. The method also relies on precise tumor region/volume-of-interest (ROI/VOI) definition and on matching voxels during potentially rapid tumor growth or shrinkage. By virtue of the

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2018.00049

underlying statistical assumptions (14), fDM analysis includes thresholding for significant change, which can be nonspecific to the ADC range and tumor density as was originally proposed in (13). Notwithstanding demonstrated promising predictive value of the fDM metrics (11, 12), its direct relation to the biophysical properties of dense versus necrotic tumor volumes has not yet been clearly established. In principle, significant changes of fDM may occur over the full range of ADC values (both for restricted and less restricted diffusion (1)).

An alternative approach that forfeits retention of spatial origin of voxels within tumor is to perform histogram analysis of ADC voxel values (6, 15). Intralesion heterogeneity is retained by the histogram, although direct spatial identification of responsive/resistant regions is lost. The histogram analysis approach has several benefits. First, this approach removes dependence on technical performance of an image volume registration step, as well as assumptions that regions of rapid tumor growth/shrinkage are adequately coregistered. Second, the ADC histogram inherently facilitates segmentation of tumor based on tissue density reflected by water mobility (6). Third, this also allows direct identification of naturally high water mobility within cystic necrotic tumor tissue present before initiation of treatment to potentially distinguish from additional necrosis (9) resultant from cytotoxic treatment.

The purpose of the present study was to evaluate predictive power of several histogram-based ADC metrics and their correlation to fDM using quantitative DWI data from a common cohort of patients with glioma treated by chemoradiation. Because the overall objective was a technical comparison of the metrics, image processing and image segmentation were held constant across metrics derivation, and "survival" was used as the sole clinical outcome.

METHODOLOGY

This study analyzed Kaplan–Meier (KM) survival prediction for multiple ADC histogram metrics versus reference fDM-derived from quantitative DWI data including pretreatment (preTx) and 3-week midtreatment (midTx) imaging of a cohort of patients with high-grade glioma that underwent chemoradiotherapy treatment with longitudinal radiological surveillance (12). The baseline preTx scan was acquired postsurgery/biopsy before the start of treatment. The survival was assessed from the time of the diagnosis. All quantitative DWI and statistical analysis was performed using home-built routines developed in MATLAB 7 (MathWorks, Natick, MA). KM estimate of cumulative distribution function (CDF) for survival probability was generated using MATLAB built-in "*ecdf*" routine. The KM stair-step graphs for CDF censoring visualization were generated using MATLAB Central "*MatSurv*" function (16).

Patient Cohort

Details on patient cohort, treatment schedule, and diffusion scans are previously reported (12). Informed consent for images and medical record use for research was approved by institutional review board and renewed over the study period from 2000 to 2011. In total, 25 additional consented study subjects (scanned between 2007 and 2011) with grade 3 and 4 primary brain tumors were included into the present analysis and were added to the 60 previously analyzed (2000 to 2006) (12). Overall patient demographics, pathology grade, treatment plans, response status, and imaging schedule were not significantly different from the original study and are not detailed here. Both patient survival (median months, 13.7 and 14.5) and pathology grade (3-to-4 ratios, 28% and 25%) were consistent between acquisition-date subgroups (Student's *t*-test, $P > .7$), ensuring nominally unbiased clinical outcome measures of the combined group. Only preTx and 3-week midTx imaging were included in this study owing to previously demonstrated relevance for early response survival prediction by fDM (12). Only survival was used and no other clinical outcomes such as time-to-progression were considered.

Imaging Studies

Clinical MRI scans including quantitative diffusion MRI and standard MRI (fluid attenuation inversion recovery, T2-weighted, and T1-weighted with gadolinium enhancement [T1Gd] and without Gd enhancement) were performed for all imaging endpoints on 1.5 T MRI system (General Electric, Waukesha, WI; n = 45 patients) and on 3 T MRI scanner (Philips, Best, The Netherlands; n = 40 patients). The 75% of the initial (2000–2006) study scans were performed on1.5 T, while 3 T scanner system was used exclusively for the (2007–2011) study subgroup. Consistent with the nominal independence on the acquisition-date, survival and pathology grade were not biased by the scanner subgroups ($P > .3$).

DWI protocol prescribed single-shot echo-planar imaging acquisition of three orthogonal–axial DWI scans with *b*-values = 0 and 1000 s/mm^2 using a 16-channel head-coil. On the 1.5 T system, 24 6-mm axial-oblique sections were acquired using a 22-cm field of view and 128 matrix (voxel size = 17.7 mm^3) repetition time = 10 000 ms; echo time = 71 to 100 ms, and number of averages (NAV) = 1. On the 3 T system, at least 28 4-mm axial– oblique sections were acquired through the brain using a 24-cm field of view and 128 matrix (voxel size = 14 mm^3; repetition time = 2.636 milliseconds; TE = 46 ms; NAV = 1 for $b = 0$, and NAV = 2 for $b = 1000$ s/mm^2. Parallel imaging (sensitivity-encoding factor = 3) was used at 3 T to reduce spatial distortion. PreTx and midTx scans for a given patient were performed on the same system.

ADC Parametric Map Generation

The diffusion images for the three orthogonal directions were combined into trace DWI to calculate an ADC map. All acquired data were stored and distributed in Digital Image Communication in Medicine (DICOM) format (17). ADC was fit as a slope of log-signal DWI as a function of *b*-value up to $b_{max} = 1000$ s/mm^2. For previously published data subset (12), image registration volumes and tumor segmentations were reused from prior analysis. For additional study subjects, the resulting low *b*-value, high *b*-value, and ADC maps were exported as Meta-image Header (MHD) format (18) for volumetric spatial registration to the anatomical pretreatment T1Gd images using the Elastix toolkit (19) with full-affine transformation. The low *b*-value DWI volume was used to drive image registration using the mutual information figure of merit, and the resultant spatial transformation was automatically applied to the corresponding high *b*-value and ADC volumes. Tumor-encompassing ROIs pre-

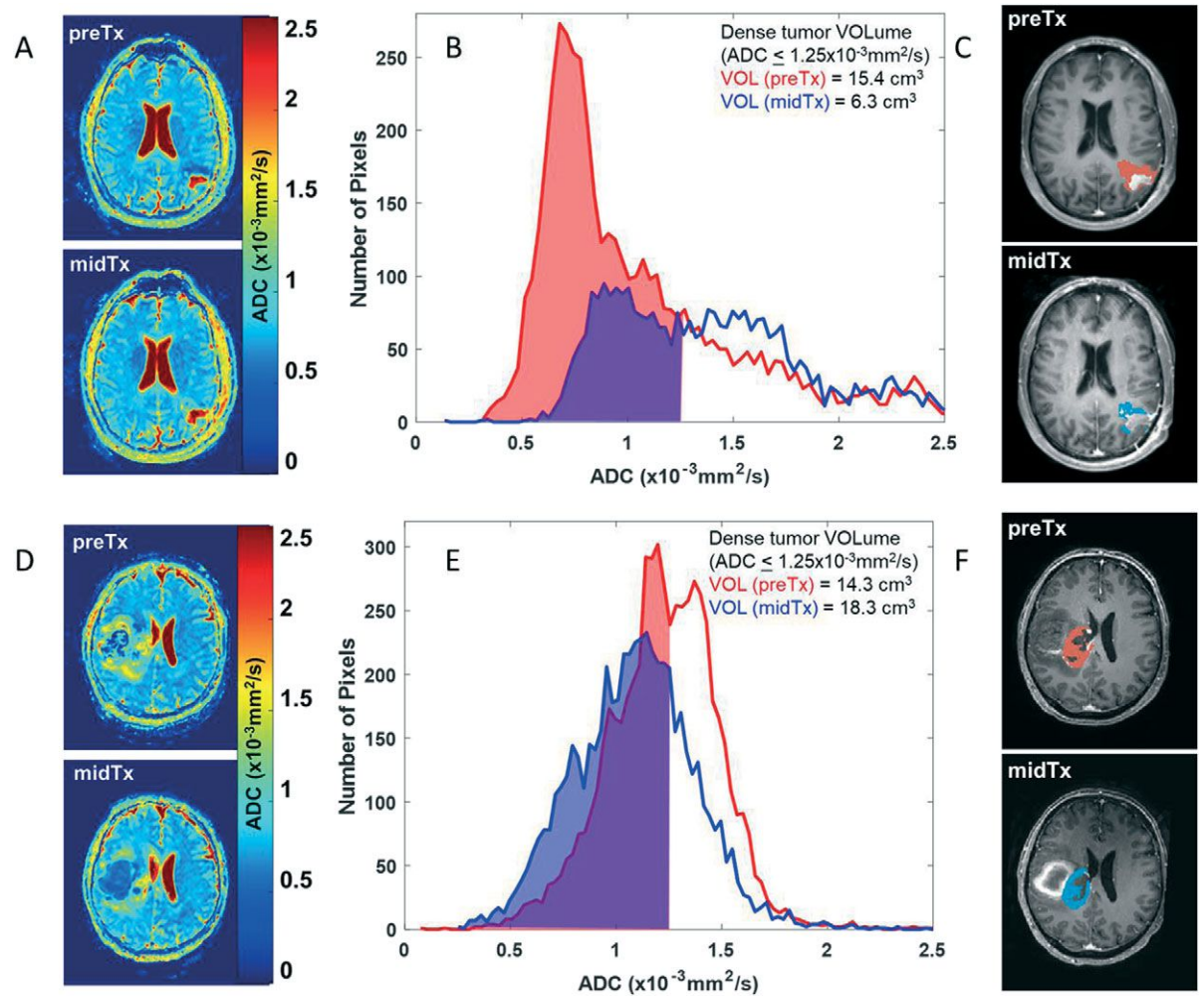


Figure 1. Left vertically arranged images (A, D) show ADC maps for preTx and midTx imaging time-points of 2 patients with glioma that responded favorably (A) and did not responded (D) to chemoradiation therapy. Common scale for the ADC maps is indicated by the color bar. The center panes (B, E) illustrate the corresponding tumor volume ADC histograms (preTx: red, and midTx: blue) and tumor voxel volumes (filled) below ADC threshold of 1.25 ($\times 10^{-3}$ mm^2/s). The corresponding integrated volumes of the dense tumor are listed in the legend. The spatial location of thresholded histogram voxels is overlaid in red and blue on a single representative slice of each patient preTx and midTx T1Gd images on the right in (C, F), used as a reference for tumor ROI definition.

viously defined by two experienced (>20 years) radiologists on the T1Gd images (coregistered to ADC maps) were imported into 3D Slicer (20) and converted to MHD ROI labels. These MHD VOI masks were then imported to MATLAB and applied to ADC maps to generate histograms of voxel ADC values within the defined tumor VOI (Figure 1). Additional VOIs (median volume, 5.4 cm^3; range, 3.6–7.6 cm^3) were defined on 3 slices for frontal normal-appearing white matter (contralateral to tumor) to confirm negligible system-specific ADC bias (21, 22) in two scanner subgroups [median ADC ($\times 10^{-3}$ mm^2/s): 0.785 (1.5 T) and 0.789 (3 T); $P = .19$].

ADC Histogram Metrics

Histogram "volume" metrics (in cubic centimeter units) were generated by numerically integrating the voxels up to specified ADC thresholds (without reference to spatial location other than being within the specified tumor VOI) and multiplying by the known image voxel volume. The upper thresholds for low-ADC histogram portion (presumably reflecting more cellular-dense tumor) were sampled from 0.25 to 1.5 in steps of 0.25 ($\times 10^{-3}$ mm^2/s). The upper sampling bound of 1.5 ($\times 10^{-3}$ mm^2/s) was set to the previously published ADC value for necrotic tumor tissue (8). The standard whole-tumor histograms metrics, including ADC mean, median, and standard deviation were likewise evaluated for preTx and midTx imaging points separately and for their fraction-change with respect to preTx. The thresholds for survival-based therapy response prediction of each ADC histogram metric were dichotomized by population median values.

fDM Reference Metrics and KM Analysis

fDM analysis was performed as previously described (12). Only voxels present both in preTx and midTx tumor VOIs were stratified according to their change in ADC value (Figure 2, A and B)

into significantly increased (Vi, red, ADC change $> 0.55 \times 10^{-3}$ mm^2/s), decreased (Vd, blue, $<0.55 \times 10^{-3}$ mm^2/s), and the remainder unchanged (Vo, green, within the 0.55×10^{-3} mm^2/s 95% confidence interval [CI]). The total percentage of tumor with significant increase in diffusion value was calculated as 100% × Vi/(Vi + Vo + Vd) and used as the reference fDM biomarker.

The KM survival probability analysis was then performed for the choice metrics with predetermined (population-median) thresholds and the corresponding log-rank P-values (P_{KM}). Median fDM threshold was Vi > 4% ($P_{KM} = 0.0008$; Figure 2C; magenta KM line), which reasonably agreed with the optimized fDM threshold of 4.7% from the previous study (12) corresponding to maximum area under (AUC) receiver operating curve (ROC). Note that compared to the typical stair-step graphical representation (Figure 2C), the actual KM CDF curves would terminate before the last "stair-step" to exclude (unchanging) probability from the last censored patients (eg,

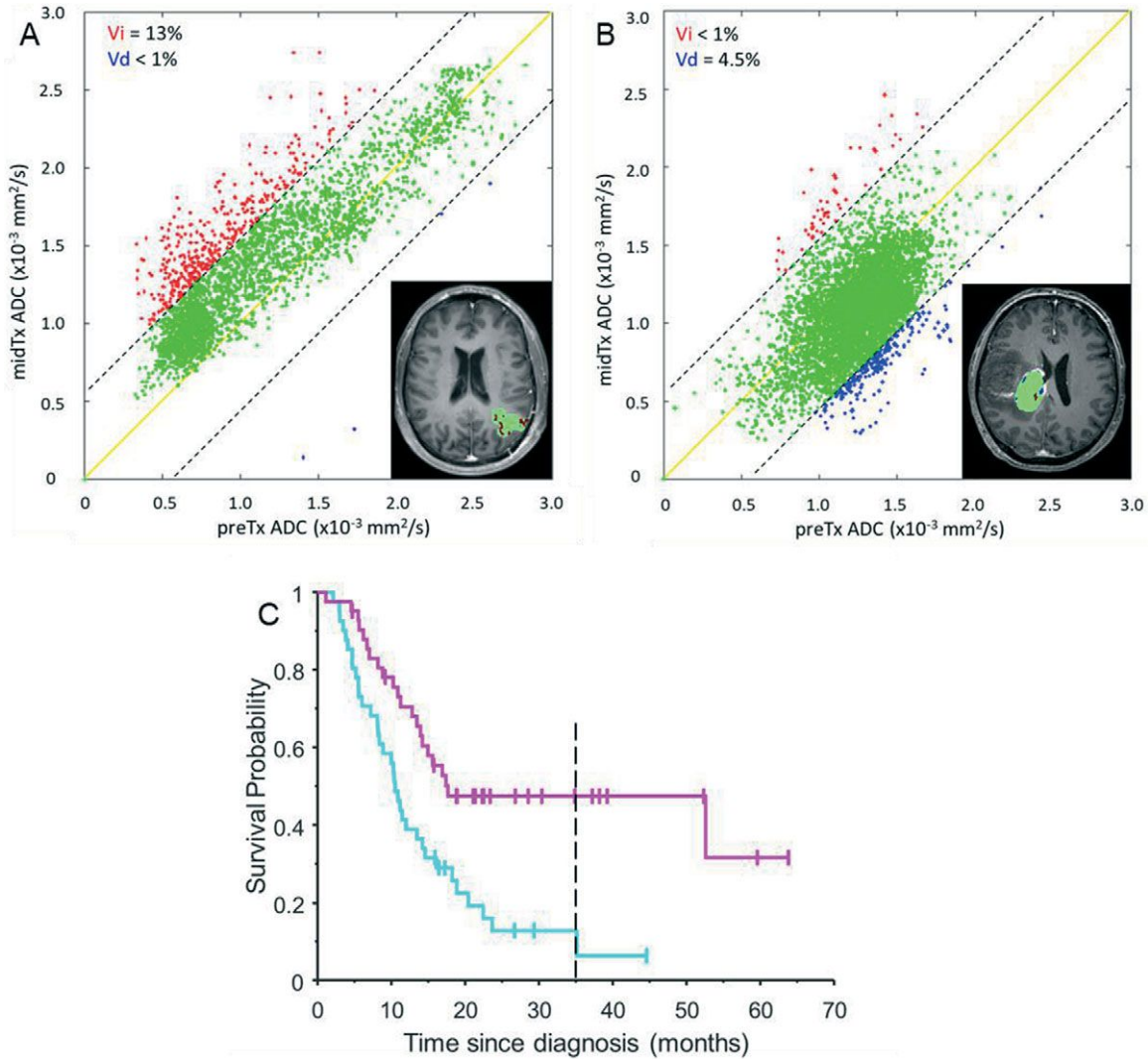


Figure 2. fDM metrics determined from midTx versus preTx ADC PRM scatter plots is overlaid on the T1Gd image inserts for the same two patients [responder (A) and nonresponder (B)] as in Figure 1 histograms. The dashed diagonal lines indicate 95% CI for the change encompassing green voxels corresponding to tumor regions not altered by therapy. The solid yellow line corresponds to the perfect fDM correlation. Red and blue areas mark tumor voxels with respective significant increase and decrease in ADC midTX verus preTx (summarized in the legends). (C) shows stair-step graph for reference fDM KM survival analysis of responders (magenta) and nonresponders (cyan) based on a median response threshold of 4% fDM-increase (magenta KM stair-step trend) for the whole glioma study population. Magenta and cyan KM trends correspond to the tumor fDM, respectively, above and below median response threshold. Vertical tick-marks along KM trends indicate individual patients whose survival times have been censored. Dashed vertical line corresponds to the minimal survival time included into the corresponding KM cumulative distribution function (CDF) probability analysis (excluding survival for the late censored patients).

at minimum CDF probability values of 0.07 and 0.3 for Figure 2C cyan and magenta trends, respectively).

Predictive power of each KM estimator was quantified by the mean cumulative probability difference (mCPD) between KM CDF curves (0.21 for reference fDM in Figure 2C). The KM curves for each sampled ADC metric were linearly interpolated to the common time-since-diagnosis axis corresponding to the fDM reference. The time-dependent survival probability differences between KM responder and nonresponder curves were correlated to that of the fDM reference to determine metrics with maximum KM "alignment" to the fDM. Pearson correlation, R_{fDM}, with $P_R < .05$ was considered significant. KM-length was determined as the minimal length of the two survival CDF curves for each metric. Similarity index was assessed by product of R_{fDM} and KM-length ratio, L_R, with respect to the fDM nonresponder reference (Figure 2C; vertical dashed line marks the end of the corresponding CDF at 35 months).

RESULTS

Figure 1 illustrates ADC histogram analysis for the representative responder and nonresponder tumors using a low-ADC volume threshold of 1.25×10^{-3} mm^2/s (ie, only counting voxels within VOI having an ADC below this value) to favor inclusion of dense tumor while excluding necrotic regions. The corresponding ADC maps (Figure 1, A and D) depict quantitative regional diffusion changes in response to therapy, more pronounced for the responder (Figure 1, A–C) (survival, >27 months), relative to the nonresponder in Figure 1, D–F (survival, <9 months). The low ADC tumor component between midTx and preTx is quantified by a 9 cm^3 decrease of integrated dense tumor volume for the responder (Figure 1B) versus a 4 cm^3 increase for nonresponder (Figure 1E). That is, the fractional change in the low-ADC component of the histogram (59% decrease) owing to an upward shift, and shape change is enhanced by exclusion of the high ADC contribution that attenuates whole-tumor volumetric change (32% decrease) and whole-tumor mean ADC (30% increase). The low-ADC histogram voxel overlays on T1Gd images (Figure 1, C and F) further illustrate how influence of the preexisting necrotic portion of the tumor is reduced by this analysis. Conversely, the nonresponder had an increase in dense tumor volume (by +28%) despite a reduction in whole-tumor volume (−6%). Although only central-tumor slices are shown in Figure 1, the histogram VOI analysis included all tumor slices.

Figure 2 illustrates fDM analysis for the same 2 subjects with diagnostic changes related to tumor response metrics (Figure 2A: Vi = 13%, red, and Figure 2B: Vd = 4.5% blue voxels) observed predominantly toward lower ADC values ($<1.5 \times 10^{-3}$ mm^2/s). The red or blue fDM voxels marking regions with respective significant increase or decrease in ADC are evidently clustered in the lower half of midTx versus preTx values for a responder (Figure 2A, red) and nonresponder (Figure 2B, blue). The voxels with significantly higher midTX ADC for responder are distributed more uniformly across the ADC range of dense and necrotic tumor ($[1.25 - 2.25] \times 10^{-3}$ mm^2/s). However, the necrotic portion of the tumor does not significantly contribute to Vi in fDM analysis owing to high baseline ADC. Much lower red fDM volume shifted toward higher (necrotic) midTX ADC ($>1.5 \times 10^{-3}$ mm^2/s) is observed for nonresponder in Figure 2B with a noticeable increase in blue fDM voxel areas corresponding to lower (dense-tumor) ADC ($<1.25 \times 10^{-3}$ mm^2/s) for midTx. As in Figure 1, fDM difference overlays are on a single slice (Figure 2, inserts), whereas the fDM analysis spans the full tumor volume.

The responder versus nonresponder KM thresholds for the select test histogram characteristics based on population-wise median values are summarized in Table 1 along with their KM mCPD and percent-similarity index to the fDM CDF reference (Figure 2C). These median thresholds were used for the corresponding KM survival analysis shown in Figure 3. Other histogram metrics (not included) has shown <50% absolute similarity to fDM KM reference. Low predictive power was observed for all preTx metrics (median response threshold, $P_{KM} > .1$, mCPD < 0.06), reflecting dependence of response on the therapy administration. As expected, the corresponding KM CDF (Figure 3, A, D, and G) have shown low absolute similarity (<35%) to refer-

Table 1. Population-wise Median KM Response-Threshold, mCPD, and Similarity to Reference KM fDM for Select ADC Histogram Metrics

Metric	Median KM Threshold (P_{KM}[a])	mCPD	Similarity Index (%)
preTx Mean ADC (10^{-3} mm^2/s)	1.19 (0.36)	0.06	20
midTx Mean ADC (10^{-3} mm^2/s)	1.25 (0.0033)	0.2	13
% Change[b] Mean ADC	1.83 (0.05)	0.17	51
preTx Volume (cm^3)	32.5 (0.75)	0.05	35
midTx Volume (cm^3)	27.6 (0.38)	0.1	13
% Change[b] Volume	−0.8 (0.011)	0.18	−87
preTx LowADC Vol[c] (cm^3)	17.6 (0.51)	0.04	−18.6
midTx LowADC Vol[c] (cm^3)	15 (0.047)	0.14	−86
% Change[b] LowADC Vol[b]	−7.8 (0.0006)	0.22	−92.5

[a] P-value of population-wise median KM response-threshold.
[b] % Change = 100% (midTx − preTx)/preTx.
[c] Volume of tumor with ADC $\leq 1.25 \times 10^{-3}$ mm^2/s.

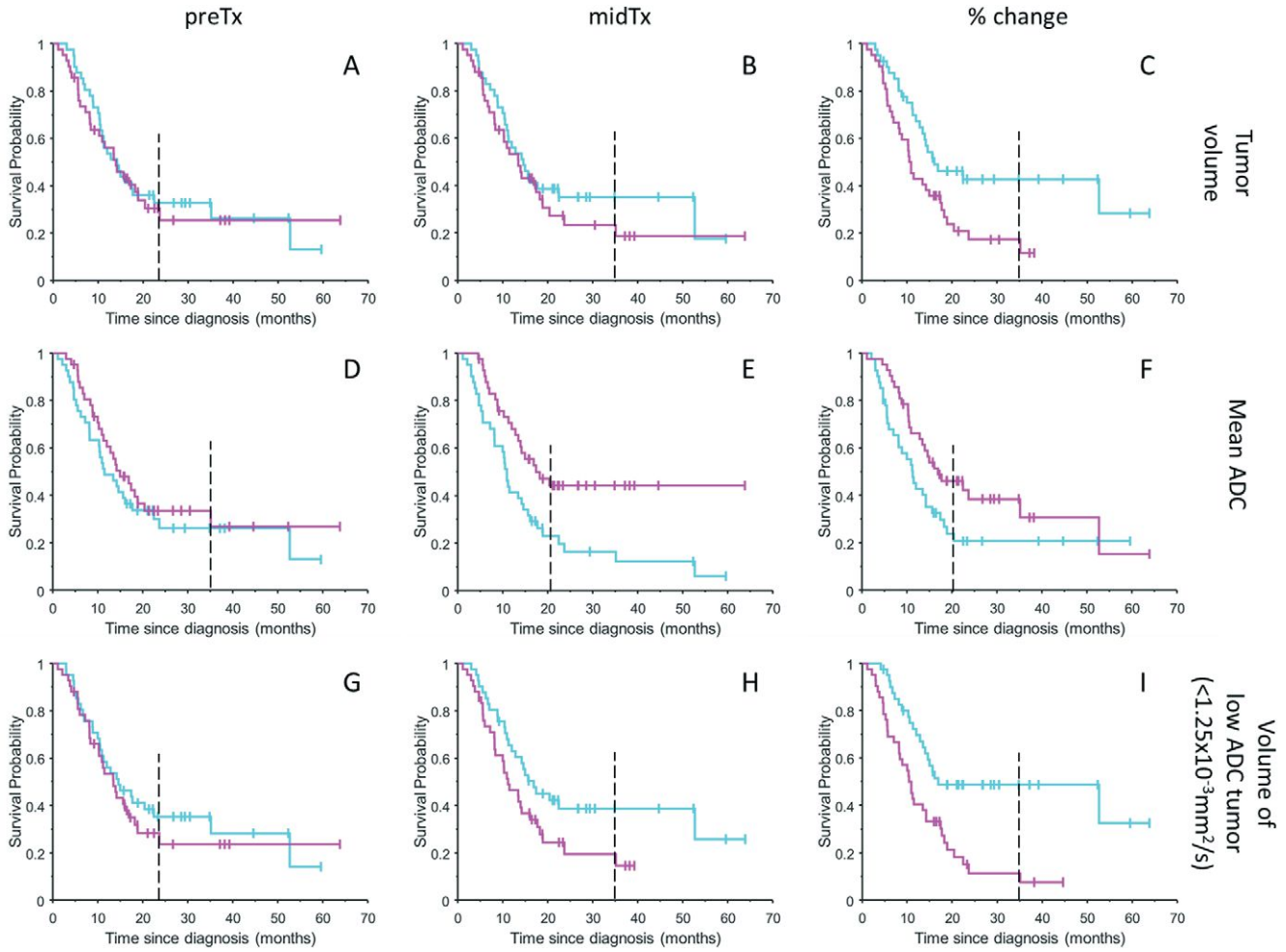


Figure 3. KM survival probability analysis results are summarized as stair-step graphs for conventional histogram metrics of total T1Gd tumor volume in (A–C), mean ADC in (D–F), and low ADC ($<1.25 \times 10^{-3}$ mm^2/s) histogram volume in (G–I). Magenta and cyan KM trends correspond to the tumor characteristics, respectively, above and below median response threshold for the studied ADC histogram metrics. The color flip from cyan to magenta for responder KM trends (with higher probability of survival) between mean ADC (D–F) and volume-based metrics (A–C, G–I) reflects negative change in tumor volume versus positive change in ADC metrics. Time-dependent distance between KM curves reports on predictive power of the studied histogram metrics. Vertical tick-marks along KM trends indicate individual patients whose survival times have been censored. Dashed vertical line corresponds to the minimal survival time included into the corresponding KM CDF probability analysis (excluding survival for the late censored patients).

ence KM fDM (Figure 2C) that was based on changes between midTx and preTx. Significant enhancement of KM CDF separation (P_{KM} = 0.003–0.05, mCPD = 0.17–0.2) was observed for midTx ADC (Figure 3E) above a median response threshold of 1.25($\times 10^{-3}$ mm^2/s), as well as for change in whole-tumor mean ADC and total tumor-volume differences above versus below 1%–2% (Figure 3, C, E, and F). However, a notably high number (fourteen) of censored patients (Figure 3E, magenta ticks) made CDF estimate for midTx ADC metric unreliable beyond 21-months survival (Figure 3E, dashed). The similarity of the fractional volume KM to reference fDM was −87%, notably higher than that for significant (midTx and fractional change) ADC metrics, consistent with volumetric nature of the fDM analysis. This is also consistent with observation of high KM similarity (−86%) for low-ADC volume midTx (Figure 3H). The general color "flip" for responder KM trends based on volume metrics (Figure 3, A–C, G–I, cyan) versus ADC metrics (Figure 3, D–F, magenta) reflected negative change in tumor volume versus positive change in ADC metrics related to higher probability of survival.

The best KM survival probability CDF estimator in Figure 3I (with maximum mCPD = 0.22 and minimum $P_{KM} < 0.001$) was based on the fraction low-ADC volume shrinkage (cyan KM trend). This estimator used combined tumor volume change and tumor density (ADC-threshold $< 1.25 \times 10^{-3}$ mm^2/s) information. The fractional low-ADC volume metric clearly showed similar predictive power (relative distance between KM CDF) as reference fDM KM (Figure 2C, mCPD = 0.21) based on the

increased fDM PRM midTx ("magenta" trend). The reliable CDF estimate for both reference (Figure 2C) and fractional low-ADC volume (Figure 3I) was confirmed by a small number (two) of patients censored beyond minimal CDF values of the corresponding KM trends (at survival probabilities of 0.3 and 0.07). The bulk of the KM differences between responders and nonresponders was evidently related to the low ADC volume midTx (Figure 3H), rather than preTX volume (Figure 3G), confirming that the functional response was triggered by treatment. The decreasing low-ADC volume midTX versus preTx (less than −8%, $P_{KM} < 0.001$) in Figure 3I, was significantly (negatively) correlated to increasing fDM (>4%, $P_{KM} < 0.001$) in Figure 2C and Table 1 (−92.5%), confirming fDM relation to shrinking tumor volume.

DISCUSSION

The decrease in low-ADC volume was found to be a good predictor of KM survival (treatment response) most similar to the fDM reference. The strong alignment between KM curves for fDM and low-ADC volume metrics confirms that the early response prediction power of increasing fDM likely stems from decreasing volume of shrinking dense tumor observed as early as 3 weeks after radiation therapy for glioma tumors. Interestingly, the fDM population-median KM threshold for responders versus nonresponders of 4% was still close to 4.7% that maximized AU-ROC as previously determined (12) despite the additional 25 subjects. Another supporting observation is that the population-median response threshold for mean ADC-based KM survival probability midTx corresponded to the dense tumor low-ADC integration limit of 1.25×10^{-3} mm^2/s. The proximity of median thresholds for fractional ADC and tumor volume changes to 0% likely reflected KM sensitivity to the sign of the effect (increasing ADC and decreasing volume) rather than absolute metric value. The fact that no significance was observed for preTx low-ADC volume itself, suggested that midTx volume change was indeed reflective of the therapy efficacy. This specific relation to reduction of the dense tumor ADC volume and treatment option provided independent evidence for the biophysical origin of the fDM predictive power. Our analysis effectively revealed that fDM portions with low-ADC midTx report on the therapy response.

The main limitation of this study was that the data analysis was restricted to only two imaging end points, precluding evaluation of relative longitudinal changes in the histogram metrics over the full course of radiological surveillance. Furthermore, the KM thresholds were not optimized by AU-ROC analysis or cross-validation. These restrictions were intentional for the largely technical aims of this study to determine the ADC histogram metrics that had early response prediction power similar to the reference fDM, as shown by previous work (12), and to maximize method consistency across histogram and fDM analyses, reducing dependence on any residual study bias. For this reason, ADC histograms were derived from the same coregistered image sets and the same tumor segmentations as used to generate the reference fDM metrics, even though ADC histogram analysis can be performed on non-coregistered images. This study design precluded evaluation of sensitivity of low-ADC histogram-based segmentation to image registration-related errors. For ADC histogram threshold method, the specific voxel locations are less important, and hence higher immunity is potentially expected to coregistration errors. This should be a topic of a future study.

Others have applied alternative ADC histogram-based analyses in the context of newly diagnosed (6, 10, 15) and recurrent (23) glioblastoma to predict response to antivascular chemotherapy used alone or in combination with radiation treatment. Technical aspects of histogram analysis varied. Bimodal mixed normal distribution fitting of the whole tumor ADC histogram into means of the low-ADC curve and high-ADC curve was performed by Pope et al. (10, 15, 23). In contrast, Wen et al. (6) analyzed specific percentile points of the ADC histogram. However, both methods consistently found greater predictive content in the low-ADC regime. Prediction metrics in both of these alternative histogram approaches were expressed in physical diffusion units (ie, square millimeter per second), whereas the method presented in this study focused on volume (ie, in cubic centimeter units) of ostensibly dense tumor defined by an ADC below a specified value, 1.25×10^{-3} mm^2/s.

The low-ADC volume approach presented here parallels similar logic used to assess traditional response metrics based on tumor shrinkage assessed by conventional neuroimaging (24-26), although it exploits tumor density segmentation qualities inherent in diffusion mapping. A common feature in these various diffusion histogram approaches and fDM (or PRM) is a framework to deal with tumor heterogeneity and to avoid inclusion of preexisting cystic/necrotic portions of the tumor that can attenuate sensitivity to therapeutic changes in viable tumor. Response to treatment (or tumor progression) can be spatially nonuniform as well, and fDM/PRM provides means to map responsive/resistant/progression regions (11, 12, 27).

The current study design amplified ADC measurement sensitivity to the therapeutic effect by performing longitudinal patient surveillance scans on the same MRI system. Although desirable, this level of control may be challenging in the clinical setting. When multiple scanners are used, systematic biases may increase between-scan variability (eg, due to spatial *b*-value bias for anatomy at different offsets from isocenter (21, 22). For longitudinal studies, these errors may potentially increase the population histogram noise and attenuate the absolute ADC measurement sensitivity to the therapeutic effect. In principle, such systematic errors should be monitored similar to normal-appearing white matter analysis in this study [or using phantoms with known ADC (21, 22)] and, when present, corrected using MRI system gradient characteristics before population ADC histogram analysis.

In conclusion, fDM changes diagnostic of early therapy response for high-grade glioma tumors are confirmed using comprehensive analysis of multiple ADC histogram metrics. Reduction in solid (non-necrotic) tumor volume correlates with low-ADC fDM changes. Histogram-based ADC segmentation facilitates elimination of high-mobility (necrotic) tissue, allowing for focusing on shrinkage of low-mobility (cellular-dense) tumor regions.

ACKNOWLEDGMENTS

This research was supported by National Institutes of Health Grants: U01CA166104, R44CA210825, and P01CA085878, and by the Swedish Cancer Society CAN 2016/365.

Disclosures: TLC, CJG, and BDR are coinventors on intellectual property assigned to and managed by the University of Michigan licensed by Imbio for histogram and fDM analysis.

REFERENCES

1. Ellingson BM, Malkin MG, Rand SD, Connelly JM, Quinsey C, LaViolette PS, Bedekar DP, Schmainda KM. Validation of functional diffusion maps (fDMs) as a biomarker for human glioma cellularity. J Magn Reson Imaging. 2010;31:538–548.
2. Le Bihan D. Molecular diffusion, tissue microdynamics and microstructure. NMR Biomed. 1995;8:375–386.
3. Squillaci E, Manenti G, Cova M, Di Roma M, Miano R, Palmieri G, Simonetti G. Correlation of diffusion-weighted MR imaging with cellularity of renal tumours. Anticancer Res. 2004;24:4175–4179.
4. Chenevert TL, Stegman LD, Taylor JM, Robertson PL, Greenberg HS, Rehemtulla A, Greenberg HS, Rehemtulla A, Ross BD. Diffusion magnetic resonance imaging: an early surrogate marker of therapeutic efficacy in brain tumors. J Natl Cancer Inst. 2000;92:2029–2036.
5. Nagane M, Kobayashi K, Tanaka M, Tsuchiya K, Shishido-Hara Y, Shimizu S, Shiokawa Y. Predictive significance of mean apparent diffusion coefficient value for responsiveness of temozolomide-refractory malignant glioma to bevacizumab: preliminary report. Int J Clin Oncol. 2014;19:16–23.
6. Wen Q, Jalilian L, Lupo JM, Molinaro AM, Chang SM, Clarke J, Prados M, Nelson SJ. Comparison of ADC metrics and their association with outcome for patients with newly diagnosed glioblastoma being treated with radiation therapy, temozolomide, erlotinib and bevacizumab. J Neurooncol. 2015;121:331–339.
7. Qu J, Qin L, Cheng S, Leung K, Li X, Li H, Dai J, Jiang T, Akgoz A, Seethamraju R, Wang Q, Rahman R, Li S, Ai L, Jiang T, Young GS. Residual low ADC and high FA at the resection margin correlate with poor chemoradiation response and overall survival in high-grade glioma patients. Eur J Radiol. 2016;85:657–664.
8. Chenevert TL, McKeever PE, Ross BD. Monitoring early response of experimental brain tumors to therapy using diffusion magnetic resonance imaging. Clin Cancer Res. 1997;3:1457–1466.
9. Higano S, Yun X, Kumabe T, Watanabe M, Mugikura S, Umetsu A, Sato A, Yamada T, Takahashi S. Malignant astrocytic tumors: clinical importance of apparent diffusion coefficient in prediction of grade and prognosis. Radiology. 2006;241:839–846.
10. Pope WB, Kim HJ, Huo J, Alger J, Brown MS, Gjertson D, Sai V, Young JR, Tekchandani L, Cloughesy T, Mischel PS, Lai A, Nghiemphu P, Rahmanuddin S, Goldin J. Recurrent glioblastoma multiforme: ADC histogram analysis predicts response to bevacizumab treatment. Radiology. 2009;252:182–189.
11. Ellingson BM, Malkin MG, Rand SD, LaViolette PS, Connelly JM, Mueller WM, Schmainda KM. Volumetric analysis of functional diffusion maps is a predictive imaging biomarker for cytotoxic and anti-angiogenic treatments in malignant gliomas. J Neurooncol. 2011;102:95–103.
12. Hamstra DA, Galban CJ, Meyer CR, Johnson TD, Sundgren PC, Tsien C, Lawrence TS, Junck L, Ross DJ, Rehemtulla A, Ross BD, Chenevert TL. Functional diffusion map as an early imaging biomarker for high-grade glioma: correlation with conventional radiologic response and overall survival. J Clin Oncol. 2008;26:3387–3394.
13. Moffat BA, Chenevert TL, Lawrence TS, Meyer CR, Johnson TD, Dong Q, Tsien C, Mukherji S, Quint DJ, Gebarski SS, Robertson PL, Junck LR, Rehemtulla A, Ross BD. Functional diffusion map: a noninvasive MRI biomarker for early stratification of clinical brain tumor response. Proc Natl Acad Sci U S A. 2005;102:5524–5529.
14. Hastie T, Tibshirani I, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, NY: Springer; 2001.
15. Pope WB, Lai A, Mehta R, Kim HJ, Qiao J, Young JR, Xue X, Goldin J, Brown MS, Nghiemphu PL, Tran A, Cloughesy TF. Apparent diffusion coefficient histogram analysis stratifies progression-free survival in newly diagnosed bevacizumab-treated glioblastoma. AJNR Am J Neuroradiol. 2011;32:882–889.
16. Berglund AE. MatSurv.m 2017. https://github.com/aebergl/MatSurv.
17. Clunie DA. DICOM structured reporting and cancer clinical trials results. Cancer Inform. 2007;4:33–56.
18. MHD: image metadata format Public Wiki2014. https://itk.org/Wiki/ITK/MetaIO/Documentation.
19. Klein S, Staring M, Murphy K, Viergever MA, Pluim JP. elastix: a toolbox for intensity-based medical image registration. IEEE Trans Med Imaging. 2010;29:196–205.
20. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin JC, Pujol S, Bauer C, Jennings D, Fennessy F, Sonka M, Buatti J, Aylward S, Miller JV, Pieper S, Kikinis R. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging. 2012;30:1323–1341.
21. Malyarenko D, Galban CJ, Londy FJ, Meyer CR, Johnson TD, Rehemtulla A, Ross BD, Chenevert TL. Multi-system repeatability and reproducibility of apparent diffusion coefficient measurement using an ice-water phantom. J Magn Reson Imaging. 2013;37:1238–1246.
22. Mulkern RV, Ricci KI, Vajapeyam S, Chenevert TL, Malyarenko DI, Kocak M, Poussaint TY. Pediatric brain tumor consortium multisite assessment of apparent diffusion coefficient z-axis variation assessed with an ice-water phantom. Acad Radiol. 2015;22:363–369.
23. Pope WB, Qiao XJ, Kim HJ, Lai A, Nghiemphu P, Xue X, Ellingson BM, Schiff D, Aregawi D, Cha S, Puduvalli VK, Wu J, Yung WK, Young GS, Vredenburgh J, Barboriak D, Abrey LE, Mikkelsen T, Jain R, Paleologos NA, Rn PL, Prados M, Goldin J, Wen PY, Cloughesy T. Apparent diffusion coefficient histogram analysis stratifies progression-free and overall survival in patients with recurrent GBM treated with bevacizumab: a multi-center study. J Neurooncol. 2012;108:491–498.
24. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, Dancey J, Arbuck S, Gwyther S, Mooney M, Rubinstein L, Shankar L, Dodd L, Kaplan R, Lacombe D, Verweij J. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009;45:228–247.
25. Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK, Ohgaki H, Wiestler OD, Kleihues P, Ellison DW. The 2016 World Health Organization classification of tumors of the central nervous system: a summary. Acta Neuropathol. 2016;131:803–820.
26. Wen PY, Macdonald DR, Reardon DA, Cloughesy TF, Sorensen AG, Galanis E, Degroot J, Wick W, Gilbert MR, Lassman AB, Tsien C, Mikkelsen T, Wong ET, Chamberlain MC, Stupp R, Lamborn KR, Vogelbaum MA, van den Bent MJ, Chang SM. Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. J Clin Oncol. 2010;28:1963–1972.
27. Ellingson BM, Cloughesy TF, Lai A, Mischel PS, Nghiemphu PL, Lalezari S, Schmainda KM, Pope WB. Graded functional diffusion map-defined characteristics of apparent diffusion coefficients predict overall survival in recurrent glioblastoma treated with bevacizumab. Neuro Oncol. 2011;13:1151–1161.

Repeatability of Quantitative Diffusion-Weighted Imaging Metrics in Phantoms, Head-and-Neck and Thyroid Cancers: Preliminary Findings

Ramesh Paudyal[1], Amaresha Shridhar Konar[1], Nancy A. Obuchowski[2], Vaios Hatzoglou[3], Thomas L. Chenevert[4], Dariya I. Malyarenko[4], Scott D. Swanson[4], Eve LoCastro[1], Sachin Jambawalikar[5], Michael Z. Liu[5], Lawrence H. Schwartz[5], R. Michael Tuttle[6], Nancy Lee[7], and Amita Shukla-Dave[1,3]

[1]Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY; [2]Department of Quantitative Health Sciences, Cleveland Clinic Foundation, Cleveland, OH; [3]Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, NY; [4]Department of Radiology, University of Michigan, Ann Arbor, MI; [5]Department of Radiology, Columbia University Irving Medical Center, and New York Presbyterian Hospital, New York, NY; [6]Departments of Medicine, and [7]Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, NY

Corresponding Author:
Amita Shukla-Dave, PhD
Departments of Medical Physics and Radiology, Memorial Sloan Kettering Cancer Center, 1275 York Avenue, NY, NY 10065;
E-mail: davea@mskcc.org

Key Words: quantitative Imaging, repeatability, diffusion-weighted imaging, head and neck cancer, thyroid, cancer, within-subject coefficient of variation

Abbreviations: Diffusion-weighted imaging (DWI), head and neck (HN), apparent diffusion coefficient (ADC), isotropic diffusion kurtosis imaging (iDKI), papillary thyroid cancer (PTC), head and neck squamous cell carcinoma (HNSCC), quantitative imaging biomarker (QIB), intra-voxel incoherent motion (IVIM), ceteryl alcohol and behentrimonium (CA-BTAC), repetition time (TR), echo time (TE), number of averages (NA), number of slices (NS), single-shot spin-echo echo planar imaging (SS-SE-EPI), reduced field of view (rFOV), regions of interest (ROIs), confidence interval (CI)

ABSTRACT

The aim of this study was to establish the repeatability measures of quantitative Gaussian and non-Gaussian diffusion metrics using diffusion-weighted imaging (DWI) data from phantoms and patients with head-and-neck and papillary thyroid cancers. The Quantitative Imaging Biomarker Alliance (QIBA) DWI phantom and a novel isotropic diffusion kurtosis imaging phantom were scanned at 3 different sites, on 1.5T and 3T magnetic resonance imaging systems, using standardized multiple b-value DWI acquisition protocol. In the clinical component of this study, a total of 60 multiple b-value DWI data sets were analyzed for test–retest, obtained from 14 patients (9 head-and-neck squamous cell carcinoma and 5 papillary thyroid cancers). Repeatability of quantitative DWI measurements was assessed by within-subject coefficient of variation (wCV%) and Bland–Altman analysis. In isotropic diffusion kurtosis imaging phantom vial with 2% ceteryl alcohol and behentrimonium chloride solution, the mean apparent diffusion ($D_{app} \times 10^{-3}$ mm^2/s) and kurtosis (K_{app}, unitless) coefficient values were 1.02 and 1.68 respectively, capturing in vivo tumor cellularity and tissue microstructure. For the same vial, D_{app} and K_{app} mean wCVs (%) were ≤1.41% and ≤0.43% for 1.5T and 3T across 3 sites. For pretreatment head-and-neck squamous cell carcinoma, apparent diffusion coefficient, D, D*, K, and f mean wCVs (%) were 2.38%, 3.55%, 3.88%, 8.0%, and 9.92%, respectively; wCVs exhibited a higher trend for papillary thyroid cancers. Knowledge of technical precision and bias of quantitative imaging metrics enables investigators to properly design and power clinical trials and better discern between measurement variability versus biological change.

INTRODUCTION

Malignant tumors of the head and neck (HN) region include a diverse group of cancers in the oral cavity, nasopharynx, oropharynx, hypopharynx, larynx, and paranasal sinuses; although salivary and thyroid carcinomas are also located within the HN region, they are typically thought of as separate tumors (1). HN tumors are heterogeneous with complex anatomy ranging between oral cavity to hypopharynx (2, 3). Accurate detection and delineation of tumor extent is critical to optimize treatment planning; patients therefore routinely undergo noninvasive imaging for careful assessment of this complex anatomy by an experienced neuroradiologist (4). Noninvasive magnetic resonance imaging (MRI) has served an important role as a diagnostic test for initial staging and follow-up of tumors in the HN region (5–8).

The quantitative MRI (qMRI) technique, diffusion-weighted imaging (DWI), assesses the Brownian motion of water mole-

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2018.00044

cules at a cellular level (9). Apparent diffusion coefficient (ADC), derived by fitting DWI data to a monoexponential model using ≥2 b-values (ie, diffusion-weighting factor), reflects tumor cellularity (10, 11). Repeatability of ADC has been tested in both phantoms and solid tumors (12-15). In previous studies, ADC has exhibited promise as a quantitative imaging biomarker (QIB) of treatment response in HN cancer (16-20). The use of ADC is helpful in differentiation between malignant and benign solitary thyroid nodules and assessing tumor aggressiveness in papillary thyroid cancer (PTC) (21, 22).

Recent literature reflects interest in acquisition of DWI data using multiple b-values, which allows the measurement of both water diffusion for higher b-values (>200 s/mm^2) and vascular perfusion fraction at lower b-values separately without contrast agent injection (23, 24). Le Bihan et al. developed a biexponential model using multiple b-value DWI data and termed it "intravoxel incoherent motion" (IVIM) (25, 26), which has shown utility for the assessment of treatment response in various cancers, including HN cancer (27, 28). Test–retest studies using IVIM-DWI metrics in normal liver and metastases have a tendency towards better repeatability of measurement of true diffusion coefficient (D), whereas use of perfusion fraction (f) and pseudo-diffusion coefficient (D*) are still exploratory in nature (23, 29).

Underlying biological structures can alter the Gaussian distribution of the water diffusion as assumed in IVIM to be non-Gaussian (NG) in nature (30). This NG behavior has been incorporated in the non-monoexponential diffusion kurtosis imaging (DKI) model which provides the kurtosis coefficient (K) metric, a surrogate QIB of tissue microstructure, in addition to diffusion coefficient (31-33). Lu et al. incorporated the NG diffusion into the IVIM-DWI model (NG IVIM-DWI) and provided estimates for all the aforementioned quantitative imaging metrics (f, D, D*, and K) (34).

QIBs are being used in oncology clinical trials to monitor the effects of treatments, identify subjects likely to benefit from treatment, and as trial endpoints. As compared with other modalities and endpoints, QIBs have the advantage of being noninvasive and requiring little or no subjective interpretation. Furthermore, for disease conditions with multiple treatment options, early detection of nonresponders enables physicians to consult patients about other treatment options earlier, to potentially improve outcomes and limit adverse effects of ineffective treatments.

Before QIBs can be used in clinical trials, their technical performance must be assessed, similarly to how sensitivity and specificity must be established for diagnostic tests (35). Technical performance includes precision, bias, and the property of linearity. Perhaps the most important QIB performance metric is precision, that is, the ability to provide the same, or nearly the same, measurement value on repeated observations (36). Once precision and performance metrics are established, they may be used to formulate a clinical trial's eligibility criteria, to determine the cut-point for defining true change over time, and to compute the sample size required for the trial (37).

There is currently a paucity of repeatability literature for DWI measurements in the clinical setting, particularly for HN cancers and PTC. Hence, it is critical to perform test–retest studies as the fundamental building blocks for QIB discovery and clinical application of these more advanced quantitative imaging methods. The objective of this study was to establish the repeatability measures of quantitative Gaussian and NG diffusion metrics using data from phantoms and from patients with HN cancers and PTC.

MATERIALS AND METHODS

Quantitative DWI Phantom

The quantitative diffusion phantom (High Precision Devices, Inc, Boulder, CO) developed by National Institute of Standards and Technology (NIST)/Radiological Society of North America (RSNA)-Quantitative Imaging Biomarker Alliance (QIBA) consists of 13 vials filled with varying concentrations of polyvinylpyrrolidone (PVP) in aqueous solution (38). The phantom was specifically designed for quantitatively mapping isotropic Gaussian diffusion of water molecules and generating physiologically relevant ADC values. The distribution of PVP concentrations in the phantom is as follows: 0% (vials 1–3), 10% (vials 4–5), 20% (vials 6–7), 30% (vials 8–9), 40% (vials 10–11), and 50% (vials 12–13). The space between the vials within the phantom was filled with an ice-water bath at 0°C to eliminate thermal variability across scanner locations and timepoints in ADC measurements. In this study, we will focus on the measurements obtained from 2 vials, that is, (1) water-only and (2) PVP-20%, as they relate to data from the novel isotropic diffusion kurtosis imaging (iDKI) phantom. Details of the NIST/QIBA DWI phantom have been published previously (38, 39).

The newly developed iDKI phantom used in this study was designed and fabricated by coauthors at the University of Michigan (40). The phantom captures a range of in vivo kurtosis values (K_{app} ranges, 0.4–1.7) (31). Here we report data from 2

Table 1. Summary of Patient Characteristics

Patient	Age (years)	Gender	Primary Cancer
1	63	M	BOT
2	58	M	NPC
3	59	M	Tonsil
4	59	M	Tonsil
5	60	M	BOT
6	68	F	BOT
7	61	F	Hypopharynx
8	75	M	BOT
9	55	M	BOT
10	51	M	PTC
11	44	M	PTC
12	44	M	PTC
13	48	M	PTC
14	44	F	PTC

Abbreviations: BOT, base of tongue; NPC, nasopharyngeal carcinoma; PTC, papillary thyroid cancer.

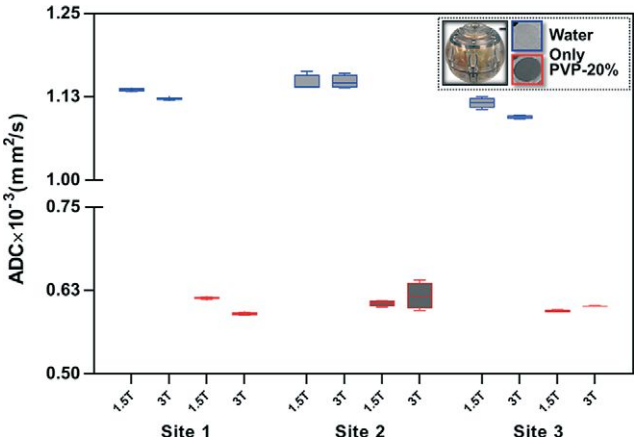


Figure 1. Box-and-whisker plot showing the test–retest mean apparent diffusion coefficient (ADC × 10^{-3} mm^2/s) values obtained from National Institute of Standards and Technology (NIST)/Quantitative Imaging Biomarker Alliance (QIBA) polyvinylpyrrolidone (PVP) diffusion phantom (at 0°C) from the 3 different sites at 1.5T and 3T. The horizontal line inside the box indicates median values. The bottom and top of the boxes indicate 25th and 75th percentiles of the values, respectively. The differences between median values across scanners reflect the differences in gradient designs.

vials in the iDKI phantom: 1 vial containing chemical ceteryl alcohol and behentrimonium (CA-BTAC), a vesicular suspension formed by water solution of 2% CA-BTAC with other (minor) stabilizing ingredients (vial #2 [V2]), and a negative control consisting of a 20% solution of PVP in water (vial #4 [V4],), similar to the vial in NIST/QIBA DWI phantom (41). The iDKI phantom has been detailed in the poster presented at the NCI/Quantitative Imaging Network (QIN) meeting (40), and its full repeatability and long-term stability study is summarized in a research paper by Malyarenko D et al. submitted to this issue of *Tomography*.

The above 2 phantoms were studied to assess the technical performance of the quantitative imaging metrics among the 3 participating sites. There was a need to compare the vials with similar chemical composition for both the standard NIST/QIBA DWI and novel iDKI phantoms to emphasize the differences between the quantitative imaging metrics values for both diffusion and kurtosis coefficients.

Patient Cohort

The institutional review board of Site 1 (Memorial Sloan Kettering Cancer Center [MSKCC]) approved this prospective study for patients with head and neck squamous cell carcinoma (HNSCC) and PTC and was compliant with the Health Insurance Portability and Accountability Act. We obtained written informed consent from all eligible patients. A total of 14 patients were enrolled in the study between December 2016 and August 2017. In total, 30 MRI examinations were performed for these 14 patients, which comprised 60 test–retest MRI data sets. Nine patients with HNSCC were enrolled. All subjects had with metastatic nodes (M/F: 7/2, mean age: 59 years, range = 55–68 years) and underwent standard chemoradiation therapy (dose, 70 Gy). MRI examinations were performed before initiation of the standard chemoradiation treatment (pre-TX) and during treatment (intra-TX weeks 1 and 2) for patients with HNSCC. One patient with pre-TX MRI did not participate in MRI examinations during treatment. Five patients with PTC who underwent surgery (M/F: 4/4, mean age: 47 years, range = 37–61 years) were studied. All patient characteristics are summarized in Table 1.

DWI Data Acquisition

Quantitative DWI Phantom. Diffusion studies were performed using the NIST/QIBA DWI phantom at 0°C on 1.5T and 3T scanners using a 16-channel head coil at all 3 sites (Site 1 [MSKCC], Site 2 [Columbia University Irving Cancer Center; CUMC] and Site 3 [University of Michigan; UMich]). Localizer images were acquired for accurate positioning of the phantom. DWI images were acquired using a single-shot echo planar imaging sequence with 4 b-values (ie, b = 0, 500, 900, 2000 s/mm^2) and the following parameters: repetition time (TR) = 15 000 milliseconds, echo time (TE) = minimum (109–110 milliseconds), number of averages (NA) = 1, acquisition matrix = 128 × 128, field of view (FOV) = 220 mm, number of slices (NS) = 36, slice thickness = 4 mm, all 3 orthogonal directions at both 1.5T and 3.0T scanners. The total acquisition time for the multiple b-value DWI data acquisition was ~2 minutes 30 seconds.

The iDKI phantom, designed and fabricated by Site 3 (UMich), was imaged by all 3 sites at different field strengths of 1.5T

Table 2. Test–Retest Repeatability Measurement of the ADC for NIST/QIBA Phantom

Metrics	Chemical (PVP) Composition	Site 1		Site 2		Site 3	
		1.5T	3T	1.5T	3T	1.5T	3T
ADC × 10^{-3} mm^2/s	0%	1.13 ± 0.008	1.12 ± 0.002	1.14 ± 0.012	1.14 ± 0.01	1.11 ± 0.007	1.09 ± 0.002
	20%	0.61 ± 0.007	0.59 ± 0.005	0.60 ± 0.004	0.61 ± 0.02	0.59 ± 0.003	0.60 ± 0.004
wCV (%)	0%	0.21 (±0.48)	0.15 (±0.34)	1.07 (±2.41)	0.84 (±1.90)	0.67 (±1.48)	0.22 (±0.49)
	20%	0.24 (±0.09)	0.32 (±0.37)	0.71 (±0.85)	3.19 (±3.86)	0.33 (±0.39)	0.10 (±0.11)

wCV data in parentheses are lower and upper 95% confidence intervals.

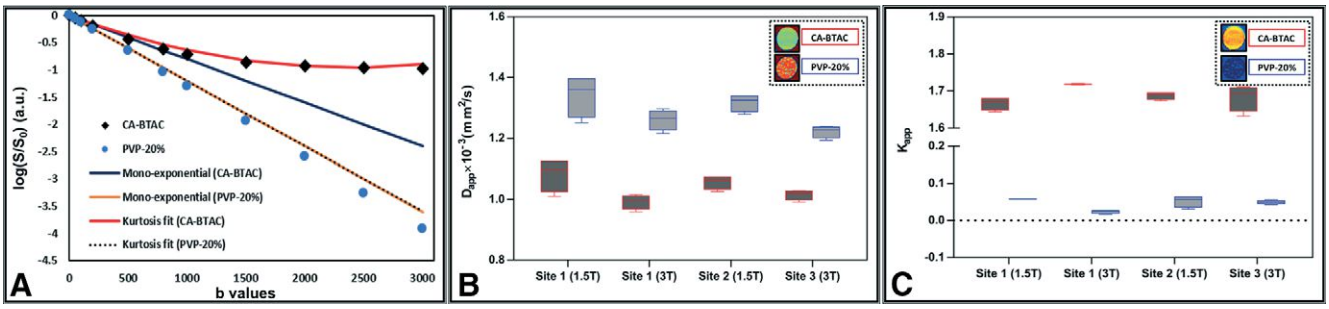


Figure 2. Representative DWI mean signal intensity decay curve vs. b-value obtained from vials of ceteryl alcohol and behentrimonium (CA-BTAC) and PVP-20% in iDKI phantom (scanned at ambient temperature) (A). The diamonds (black) and circles (blue) represent the experimental data, the monoexponential fit is represented by solid blue and yellow lines, and the solid red and dotted black lines are the fitted curves for the diffusion kurtosis model. Box-and-whisker plots show the test–retest for mean values of diffusion coefficient ($D_{app} \times 10^{-3}$ mm^2/s) (B), and kurtosis coefficient (K_{app}, no unit) for the iDKI phantom (C). The horizontal line inside the box indicates median values. The bottom and top of the boxes indicate 25th and 75th percentiles of the values, respectively. The differences between median values across scanners reflect both different scanner room temperatures and system gradient designs.

and/or 3T MRI scanners using a 16-channel head coil at ambient temperature. Localizer images were acquired for accurate positioning of the phantom. DWI images were acquired using a single-shot spin-echo echo planar imaging (SS-SE-EPI) sequence with 11 b-values (ie, b = 0, 50, 100, 200, 500, 800, 1000, 1500, 2000, 2500, 3000 s/mm^2) and parameters on both 1.5T and 3T scanners were kept similar as follows: TR = 10 000 milliseconds, TE = minimum (93–107 milliseconds), NA = 1, matrix = 128 × 128, FOV = 220 mm, NS = 5, slice thickness = 5 mm, all 3 orthogonal directions. The total acquisition time for the multiple b-value DWI data acquisition was ~5 minutes 20 seconds.

Four repeatability experiments for the NIST/QIBA DWI phantom in the study and 2 test–retests for iDKI phantoms with physical repositioning of the phantoms after each diffusion acquisition were performed.

Patient Cohort. MRI examinations were performed at Site 1 for patients with HNSCC on a Philips 3T MRI scanner (Ingenia, Philips Healthcare, The Netherlands) with a neurovascular phased-array coil (maximum number of channels: 20). Standard T1W and T2W imaging was followed by a multiple b-value DWI sequence (28). The DWI data were acquired using a SS-SE-EPI sequence with 10 b-values (ie, b = 0, 20, 50, 80, 200, 300, 500, 800, 1500, 2000 s/mm^2) with TR = 4000 milliseconds, TE = 80 (minimum) milliseconds, NA = 2, matrix = 128 × 128, FOV = 200–240 mm, NS = 8–10, and slice thickness = 5 mm. For patients with HNSCC, DWI was acquired with full field of view as part of the standard clinical imaging protocol. The total acquisition time for the multiple b-value DWI data acquisition was ~5 min. Two multi b-value DWI data sets were acquired at the same MR examination for each patient with HNSCC to test for the repeatability of the derived quantitative imaging metrics. Eighteen multiple b-value DWI data set were acquired at pre-TX (week 0). In addition, 32 multiple b-value DWI data sets were acquired at intra-TX week 1 and week 2 (during chemoradiation therapy). A total of 50 multiple b-value DWI examinations (pre-TX [9 patients], intra-TX week 1 [8 patients], and intra-TX week 2 [8 patients]) were performed (2 MR examinations at each session). As a note, these DWI data sets were acquired with full FOV (phase FOV factor = 1.0).

Table 3. Test–Retest Repeatability Measurement of the D_{app} and K_{app} for Isotropic Diffusion Kurtosis Phantom

Metrics	Chemical Composition	1.5T (Site 1)	3T (Site 1)	1.5T (Site 2)	3T (Site 3)
$D_{app} \times 10^{-3}$ mm^2/s	CA-BTAC	1.06 ± 0.08	0.99 ± 0.029	1.05 ± 0.034	1.01 ± 0.021
	PVP20%	1.32 ± 0.10	1.26 ± 0.041	1.30 ± 0.043	1.22 ±0.024
wCV (%)	CA-BTAC	1.41 (±2.94)	1.18 (±2.32)	1.18 (±2.47)	0.70 (±1.41)
	PVP-20%	1.01 (±2.67)	0.63 (±1.58)	0.31 (±0.79)	0.84 (±1.97)
K_{app}	CA-BTAC	1.66 ± 0.026	1.71 ± 0.001	1.68 ± 0.015	1.68 ± 0.044
	PVP-20%	0.06 ± 0.003	0.03 ± 0.005	0.05 ± 0.023	0.05 ± 0.006
wCV (%)	CA-BTAC	0.35 (±1.17)	0.42 (±1.41)	0.36 (±1.20)	0.43 (±1.43)
	PVP-20%	19.35 (±2.21)	11.12 (±0.57)	7.13 (±0.64)	25.06 (±2.41)

wCV data in parentheses are lower and upper 95% confidence intervals.

MRI examinations were performed at Site 1 for patients with PTC (n = 5) on a 1.5T (n = 2) or 3T (n = 3) scanner (General Electric, Milwaukee, WI), with a neurovascular phased-array coil and consisted of standard T1W and T2W imaging scans followed by multiple b-value DWI data acquisition. This was a feasibility test for the MRI of patients with PTC, which was performed as part of an ongoing research imaging protocol. Data were acquired with reduced field of view (rFOV) DWI technique, using a 2-dimensional spatially selective excitation (42). The acquisition parameters of rFOV DWI scans with the SS-SE-EPI sequence were as follows: 10 b-values (ie, b = 0, 20, 50, 80, 200, 300, 500, 800, 1500, 2000 s/mm^2), TR = 4000 milliseconds, TE = 80 (minimum) milliseconds, NA = 2, matrix = 128 × 64, FOV = 200–240 cm, NS = 8–10, slice thickness = 5 mm, and phase FOV factor = 0.5. The total time for rFOV DWI data acquisition was ~5 min.

Repeatability measures were tested on the multiple b-value DWI data sets obtained from patients with HNSCC at pre-TX, and during intra-TX weeks 1 and 2 of standard chemoradiation therapy. Pretreatment DWI repeatability data were obtained for patients with PTC who underwent surgery.

DWI Data Analysis

All DWI data postprocessing and quantitative metrics map generation, detailed below, were performed using in-house–developed software entitled MRI-QAMPER (MRI Quantitative Analysis of Multi-Parametric Evaluation Routines). The MRI-QAMPER package includes the algorithm routines for DWI data analyses (ADC, diffusion kurtosis, IVIM, and NG-IVIM), implemented in MATLAB (The MathWorks, Natick, MA). The MRI-QAMPER tool is approved by National Cancer Institute/Quantitative Imaging Network (QIN) with pre-benchmark status, which facilitates its use by other QIN site colleagues for analysis of multiple b-value DWI data.

For NIST/QIBA DWI phantom data analysis, 3 distinct circular regions of interest (ROIs) were manually placed (9 mm in diameter) on the selected vials, with water only and PVP-20%, in ADC maps avoiding boundaries; the mean pixel value across the ROIs in each vial was used to measure repeatability.

For iDKI phantom data analysis, 2 distinct circular ROIs (12 mm in diameter, single-plane) were placed on vials with CA-BTAC solution and PVP-20% in the phantom images; the mean pixel value across the ROIs in each vial was used for the test–retest study. To guarantee model convergence, a bmax constraint value for fitting the kurtosis expression in the CA-BTAC phantom vial was set to 1500 s/mm^2 (bmax × D_{app} × K_{app} <3) (43).

For DWI patient data, ROIs were manually delineated on the DWI images (b = 0 s/mm^2) on the metastatic neck node in HNSCC, normal thyroid gland, and PTC. ROIs were placed on thyroid glands avoiding obvious cystic, hemorrhagic, or calcified portions, whereas for normal thyroid tissue, ROIs were placed on the selected contralateral side to the PTC. ROIs were contoured by an experienced neuroradiologist based on the clinical information and T1W/T2W images using ImageJ (44).

Multiple b-value DWI data sets were analyzed using the following models:

1. Mono-exponential (ADC): All b-value DWI signal intensity data obtained from each voxel in the ROI were fitted to a

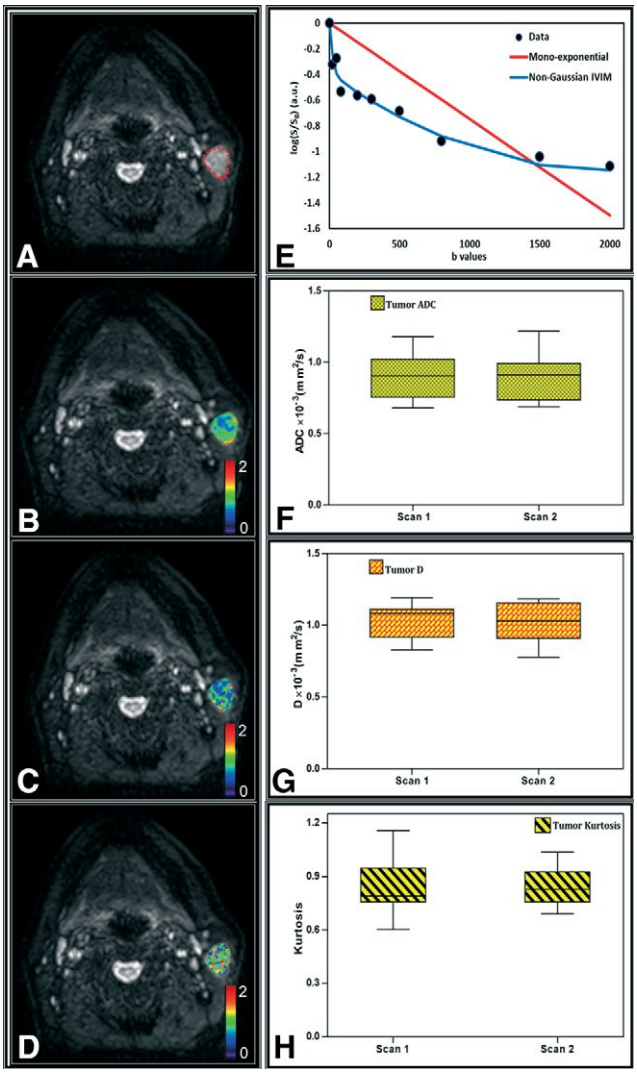


Figure 3. Representative intra-TX week 1 magnetic images (MR) images of a patient with head and neck squamous cell carcinoma (HNSCC) (76 years, male). Diffusion-weighted (b = 0 s/mm^2) image (A), apparent diffusion coefficient (ADC × 10^{-3} mm^2/s) (B), diffusion coefficient (C) (D × 10^{-3} mm^2/s), and kurtosis metric maps overlaid on DWI (b = 0 s/mm^2) image (D). Representative plot of the logarithm of signal intensity vs. b-values (E). The circle (black) represents the experimental data, and the solid lines are the fitted curves with the monoexponential (red) and extended non-Gaussian (NG) intravoxel incoherent motion (NG-IVIM) model (blue). Box-and-whisker plot shows the mean value of (F) apparent diffusion coefficient (ADC × 10^{-3} mm^2/s) and the NG-IVIM model derived metrics: (G) diffusion coefficient (D × 10^{-3} mm^2/s and (H) kurtosis coefficient ("K"). The bottom and top of the boxes indicate 25th and 75th percentiles of the values, respectively. The horizontal line inside the box indicates median values. Note: Data were acquired using standard full FOV DWI sequence.

monoexponential model to calculate ADC (mm^2/s) as follows (45):

$$S(b) = S_0 e^{-bADC} \quad (1)$$

where S(b) and S_0 are the signal intensities with and without diffusion weighting, and the quantity b is the diffusion-weighting factor (s/mm^2).

2. DKI: The signal intensity versus b-value DWI data were fitted to non-monoexponential diffusion kurtosis imaging model (DKI) of the following form (43):

$$S(b) = S_0\left[e^{-bD_{app}+\frac{1}{6}K_{app}b^2D^2}\right] \quad (2)$$

where D_{app} is the ADC (mm^2/s) and K_{app} (no unit) is a dimensionless apparent kurtosis coefficient. D_{app} and K_{app} are associated with the NG behavior of a signal in tissue. As a note, $K_{app} = 0$ is equivalent to equation (1).

3. NG-IVIM: The signal intensity versus b-value DWI signal were fitted to biexponential NG-IVIM DWI model as follows (34, 46):

$$S(b) = S_0\left[fe^{-bD^*} + (1-f)e^{-bD+\frac{1}{6}Kb^2D^2}\right] \quad (3)$$

Where D is the diffusion coefficient (mm^2/s), perfusion fraction (f), and D* is the pseudo-diffusion coefficient (mm^2/s), and K is the kurtosis coefficient.

The NIST/QIBA DWI phantom was analyzed using monoexponential diffusion model equation (1), the iDKI phantom using DKI model [equation (2)], and HNSCC (tumor), and PTC (tumor and normal) using DKI model [equation (2)] and extended NG-IVIM model [equation (3)]. Mean metric values of ADC, DKI-derived metrics (D_{app} and K_{app}), and NG-IVIM-derived metrics (D, D*, f, and K) calculated from each ROI were compared between repeated measurements.

Statistical Analysis

Technical precision of QIBs was evaluated based on the framework proposed by the RSNA/QIBA (https://www.rsna.org/uploadedFiles/RSNA/Content/Science_and_Education/QIBA/QIBA_Process_05Jan2015.pdf). The within-subject coefficient of variation (wCV, %) was used as the measure of precision; it was estimated from the phantom and clinical data as follows (22, 47-49):

$$wCV\,(\%) = \frac{\sigma_w}{\mu} \times 100 \quad (4)$$

where σ_w is the within-subject standard deviation and μ is the mean. A 95% confidence interval (CI) for the wCV was constructed using χ^2 as the pivotal statistic as follows:

$$CI\,(95\,\%) = \sqrt{\frac{N \times wCV^2}{\chi^2_{N,\alpha}}} \quad (5)$$

where N is the number of patients, each having 2 replicate observations and $\chi^2_{N,\alpha}$ is the αth percentile of the chi-square distribution with N degrees of freedom. For the lower bound, α is 0.975, and for the upper bound, α is 0.025. Bland–Altman plots were constructed to measure the repeatability of the quantitative imaging metrics.

Statistical analysis for the data was conducted in R (50) and MATLAB (The MathWorks, Inc., Natick, MA).

RESULTS

Quantitative DWI Phantom

Mean ADC values obtained from the NIST/QIBA DWI phantom (scanned at 0°C) at all 3 different sites on 1.5T and 3T MRI scanners are displayed in a box-and-whisker plot (Figure 1). ADC values are reported for 2 vials only (water-only and PVP-20%). The mean wCV (%) for vial with water-only were ≤1.07% and ≤0.84% and that for vial with PVP-20% were ≤0.71% and ≤3.19% at 1.5T and 3T MRI across the 3 sites, respectively. Results of ADC wCV and 95% CIs are summarized in Table 2.

Figure 2A shows the representative plot of the DWI logarithmic signal intensity versus b-value, fitted by both monoexponential and DKI models obtained from the iDKI phantom ROI for the vials with CA-BTAC (V2) and PVP-20% (V4). The box-and-whisker plots show the mean values of $D_{app} \times 10^{-3}$ mm^2/s (Figure 2B) and K_{app} (no unit) (Figure 2C) obtained from V2 (captures both in vivo tumor cellularity and tissue microstructure) and V4 (captures in vivo tumor cellularity but negative control for kurtosis).

The wCV (%) mean values of D_{app} and K_{app} for V2 were ≤1.41% and ≤0.43% on both 1.5T and 3T MRI. The wCV (%) mean values of D_{app} and K_{app} for V4 were ≤1.01% and ≤25.06% respectively, on both 1.5T and 3T MRI. Table 3 summarizes the

Table 4. Test–Retest Repeatability Measurement of Diffusion Kurtosis Model-Derived Metrics for Patients With HNSCC

Treatment	Measurement	D_{app}	K_{app}
Pre-TX	Mean	$(1.54 \pm 0.02) \times 10^{-3}$ mm^2/s	(0.94 ± 0.01)
	wCV (%)	5.62 (3.87, 10.30)	5.18 (3.59, 9.47)
Intra-TX Week 1	Mean	$(1.56 \pm 0.02) \times 10^{-3}$ mm^2/s	0.96 (±0.01)
	wCV (%)	2.99 (2.10, 5.72)	8.12 (3.50, 15.56)
Intra-TX Week 2	Mean	$(1.68 \pm 0.06) \times 10^{-3}$ mm^2/s	(0.85 ± 0.01)
	wCV (%)	4.29 (2.90, 8.22)	6.01 (4.06, 11.51)

wCV data in parentheses are lower and upper 95% confidence intervals.

Table 5. Test–Retest Repeatability Measurement of the ADC- and NG-IVIM DWI-Derived Metrics for Patients With HNSCC

Treatment	Measurement	ADC	D	D*	K	f
Pre-TX	Mean	$(0.90 \pm 0.04) \times 10^{-3}$ mm^2/s	$(1.03 \pm 0.07) \times 10^{-3}$ mm^2/s	$(2.51 \pm 0.19) \times 10^{-3}$ mm^2/s	(0.84 ± 0.13)	(0 .19 ± 0.04)
	wCV (%)	2.38 (1.67, 4.34)	3.55 (2.44, 6.48)	3.88 (2.67, 7.10)	8.00 (5.57, 14.61)	9.92 (6.8, 18.12)
Intra-TX Week 1	Mean	$(0.93 \pm 0.02) \times 10^{-3}$ mm^2/s	$(1.09 \pm 0.07) \times 10^{-3}$ mm^2/s	$(2.46 \pm 0.11) \times 10^{-3}$ mm^2/s	(0.87 ± 0.08)	(0.18 ± 0.02)
	wCV (%)	0.86 (0.58, 1.66)	3.46 (2.39, 6.63)	2.24 (2.62, 4.28)	4.74 (5.40, 9.09)	9.92 (67.0, 10.96)
Intra-TX Week 2	Mean	$(0.96 \pm 0.04) \times 10^{-3}$ mm^2/s	$(1.16 \pm 0.08) \times 10^{-3}$ mm^2/s	$(2.47 \pm 0.19) \times 10^{-3}$ mm^2/s	(0.86 ± 0.14)	(0.19 ± 0.04)
	wCV (%)	1.18 (0.79, 2.26)	5.57 (3.76, 10.67)	1.20 (0.81, 2.28)	8.36 (5.64, 16.01)	3.01 (2.02, 5.74)

wCV data in parentheses are lower and upper 95% confidence intervals.

D_{app} and K_{app} mean wCV and 95% CIs values for vials with CA-BTAC and PVP-20%. The absolute $K_{app} < 0.05$ value observed for ROI in vial with PVP-20% samples indicates minor bias of the NG model for this monoexponential material.

Patient Cohort. The pre-TX tumor volume (mean ± SD) in patients with HNSSC and PTC were 9.13 ± 6.22 cm^3 and 0.35 ± 0.39 cm^3, respectively.

Figure 3, A–D shows a representative DWI (b = 0 s/mm^2) image, ADC × 10^{-3} mm^2/s, D × 10^{-3} mm^2/s, and K metric maps for a patient with HNSCC. Figure 3E depicts a representative logarithmic DWI signal as a function of the b-value obtained from the metastatic node of the HNSCC patient. The DWI signal was fitted to the monoexponential and NG IVIM model. Figure 3, F–H also displays the box-and-whisker plots for pre-TX test–retest mean values of the same quantitative imaging metrics detailed above.

The wCV (%) mean values of D_{app} and K_{app} at Pre-TX were 5.62% and 5.18%, respectively. Table 4 summarizes the mean wCV (%) and 95% CIs for D_{app} and K_{app} at pre-TX and intra-TX weeks in patients with HNSCC.

The mean wCV (%) values for pre-TX ADC, D, D*, K, and f were 2.38%, 3.55%, 3.88%, 8.0%, and 9.92%, respectively. Table 5 summarizes mean wCV (%) and 95% CIs for ADC- and NG-IVIM-derived metrics (D, D*, K, and f) at pre-TX and intra-TX weeks in patients with HNSCC.

Bland–Altman plots are shown for selective quantitative imaging metrics, ADC, D, and K, obtained from the pre-TX neck nodal metastases of patients with HNSCC (Figure 4). In each panel, the differences in mean values of ADC, D, and K were plotted between the repeated MRI examinations against the combined mean values of ADC, D, and K.

The results from patients with PTC are part of ongoing feasibility testing in the research setting for thyroid MRI imaging using rFOV multiple b-value DWI. Figure 5, A–D displays a representative DWI (b = 0 s/mm^2) image, ADC × 10^{-3} mm^2/s, D × 10^{-3} mm^2/s, and K metric maps for a patient with PTC. Figure 5E shows a representative logarithmic DWI signal as a function of the b-value obtained from the normal thyroid tissue and tumor of the patient with PTC.

The wCV (%) mean values of D_{app} and K_{app} for normal tissue were 12.87% and 17.46%, respectively, whereas these metric values in tumor tissue were 22.42% and 25.94% in patients with PTC. Table 6 summarizes mean D_{app} and K_{app} wCV (%) and 95% CIs for normal and tumor region in patients with PTC.

ADC mean wCV (%) were 11.86% and 10.04%, respectively, for tumor and normal thyroid tissue ROIs. The wCV (%) for NG IVIM-derived metrics (D, D*, K, and f) from tumors were 14.98%, 4.31%, 11.09%, and 13.31%, respectively. Preliminary mean values for ADC, D, D*, K, and f are summarized in Table 7 for normal and tumor tissue in patients with PTC.

Bland–Altman plots are shown for ADC, D, and K, obtained from normal and tumor regions in the PTC patients (Figure 6).

DISCUSSION

In this preliminary study, we measured the repeatability of the quantitative diffusion imaging metrics for Gaussian and NG models using 2 phantoms (the temperature-controlled NIST/QIBA DWI phantom and a novel iDKI phantom at ambient

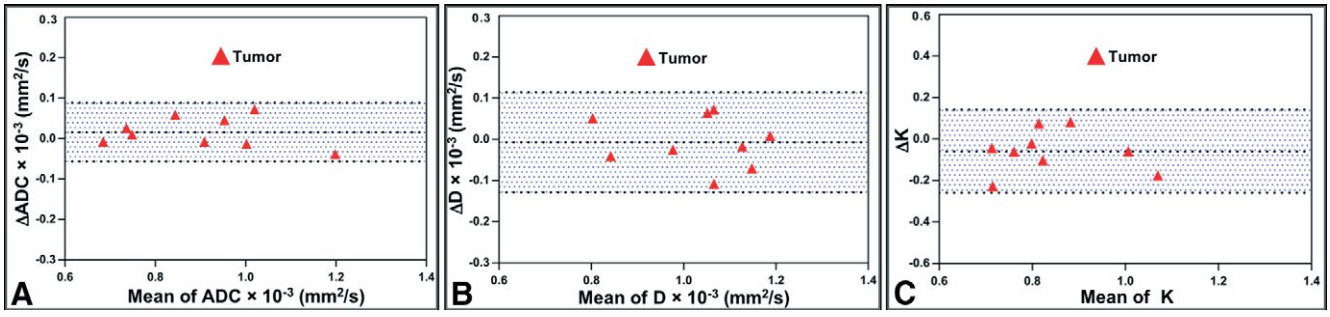


Figure 4. Bland–Altman plots of apparent diffusion coefficient (ADC × 10^{-3} mm^2/s) (A), diffusion coefficient (D × 10^{-3} mm^2/s) (B), and kurtosis coefficient (K) obtained from the metastatic neck node in patients with HNSCC on pre-treatment (C). The solid lines correspond to the mean differences between 2 estimates and the dashed lines show the 95% limits of agreement. Note: Δ represent the change in mean difference between 2 scans.

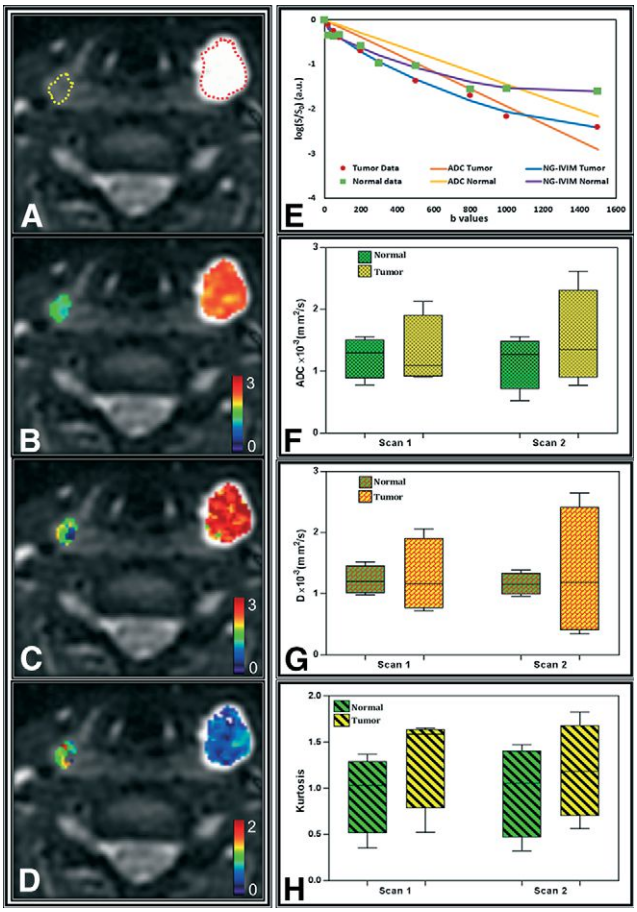


Figure 5. Representative MR images of a patient with papillary thyroid cancer (PTC) (76 years, male). Diffusion-weighted (b = 0 s/mm^2) image (A), apparent diffusion coefficient (ADC × 10^{-3} mm^2/s) (B), diffusion coefficient (D × 10^{-3} mm^2/s) (C), kurtosis metric maps overlaid on DW (b = 0 s/mm^2) images (D), representative plot of the logarithm of DWI signal intensity vs. b values (E). The squares (green) and circles (red) represent the experimental data in normal and tumor; the solid lines are the fitted curves with the monoexponential (yellow and orange) and NG intravoxel incoherent motion (NG-IVIM) (purple and blue). Box-and-whisker plot shows the mean value in normal tissue and in tumor for (F) apparent diffusion coefficient (ADC × 10^{-3} mm^2/s) and the NG-IVIM model-derived metrics: diffusion coefficient (D × 10^{-3} mm^2/s) (G) and kurtosis coefficient (K, unitless) (H). The horizontal line inside the box indicates median values. The bottom and top of the boxes indicate 25th and 75th percentiles of the values, respectively. Note: The DWI images were acquired with reduced FOV DWI sequence.

Table 6. Test–Retest Repeatability Measurement of Diffusion Kurtosis Model-Derived Metrics for Patients With PTC

Treatment	Measurement	D_{app}	K_{app}
Normal	Mean	(2.51 ± 0.32) × 10^{-3} mm^2/s	(1.08 ± 0.19)
	wCV (%)	12.87 (7.71, 37.00)	17.46 (10.46, 50.19)
Tumor	Mean	(2.52 ± 0.57) × 10^{-3} mm^2/s	(1.14 ± 0.29)
	wCV (%)	22.42 (13.43, 64.46)	25.94 (15.54, 74.57)

wCV data in parentheses are lower and upper 95% confidence intervals.

temperature) in a multisite setting, as well as for a small cohort of patients with HNSCC and PTC using the DKI model and the extended NG IVIM model.

For the NIST/QIBA DWI phantom, repeatability of mean ADC wCV (%) and 95% CIs values was excellent for the studied phantom vials with water-only and PVP-20% (≤3.19% and ≤4.0% respectively), for all 3 sites. The results reported herein are comparable to results from similar test–retest repeatability studies (51, 52). The D_{app} and K_{app} wCVs (%) and 95% CIs from all 3 sites were comparable at both 1.5T and 3T MRI. The novel iDKI phantom has been designed and fabricated with the purpose to better understand the performance of the quantitative diffusion metric kurtosis (K) as a surrogate of tissue microstructure and the stability of K over time. Performing appropriate phantom testing is a prerequisite for the QIB pipeline for clinical trials that use quantitative NG diffusion imaging metrics (37). Our phantom results confirmed adequate baseline technical performance of the MRI scanner systems and multiple b-value DWI protocols used for the quantitative DWI studies for patients with HNSCC and PTC.

There is currently paucity of repeatability measures for quantitative Gaussian and NG DWI in cancers of the HN region, despite availability of ADC test–retest data for organs such as brain (wCV = 3.97%), liver (wCV = 9.38%), and prostate (wCV = 16.97%) (13-15, 53). Only a few studies have reported test–retest data for IVIM in organs such as liver (23).

The preliminary findings for test–retest data in HNSCC showed that the mean wCV (%) for ADC-derived metric, DKI-derived metric (D_{app}), and NG IVIM-derived metrics (D and D*) were ≤6% for pre-TX, intra-TX weeks 1 and 2. For f, K_{app}, and K, the mean wCV(%) were ≤10%. Both ADC and D, quantitative imaging metrics, are surrogate biomarkers of tumor cellularity, while f and D* are still exploratory in nature (23, 29, 54). There is keen interest in furthering description of tissue microstructure using the quantitative imaging metric K (31, 34, 55, 56). The uncertainties from clinical HNSCC data slightly exceeded baseline repeatability achieved for phantoms due to additional patient-related variability.

Our clinical repeatability measurements for normal thyroid tissue and PTC are preliminary findings. Lu et al. reported that the ADC mean wCV (%) for the normal thyroid tissue in healthy volunteers is ≤10% using rFOV DWI at 3T (42). The present study found consistent results for the normal thyroid tissue (ADC mean wCV (%) = ≤10%) acquired with rFOV DWI at 1.5T and 3T MRI

Table 7. Test–Retest Repeatability Measurement of the ADC and NG-IVIM DWI-Derived Metrics for Patients With PTC

ROI	Measurement	ADC	D	D*	K	f
Normal	Mean	(1.23 ± 0.24) ×10^{-3} mm^2/s	(1.16 ± 0.58) ×10^{-3} mm^2/s	(2.89 ± 0.43) ×10^{-3} mm^2/s	(0.96 ± 0.36)	(0.26 ± 0.08)
	wCV (%)	10.05 (6.02, 28.90)	25.80 (15.46, 74.17)	7.63 (4.57, 21.94)	19.25 (11.53, 55.34)	16.54 (9.91, 47.57)
Tumor	Mean	(1.31 ± 0.30) ×10^{-3} mm^2/s	(1.54 ± 0.45) ×10^{-3} mm^2/s	(2.87 ± 0.24) ×10^{-3} mm^2/s	(1.21 ± 0.26)	(0.22 ± 0.06)
	wCV (%)	11.86 (7.11, 34.10)	14.98 (8.97, 43.06)	4.31 (2.58, 12.38)	11.09 (6.65, 31.89)	13.31 (7.97, 38.26)

wCV data in parentheses are lower and upper 95% confidence intervals.

(42). Kim et al. reported that mean ADC values obtained at 2 different MRI field strengths (1.5T and 3T) were not significantly different (19). A relatively high wCV was observed for DKI- and NG IVIM-derived metrics that may likely be related to the limited sample size and the biology of the tumors in the thyroid gland.

Establishing the technical performance of a QIB allows us to better understand a patient's measurement at a single time point, especially the changes in measurements over time, by constructing a CI for the true value or the true change. For example, suppose we measure ADC of 1.22 × 10^{-3} mm^2/s for PTC. If we know from our technical performance studies that the measurements are made with negligible bias and precision of wCV (%) = 11.86%, then a 95% CI for the patient's true ADC value is (1.22 ± 1.96 × (0.1186 × 1.22) × 10^{-3} mm^2/s or (0.94 to 1.50) × 10^{-3} mm^2/s. The CI helps differentiate between the true change of the parameter value versus the measurement uncertainty. Now suppose that on a second visit, the patient's tumor has an ADC of 1.31 × 10^{-3} mm^2/s. Has the ADC value truly increased or is the observed change attributable to measurement error? The 95% CI for the true change is $[(1.31-1.22) \pm 1.96 \times \sqrt{(0.1186 \times 1.22)^2 + (0.1186 \times 1.31)^2}] \times 10^{-3}$ mm^2/s or [−0.32, 0.50] × 10^{-3} mm^2/s. Thus, given the known imprecision in the ADC measurements, we cannot conclude that a true change has occurred with 95% confidence.

Once the technical performance of a QIB is known, investigators are better able to design their clinical trials effectively. For example, a measured change in a patient's quantitative imaging metrics (eg, D or K) must exceed 10% (ie, 2.77 × wCV) to be 95% confident that a true change has occurred (37). Thus, in a drug trial using changes in D or K (QIBs) as a measure of therapeutic effect, a ≥10% cut-point should be used to define whether a treatment effect should be used to define treatment success and determine when a change in treatment is warranted. Similarly, in planning a clinical trial where D or K values will be compared across treatment arms, the imprecision in the QIB values affects trial sample size by increasing it relative to its magnitude and the magnitude of the between-subject variability.

There are a few known limitations to this study. This is the first feasibility test–retest study of Gaussian- and NG diffusion-derived metrics from multisite phantom and single-site clinical data testing. A larger cohort of patients (>30) is necessary to confirm statistical significance of the preliminary findings (57). Susceptibility artifacts caused by SS-SE-EPI, voluntary and involuntary bulk motion, are still an issue in the HN region, limiting repeatability. Thus, rFOV DWI for incremental improvement may be an option, exciting only a limited FOV and not surrounding regions that potentially cause interference (42). For the test–retest data set, technically the patients should be scanned, removed from the scanner for a few minutes and scanned again, referred to as a "coffee break" study. Here, the patients were repositioned between scans on the table but not removed from the scanner ("coffee break") owing to practical reasons relating to patient comfort and workflow at the MR scanner. The results reported here provide insights into what is needed and must be paid attention to in test–retest studies in clinical oncology trials. For example, the test–retest studies for ADC in brain tumors derived from monoexponential mod-

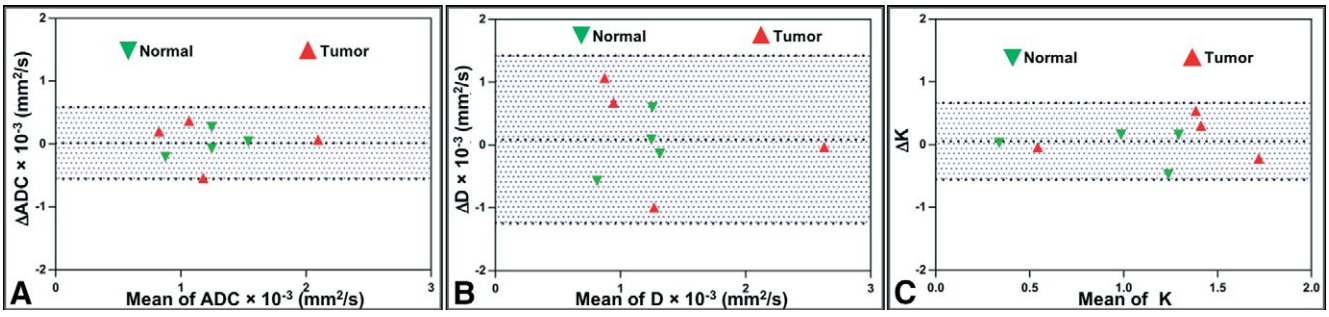


Figure 6. Bland–Altman plots of apparent diffusion coefficient (ADC × 10^{-3} mm^2/s) (A), diffusion coefficient (D × 10^{-3} mm^2/s) (B), and kurtosis coefficient (K) obtained from the papillary thyroid cancer (C). The solid lines correspond to the mean differences between 2 estimates, and the dashed lines show the 95% limits of agreement. Note: Δ represent the change in mean difference between 2 scans.

eling of DWI data reports a wCV of 3.97% (53, 58, 59). A smaller wCV value (<5%) indicates less variation in repeatability measurements.

CONCLUSION

In conclusion, we have shown repeatability of measurements for quantitative Gaussian and NG diffusion imaging metrics using multiple b-value acquisitions for NIST/QIBA DWI phantom and iDKI phantom, across multisite MRI systems, and used in HNSCC and PTC clinical trials. The preliminary results for the repeatability measurement of NG IVIM-derived metrics in HNSCC and PTC show promise and need additional validation with a larger subject cohort. In short, the precision of QIBs must be established for oncology clinical trials to noninvasively monitor the effects of treatment, to identify subjects likely to benefit from treatment and define trial endpoints.

ACKNOWLEDGMENTS

We acknowledge funding from NIH U01 CA211205 (ASD, LHS) and NIH/NCI Cancer Center Support Grant P30 CA008748, and U01CA166104 (TLC, DIM, SDS).

Disclosure: S.D. Swanson, D.I. Malyarenko, and T.L. Chenevert are coinventors on intellectual property assigned to and managed by the University of Michigan for the technology underlying the manufacturing of the quantitative isotropic diffusion kurtosis imaging (iDKI) phantoms.

Conflict of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. Vogel DWT, Thoeny HC. Cross-sectional imaging in cancers of the head and neck: how we review and report. Cancer Imaging. 2016;16:20.
2. King AD, Vlantis AC, Tsang RK, Gary TM, Au AK, Chan CY, Kok SY, Kwok WT, Lui HK, Ahuja AT. Magnetic resonance imaging for the detection of nasopharyngeal carcinoma. AJNR Am J Neuroradiol. 2006;27:1288–1291.
3. Klussmann JP. Head and neck cancer—new insights into a heterogeneous disease. Oncol Res Treat. 2017;40:318–319.
4. Vairaktaris E, Yapijakis C, Psyrri A, Spyridonidou S, Yannopoulos A, Lazaris A, Vassiliou S, Ferekidis E, Vylliotis A, Nkenke E, Patsouris E. Loss of tumour suppressor p16 expression in initial stages of oral oncogenesis. Anticancer Res. 2007;27:979–984.
5. King AD, Mo FK, Yu KH, Yeung DK, Zhou H, Bhatia KS, Tse GM, Vlantis AC, Wong JK, Ahuja AT. Squamous cell carcinoma of the head and neck: diffusion-weighted MR imaging for prediction and monitoring of treatment response. Eur Radiol. 2010;20:2213–2220.
6. Lenz M, Greess H, Baum U, Dobritz M, Kersting-Sommerhoff B. Oropharynx, oral cavity, floor of the mouth: CT and MRI. Eur J Radiol. 2000;33:203–215.
7. Zima AJ, Wesolowski JR, Ibrahim M, Lassig AA, Lassig J, Mukherji SK. Magnetic resonance imaging of oropharyngeal cancer. Top Magn Reson Imaging. 2007;18:237–242.
8. Noda Y, Kanematsu M, Goshima S, Kondo H, Watanabe H, Kawada H, Bae KT. MRI of the thyroid for differential diagnosis of benign thyroid nodules and papillary carcinomas. Am J Roentgenol. 2015;204:W332–W335.
9. Le Bihan D. From Brownian motion to mind imaging: diffusion MRI. Bull Acad Natl Med. 2006;190:1605–1627; discussion 1627. [Article in French
10. Chenevert TL, McKeever PE, Ross BD. Monitoring early response of experimental brain tumors to therapy using diffusion magnetic resonance imaging. Clin Cancer Res. 1997;3:1457–1466.
11. Partridge SC, Nissan N, Rahbar H, Kitsch AE, Sigmund EE. Diffusion-weighted breast MRI: clinical applications and emerging techniques. J Magn Reson Imaging. 2017;45:337–355.
12. Boss M, Chenevert T, Waterton J, Morris D, Ragheb H, Jackson A, de Souza N, Collins DJ, van Beers BV, Garteiser P, Doblas S, Persigehl T, Hedderich D, Martin A, Mukherjee P, Keenan K, Russek S, Jackson E, Zahlmann G. Thermally-stabilized isotropic diffusion phantom for multisite assessment of apparent diffusion coefficient reproducibilty. Med Phys. 2014;41:464.
13. Braithwaite AC, Dale BM, Boll DT, Merkle EM. Short- and midterm reproducibility of apparent diffusion coefficient measurements at 3.0-T diffusion-weighted imaging of the abdomen. Radiology. 2009;250:459–465.
14. Gibbs P, Pickles MD, Turnbull LW. Repeatability of echo-planar-based diffusion measurements of the human prostate at 3 T. Magn Reson Imaging. 2007;25:1423–1429.
15. Jambor I, Merisaari H, Aronen HJ, Jarvinen J, Saunavaara J, Kauko T, Borra R, Pesola M. Optimization of b-value distribution for biexponential diffusion-weighted MR imaging of normal prostate. J Magn Reson Imaging. 2014;39:1213–1222.
16. Vandecaveye V, De Keyzer F, Dirix P, Lambrecht M, Nuyts S, Hermans R. Applications of diffusion-weighted magnetic resonance imaging in head and neck squamous cell carcinoma. Neuroradiology. 2010;52:773–784.
17. Thoeny HC, Ross BD. Predicting and monitoring cancer treatment response with diffusion-weighted MRI. J Magn Reson Imaging. 2010;32:2–16.
18. Dirix P, Vandecaveye V, De Keyzer F, Op de Beeck K, Poorten VV, Delaere P, Verbeken E, Hermans R, Nuyts S. Diffusion-weighted MRI for nodal staging of head and neck squamous cell carcinoma: impact on radiotherapy planning. Int J Radiat Oncol Biol Phys. 2010;76:761–766.
19. Kim S, Loevner L, Quon H, Sherman E, Weinstein G, Kilger A, Poptani H. Diffusion-weighted magnetic resonance imaging for predicting and detecting early response to chemoradiation therapy of squamous cell carcinomas of the head and neck. Clin Cancer Res. 2009;15:986–994.
20. Galban CJ, Mukherji SK, Chenevert TL, Meyer CR, Hamstra DA, Bland PH, Johnson TD, Moffat BA, Rehemtulla A, Eisbruch A, Ross BD. A feasibility study of parametric response map analysis of diffusion-weighted magnetic resonance imaging scans of head and neck cancer patients for providing early detection of therapeutic efficacy. Transl Oncol. 2009;2:184–190.
21. Lu Y, Moreira AL, Hatzoglou V, Stambuk HE, Gonen M, Mazaheri Y, Deasy JO, Shaha AR, Tuttle RM, Shukla-Dave A. Using diffusion-weighted MRI to predict aggressive histological features in papillary thyroid carcinoma: a novel tool for preoperative risk stratification in thyroid cancer. Thyroid. 2015;25:672–680.
22. Razek AA, Sadek AG, Kombar OR, Elmahdy TE, Nada N. Role of apparent diffusion coefficient values in differentiation between malignant and benign solitary thyroid nodules. Am J Neuroradiol. 2008;29:563–568.
23. Andreou A, Koh DM, Collins DJ, Blackledge M, Wallace T, Leach MO, Orton MR. Measurement reproducibility of perfusion fraction and pseudodiffusion coefficient derived by intravoxel incoherent motion diffusion-weighted MR imaging in normal liver and metastases. Eur Radiol. 2013;23:428–434.
24. Hauser T, Essig M, Jensen A, Laun FB, Munter M, Maier-Hein KH, Stieltjes B. Prediction of treatment response in head and neck carcinomas using IVIM-DWI: evaluation of lymph node metastasis. Eur J Radiol. 2014;83:783–787.
25. Le Bihan D, Breton E, Lallemand D, Aubin ML, Vignaud J, Laval-Jeantet M. Separation of diffusion and perfusion in intravoxel incoherent motion MR imaging. Radiology. 1988;168:497–505.
26. Le Bihan D, Breton E, Lallemand D, Grenier P, Cabanis E, Laval-Jeantet M. MR imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders. Radiology. 1986;161:401–407.
27. Ding Y, Hazle JD, Mohamed AS, Frank SJ, Hobbs BP, Colen RR, Gunn GB, Wang J, Kalpathy-Cramer J, Garden AS, Lai SY, Rosenthal DI, Fuller CD. Intravoxel incoherent motion imaging kinetics during chemoradiotherapy for human papillomavirus-associated squamous cell carcinoma of the oropharynx: preliminary results from a prospective pilot study. NMR Biomed. 2015;28:1645–1654.
28. Paudyal R, OH JH, Riaz N, Venigalla P, Li J, Hatzoglou V, Leeman J, Nunez DA, Lu Y, Deasy JO, Lee N, Shukla-Dave A. Intravoxel incoherent motion diffusion-weighted MRI during chemoradiation therapy to characterize and monitor treatment response in human papillomavirus head and neck squamous cell carcinoma. J Magn Reson Imaging. 2017;45:1013–1023.
29. Dyvornea H, Jajamovicha G, Kakitea S, Kuehn B, Taoulia B. Intravoxel incoherent motion diffusion imaging of the liver:Optimal b-value subsampling and impact on parameterprecision and reproducibility. Eur J Radiol. 2014;83:2109–2113.
30. Le Bihan D. Apparent diffusion coefficient and beyond: what diffusion MR imaging can tell us about tissue structure. Radiology. 2013;268:318–322.
31. Jansen JF, Stambuk HE, Koutcher JA, Shukla-Dave A. Non-Gaussian analysis of diffusion-weighted MR imaging in head and neck squamous cell carcinoma: a feasibility study. AJNR Am J Neuroradiol. 2010;31:741–748.
32. Jensen JH, Helpern JA. MRI quantification of non-Gaussian water diffusion by kurtosis analysis. NMR Biomed. 2010;23:698–710.

33. Yuan J, Yeung DKW, Mok GSP, Bhatia KS, Wang YXJ, Ahuja AT, King AD. Non-gaussian analysis of diffusion weighted imaging in head and neck at 3T: a pilot study in patients with nasopharyngeal carcinoma. PLoS One. 2014;9:e87024.
34. Lu Y, Jansen JF, Mazaheri Y, Stambuk HE, Koutcher JA, Shukla-Dave A. Extension of the intravoxel incoherent motion model to non-Gaussian diffusion in head and neck cancer. J Magn Reson Imaging. 2012;36:1088–1096.
35. Raunig DL, McShane LM, Pennello G, Gatsonis C, Carson PL, Voyvodic JT, Wahl RL, Kurland BF, Schwarz AJ, Gönen M, Zahlmann G, Kondratovich MV, O'Donnell K, Petrick N, Cole PE, Garra B, Sullivan DC; QIBA Technical Performance Working Group. Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment. Stat Methods Med Res. 2015;24:27–67.
36. Kessler LG, Barnhart HX, Buckler AJ, Choudhury KR, Kondratovich MV, Toledano A, Guimaraes AR, Filice R, Zhang Z, Sullivan DC; QIBA Terminology Working Group. The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions. Stat Methods Med Res. 2015;24:9–26.
37. Obuchowski NA, Reeves AP, Huang EP, Wang XF, Buckler AJ, Kim HJ, Barnhart HX, Jackson EF, Giger ML, Pennello G, Toledano AY, Kalpathy-Cramer J, Apanasovich TV, Kinahan PE, Myers KJ, Goldgof DB, Barboriak DP, Gillies RJ, Schwartz LH, Sullivan DC; Algorithm Comparison Working Group. Quantitative imaging biomarkers: a review of statistical methods for computer algorithm comparisons. Stat Methods Med Res. 2015;24:68–106.
38. Boss MA, editor. Multicenter Study of Reproducibility of Wide Range of ADC at 0°C. Chicago IL: RSNA; 2015.
39. Palacios EM, Martin AJ, Boss MA, Ezekiel F, Chang YS, Yuh EL, Vassar MJ, Schnyer DM, MacDonald CL, Crawford KL, Irimia A, Toga AW, Mukherjee P; TRACK-TBI Investigators. Toward precision and reproducibility of diffusion tensor imaging: a multicenter diffusion phantom and traveling volunteer study. Am J Neuroradiol. 2017;38:537–545.
40. Swanson SD, Malyarenko DI, Fabiilli ML, Paudyal R, LoCastro E, Jambawalikar SR, Liu MZ, Schwartz LH, Shukla-Dave A, Chenevert TL, editors. Design and Development of a Novel Phantom to Assess Quantitative Diffusion Kurtosis Imaging, a Multi-Site Initiative. Bethesda, MD: NCI/Quantitative Imaging Network (QIN) Annual Face-to-Face meeting; 2018.
41. Pierpaoli C, Sarlls J, Nevo U, Basser PJ, Horkay F, editors. Polyvinylpyrrolidone (PVP) Water Solutions as Isotropic Phantoms for Diffusion MRI Studies. Proceedings of the 17th Annual Meeting of the ISMRM; 2009; Honolulu, HW.
42. Lu Y, Hatzoglou V, Banerjee S, Stambuk HE, Gonen M, Shankaranarayanan A, Mazaheri Y, Deasy JO, Shaha AR, Tuttle RM, Shukla-Dave A. Repeatability investigation of reduced field-of-view diffusion-weighted magnetic resonance imaging on thyroid glands. J Comput Assist Tomography. 2015;39:334–339.
43. Jansen JF, Koutcher JA, Shukla-Dave A. Non-invasive imaging of angiogenesis in head and neck squamous cell carcinoma. Angiogenesis. 2010;13:149–160.
44. Rasband WS. ImageJ. Bethesda, MD: U. S. National Institutes of Health; 1997–2016.
45. Chenevert TL, Galban CJ, Ivancevic MK, Rohrer SE, Londy FJ, Kwee TC, Meyer CR, Johnson TD, Rehemtulla A, Ross BD. Diffusion coefficient measurement using a temperature-controlled fluid for quality control in multicenter studies. J Magn Reson Imaging. 2011;34:983–987.
46. Le Bihan D. Intravoxel incoherent motion imaging using steady-state free precession. Magn Reson Med. 1988;7:346–351.
47. Nakahira M, Saito N, Yamaguchi H, Kuba K, Sugasawa M. Use of quantitative diffusion-weighted magnetic resonance imaging to predict human papilloma virus status in patients with oropharyngeal squamous cell carcinoma. Eur Arch Otorhinolaryngol. 2014;271:1219–1225.
48. Arlinghaus LR, Dortch RD, Whisenant JG, Kang H, Abramson RG, Yankeelov TE. Quantitative magnetization transfer imaging of the breast at 3.0 T: reproducibility in healthy volunteers. Tomography. 2016;2:260–266.
49. Galbraith SM, Lodge MA, Taylor NJ, Rustin GJ, Bentzen S, Stirling JJ, Padhani AR. Reproducibility of dynamic contrast-enhanced MRI in human muscle and tumours: comparison of quantitative and semi-quantitative analysis. NMR Biomed. 2002;15:132–142.
50. RstudioTeam. RStudio: Integrated Development for R. Boston, MA: RStudio, Inc; 2015.
51. Malyarenko D, Fedorov A, Bell L, Prah M, Hectors S, Arlinghaus L, Muzi M, Solaiyappan M, Jacobs M, Fung M, Shukla-Dave A, McManus K, Boss M, Taouli B, Yankeelov TE, Quarles CC, Schmainda K, Chenevert TL, Newitt DC. Toward uniform implementation of parametric map Digital Imaging and Communication in Medicine standard in multisite quantitative diffusion imaging studies. J Med Imaging. 2018;5:011006.
52. Newitt DC, Malyarenko D, Chenevert TL, Quarles CC, Bell L, Fedorov A, Fennessy F, Jacobs MA, Solaiyappan M, Hectors S, Taouli B, Muzi M, Kinahan PE, Schmainda KM, Prah MA, Taber EN, Kroenke C, Huang W, Arlinghaus LR, Yankeelov TE, Cao Y, Aryal M, Yen YF, Kalpathy-Cramer J, Shukla-Dave A, Fung M, Liang J, Boss M, Hylton N. Multisite concordance of apparent diffusion coefficient measurements across the NCI Quantitative Imaging Network. J Med Imaging. 2018;5:011003.
53. Bonekamp D, Nagae LM, Degaonkar M, Matson M, Abdalla WM, Barker PB, Mori S, Horská A. Diffusion tensor imaging in children and adolescents: reproducibility, hemispheric, and age-related differences. Neuroimage. 2007;34:733–742.
54. Lee Y, Lee SS, Kim N, Kim E, Kim YJ, Yun SC, Kühn B, Kim IS, Park SH, Kim SY, Lee MG. Intravoxel incoherent motion diffusion-weighted MR imaging of the liver: effect of triggering methods on regional variability and measurement repeatability of quantitative parameters. Radiology. 2015;274:405–415.
55. Le Bihan D. Diffusion MRI: what water tells us about the brain. EMBO Mol Med. 2014;6:569–73.
56. Padhani AR, Koh DM. Diffusion MR imaging for monitoring of treatment response. Magn Reson Imaging Clin North Am. 2011;19:181–209.
57. Lachin JM. Introduction to sample size determination and power analysis for clinical trials. Control Clin Trials. 1981;2:93–113.
58. Paldino MJ, Barboriak D, Desjardins A, Friedman HS, Vredenburgh JJ. Repeatability of quantitative parameters derived from diffusion tensor imaging in patients with glioblastoma multiforme. J Magn Reson Imaging. 2009;29:1199–1205.
59. Pfefferbaum A, Adalsteinsson E, Sullivan EV. Replicability of diffusion tensor imaging measurements of fractional anisotropy and trace in brain. J Magn Reson Imaging. 2003;18:427–433.

TOMOGRAPHY® RESEARCH ARTICLE

Quantitative Non-Gaussian Intravoxel Incoherent Motion Diffusion-Weighted Imaging Metrics and Surgical Pathology for Stratifying Tumor Aggressiveness in Papillary Thyroid Carcinomas

David Aramburu Núñez[1], Yonggang Lu[2], Ramesh Paudyal[1], Vaios Hatzoglou[4], Andre L. Moreira[3], Jung Hun Oh[1], Hilda E. Stambuk[4], Yousef Mazaheri[1], Mithat Gonen[5], Ronald A. Ghossein[6], Ashok R. Shaha[7], R. Michael Tuttle[8], and Amita Shukla-Dave[1,4]

[1]Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY; [2]Department of Radiology, Medical College of Wisconsin, Milwaukee, WI; [3]Department of Pathology, NYU Langone Medical Center, New York, NY; Departments of [4]Radiology, [5]Epidemiology and Biostatistics, [6]Pathology, [7]Surgery, and [8]Medicine, Memorial Sloan Kettering Cancer Center, New York, NY

Corresponding Author:
Amita Shukla-Dave, PhD
Memorial Sloan Kettering Cancer Center,
1275 York Avenue, NY, New York 10065;
E-mail: davea@mskcc.org

Key Words: diffusion-weighted imaging, multi b-value, Gaussian and non-Gaussian, papillary thyroid carcinoma, tumor aggressiveness

Abbreviations: Papillary thyroid cancer (PTC), apparent diffusion coefficient (ADC), papillary thyroid carcinoma (PTC), ultrasonography (US), diffusion-weighted imaging (DWI), intravoxel incoherent motion (IVIM), non-Gaussian (NG), repetition time (TR), echo time (TE), echo planar imaging (EPI), regions of interest (ROIs), magnetic resonance imaging (MRI), leave-1-out cross-validation (LOOCV), extrathyroidal extension (ETE), receiver operating characteristic (ROC), area under the ROC curve (AUC)

ABSTRACT

We assessed a priori aggressive features using quantitative diffusion-weighted imaging metrics to preclude an active surveillance management approach in patients with papillary thyroid cancer (PTC) with tumor size 1–2 cm. This prospective study enrolled 24 patients with PTC who underwent pretreatment multi-b-value diffusion-weighted imaging on a GE 3 T magnetic resonance imaging scanner. The apparent diffusion coefficient (ADC) metric was calculated from monoexponential model, and the perfusion fraction (f), diffusion coefficient (D), pseudo-diffusion coefficient (D^*), and diffusion kurtosis coefficient (K) metrics were estimated using the non-Gaussian intravoxel incoherent motion model. Neck ultrasonography examination data were used to calculate tumor size. The receiver operating characteristic curve assessed the discriminative specificity, sensitivity, and accuracy between PTCs with and without features of tumor aggressiveness. Multivariate logistic regression analysis was performed on metrics using a leave-1-out cross-validation method. Tumor aggressiveness was defined by surgical histopathology. Tumors with aggressive features had significantly lower ADC and D values than tumors without tumor-aggressive features ($P < .05$). The absolute relative change was 46% in K metric value between the 2 tumor types. In total, 14 patients were in the critical size range (1–2 cm) measured by ultrasonography, and the ADC and D were significantly different and able to differentiate between the 2 tumor types ($P < .05$). ADC and D can distinguish tumors with aggressive histological features to preclude an active surveillance management approach in patients with PTC with tumors measuring 1–2 cm.

INTRODUCTION

Currently, many clinicians continue to recommend an aggressive initial management approach to all but the patients with the most low-risk papillary thyroid carcinoma (PTC), which usually includes thyroid surgery and radioactive iodine adjuvant therapy (1). Despite this, a more-stratified, risk-adapted initial management approach has been strongly recommended in the recent American Thyroid Association thyroid cancer clinical practice guidelines (2). The recommendations for either a limited thyroid surgery option (thyroid lobectomy without adjuvant therapy) or an active surveillance management approach (serial observation with neck ultrasonography [US] with surgical intervention deferred until documented disease progression) may be less-drastic incremental options for patients with intrathyroidal papillary thyroid cancers thought to be at low risk for disease-specific mortality and recurrence. These treatment options are being offered on the basis of abundance of data showing excellent clinical outcomes following either thyroid lobectomy or active

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2018.00054

surveillance in properly selected low-risk PTC patients with intrathyroidal disease (1-3). Studies have shown that patients with micropapillary carcinomas having small tumors (size, <1 cm) well confined to the thyroid and without presence of extrathyroidal extension and/or lymph node metastases are good candidates for active surveillance (2). The cumulative risk of extrathyroidal extension and lymph node metastases both increased linearly as the primary tumor increases from 1 to 2 cm (4). It is therefore even more important to ensure that these larger thyroid cancers are confined to the thyroid before considering an observation management approach. US alone is an adequate study for selection of those PTCs with smaller tumors (<1 cm) for active surveillance (5-8). However, for those PTCs with larger tumors (>1 cm up to 2 cm), US has suboptimal sensitivity and specificity in the detection of extrathyroidal extension (9, 10) and cannot reliably detect cervical lymph node metastases deep to the intact thyroid gland or in the infraclavicular, retropharyngeal, and parapharyngeal regions (1). Therefore, some experts have suggested that additional imaging methods be used to verify the absence of disease outside the thyroid when considering a conservative management approach in larger tumors.

Quantitative magnetic resonance imaging (qMRI) is a noninvasive technique that provides images of high spatial resolution with excellent tissue contrast. Quantitative diffusion-weighted imaging (DWI) measures the Brownian motion of water molecules in tumor tissue, which is highly reflective of the cellular organization and membrane integrity (11). DWI has shown promise in the detection, staging, prognosis, and monitoring of thyroid cancers (12-19). Quantitative apparent diffusion coefficient (ADC) metric derived from monoexponential modeling of DWI data, under a Gaussian behavior, using ≥2 b-values (ie, diffusion-weighting factor) (11) reflects tumor cellularity. Recently, clinical relevance for ADC has been shown in assessing extrathyroidal extension in discernable intrathyroidal papillary microcarcinomas (tumor size, <1 cm), an aggressive tumor feature that was limited to identification by surgery only (20).

Le Bihan et al. developed the intravoxel incoherent motion (IVIM) model to describe diffusion in the capillary and tissue compartments separately using multiple b value DWI data set (21, 22). IVIM is a biexponential model, which is based on a Gaussian distribution, and it provides estimates of pseudo-diffusion coefficient (D^*), perfusion fraction (f) within the capillary network, and true diffusion coefficient of the tissue (D) metrics (22). Recent studies have shown the utility of IVIM-DWI in clinical oncology (23-31).

Diffusion in biological tissue is hindered and complex and therefore lends itself well to a non-Gaussian (NG) nature, which has been readily observable via noninvasive imaging at high b-values (32, 33). Using multi-b-value DWI data, NG models [ie, diffusional kurtosis (34, 35) and extension of biexponential IVIM with kurtosis, called NG-IVIM (22, 36)] have been developed to account for hindered and restricted diffusion in tumor tissue. The dimensionless imaging metric K characterizes NG diffusion behavior in tissue microstructure. The quantitative metric K obtained from both diffusion kurtosis and NG-IVIM models has shown feasibility to quantify tissue microstructure in head and neck (HN) cancer (35, 36). Given the known microstructural complexity in PTC, we hypothesized that the NG-IVIM may have greater utility than Gaussian models in risk stratification for active surveillance candidates. The purpose of this study was to identify a priori aggressive histological features using NG-IVIM to preclude an active surveillance management approach in patients with PTC, with tumor diameter size 1–2 cm as measured by US.

MATERIALS AND METHODS

Patients

This clinical study was approved by our institutional review board, which was compliant with the Health Insurance Portability and Accountability Act. In total, 24 patients (age, 27–78 years; male/female, 8/16) were enrolled in this prospective clinical trial, before surgery. All patients who underwent the study signed a form of written consent.

DWI Data Acquisition

MRI examinations were performed on a 3-Tesla GE scanner (General Electric, Milwaukee, WI), with a neurovascular phased-array coil and consisted of standard multiplanar (sagittal, axial, coronal) T1- and T2-weighted imaging scans followed by multi-b-value DWI scans. The T1- and T2-weighted MRI scans covered the whole thyroid gland with a field of view (FOV) of 20–24 cm, slice thickness of 5 mm, and acquisition matrix of 256 × 256. The repetition time (TR)/echo time (TE) for T1-weighted scans were 500 milliseconds (ms)/15 ms; and TR/TE for T2-weighted scans were 4000 ms/80 ms.

Multi-b-value DWI images were acquired using a single-shot spin-echo echo planar imaging (SS-SE-EPI) sequence with TR = 4000 ms, TE = minimum (100–110 ms), number of excitations (NEX) = 4, slice thickness = 5 mm, gap = 0 mm, field of view = 20–24 cm, acquisition matrix of 128 × 128, which was zero-filled and reconstructed to 256 × 256 pixels, with 10 b values of 0, 20, 50, 80, 200, 300, 500, 800, 1000, and 1500 s/mm^2. Images were acquired using 3 orthogonal diffusion gradient direction. The acquisition minimum TE varied between patients (minimum TE 100–110 ms) because of slight differences in obliquity of the prescription. A calibration scan was performed before multi-b-value acquisition to reduce Nyquist (N/2) ghosting artifacts (20, 37). Fat suppression, shimming, and parallel imaging (acceleration factor = 2) techniques were used to reduce imaging artifacts.

DWI Data Processing

The regions of interest (ROIs) on thyroid glands for the PTCs were drawn on the DWI images (b = 0 s/mm^2) by an experienced neuroradiologist (>10 years' experience) using ImageJ (38), in conjunction with the radiological, clinical, and pathological information. All ROIs avoid obvious cystic, hemorrhagic, or calcified portions. All data analyses were performed using an in-house-developed software package, MRI-QAMPER (Quantitative Analysis Multi-Parametric Evaluation Routines) implemented in MATLAB (The MathWorks, Natick, MA). Metric values were estimated on a voxel-by-voxel basis to generate parametric maps, and ROI-averaged values for each quantitative imaging metric were calculated.

The voxel-wise apparent diffusion coefficient (ADC) map within the ROI was calculated from the multi-b-value DWI data, using a monoexponential model given by:

$$S_b = S_0 e^{-b\,ADC} \quad (1)$$

where S_b and S_O are signal intensities with and without diffusion weighting, and b is the diffusion-weighting factor (s/mm^2).

The quantitative imaging metrics estimated from NG-IVIM using the multi-b-value DW-MRI data are given by the following equation (21, 36):

$$S_b = S_0\left[f e^{-bD^*} + (1-f)e^{-bD+\frac{1}{6}Kb^2D^2}\right] \quad (2)$$

where f is the vascular volume fraction; D is the diffusion coefficient (mm^2/s); D^* is the pseudo-diffusion coefficient (mm^2/s) associated with blood perfusion, and K is the diffusion kurtosis coefficient. Under the assumption of a Gaussian distribution ($K = 0$), equation (2) is equivalent to IVIM model equation (21).

Because multi-b-value DWI images are inherently noisy owing to thermal or physiological factors, a noise-rectified method was used for metric estimation, as detailed elsewhere (36). For image processing, DWI data were fitted using a nonlinear least-square fitting method using MRI-QAMPER (24, 36).

Histopathological Examination

Surgical papillary thyroid tumor specimens after radical thyroidectomy or lobectomy were collected under the supervision of an experienced (>10 years) pathologist. Paraffin-embedded tissue blocks were obtained for each surgically resected tumor specimen and stained with hematoxylin and eosin. The hematoxylin and eosin section of each papillary thyroid tumor was reviewed by the same excising pathologist, using established criteria for evaluating tumor aggressiveness (39, 40). The histopathological characteristics of tumor aggressiveness were evaluated individually using the following 6 features: tall cell variant, necrosis, vascular and/or tumor capsular invasion, extrathyroidal extension, regional metastases, and distant metastases. A tumor identified with the presence of any 1 of these features was considered to be aggressive.

US Examination

US examinations were performed according to a standard protocol that includes grayscale and color Doppler US assessment of the thyroid bed and cervical lymph nodes in all neck compartments. US reports include information about size, location, and structure of thyroid nodules and cervical lymph nodes. Size was defined as the largest diameter among the 3 dimensions observed. The US studies were performed with Siemens Acuson S2000 or SEQUOIA (Siemens Medical Solutions, Mountain View, CA), or the GE Logiq 9 (GE Healthcare, Little Chalfont, UK) units, using 8- to 15-MHz linear transducers.

Statistical Analysis

Quantitative imaging metrics ADC, D, f, D^*, and K from NG-IVIM analysis and US measurement values were reported as ROI-based mean ± standard deviation (SD). To compare metric value differences among groups of PTCs with and without features of aggressiveness, a nonparametric Wilcoxon rank-sum test was used. A Spearman correlation analysis was performed between quantitative imaging metrics. The significance level was set at $P \leq .05$.

Finally, the relative percentage change (rc, %) in imaging metrics mean values were calculated as:

$$rc(\%) = \frac{(X_{ag} - X_{nag})}{X_{nag}} \times 100 \quad (3)$$

where X_{ag} and X_{nag} are the quantitative imaging metrics mean values (ie, ADC, D, f, D^* and K) of tumors with and without aggressive features, respectively.

Receiver operating characteristic (ROC) curve analysis was performed for each metric to assess its capability to discriminate between PTC groups with and without aggressive features, resulting in area under the ROC curve (AUC) evaluation. Youdon's index was used to estimate the optimal cutoff values for individual metrics (41, 42). Multivariate logistic regression analysis was performed on relevant metrics using a leave-one-out cross-validation (LOOCV) method for unbiased assessment of the modeling.

All statistical analyses were performed using R software and Stata (43, 44).

RESULTS

Patient characteristics are summarized in Table 1. Of the 24 patients, 13 patients were found to have locoregional metastases by preoperative US imaging. Based on surgical pathology analysis, all 24 patients had PTCs, including 6 patients with the tall cell variant, 1 patient with vascular and/or capsular invasion, 9 patients with extrathyroidal extension, and 16 patients with locoregional metastases. The mean size of the lesion based on US was 16 ± 6 mm and ranged from 6–26 mm.

Figure 1 shows a representative plot of signal intensity decay curve as a function of the b-values (s/mm^2) obtained from a patient with aggressive feature of extrathyroidal extension (ETE) confirmed at surgical pathology.

Figures 2 and 3 show NG-IVIM metric maps overlaid on the DWI images from a representative patient with PTC with aggressive tumor features (female; 28 years; US tumor maximum diameter, 2.1 cm) and a representative patient with PTC without aggressive tumor features (female; 48 years; US tumor maximum diameter, 2.1 cm), respectively. It is interesting to note that maximum tumor diameter in preoperative US was the same for both tumors shown in Figures 2 and 3. However, at surgical pathology, the tumor with aggressive features was found to be in the size range of >2 cm (Figure 2), while tumor with nonaggressive feature was in the size range of 1–2 cm (Figure 3).

Tumors with aggressive features (tall cell variant, necrosis, vascular and/or tumor capsular invasion, ETE, regional metastases, or distant metastases) had significantly lower ADC and D values and higher f values than tumors without aggressive features ($P < .05$) (Figure 2), whereas K and D^* values were not significantly different ($P > .05$) for the 2 groups (Table 2).

Out of the 24 patients, 14 patients were in the critical size range (1–2 cm), and ADC and D were significantly different (Table 2), differentiating between tumors with (n = 10) versus without (n = 4) aggressive features ($P < .05$). The ADC values

Table 1. Patient Characteristics

Characteristic	Values
Age at diagnosis (years)	41 ± 7 (range, 27–78)
Sex	
Female	16 (67%)
Male	8 (33%)
Fine-needle aspiration cytology	
Papillary thyroid cancer	16 (67%)
Suspicious for papillary thyroid cancer	8 (33%)
Preoperative US	
Subcapsular location of tumor	19 (79%)
Extrathyroidal Extension	2 (8%)
Evidence of Lymph node metastases	13 (54%)
Size of papillary carcinoma (mm)	16 ± 6 (range, 6–26)
Histology	
Classic papillary thyroid cancer (cPTC)	13 (54%)
Follicular variant papillary thyroid cancer (fvPTC)	3 (12%)
Diffuse sclerosing PTC (dsPTC)	1 (4%)
Tall cell variant PTC (tPTC)	4 (16%)
Multifocal (cPTC+fvPTC; cPTC+tPTC)	1 + 2 (12%)
Size of papillary carcinoma (mm)	15 ± 6 (range, 5–25)
Aggressive features based on pathology	
Tall cell	6 (25%)
Extrathyroidal extension	9 (38%)
Necrosis	0 (0%)
Vascular and/or tumor capsular invasion	1 (4%)
Regional metastases	16 (67%)
Distant metastases	0 (0%)
Pathology T	
T1a	2 (8%)
T1b	10 (42%)
T2	3 (13%)
T3	9 (38%)
Pathology N	
N0	8 (33%)
N1a	6 (25%)
N1b	10 (42%)
Clinical M	
M0	24 (100%)
M1	0 (0%)
AJCC Stage	
I	17 (71%)
II	2 (8%)
III	3 (13%)
IVA	2 (8%)

were $1.3 \pm 0.3 \times 10^{-3}$ mm^2/s vs $1.9 \pm 0.5 \times 10^{-3}$ mm^2/s for tumors with and without aggressive features, respectively. The D values were $1.3 \pm 0.3 \times 10^{-3}$ mm^2/s vs $2.1 \pm 0.6 \times 10^{-3}$ mm^2/s for tumors with and without aggressive features, respectively. The K, D^*, and f metrics were not significantly different in this cohort ($P > .05$).

Figure 4 boxplot compares the quantitative imaging metrics mean values for ADC, D, and US-measured tumor size (mm) between tumors with and without aggressive features. The absolute relative percentage change (rc, %) in the quantitative imaging metric ADC, D, K, D^*, and f metric values for tumors with aggressive features were 31%, 40%, 46%, 7%, and 31% respectively, in comparison to tumors without aggressive features.

Figure 5 displays the scatter plot between NG-IVIM estimates of two quantitative imaging metrics D and K. The Spearman rank-order correlation coefficient (ρ) was -0.46 ($P < .05$), indicating a significant correlation between the D and K.

Figure 6A shows the estimated ROC curves for quantitative imaging metrics ADC, D, and K. Using ROC analysis, the best cutoff values of ADC, D, and K that discriminate between aggressive PTCs with and without aggressive features were determined as follows: ADC $= 1.79 \times 10^{-3}$ mm^2/s, $D = 1.35 \times 10^{-3}$, and $K = 0.68$. The sensitivity, specificity, and AUC obtained from the ROC curve were 100%, 75%, and 0.875, respectively, for ADC; 80%, 100%, and 0.95, respectively, for D; and 70%, 75%, and 0.725, respectively, for K. The AUC is the highest for metric D, followed by metrics ADC and K. Figure 6B resulted from logistic regression on combined 2 metrics (ADC and D) and 3 metrics which included K based on the LOOCV method. Sensitivity, specificity, and AUC obtained from the LOOCV analysis combining 2 and 3 metrics were as follows: 90%, 75% and 0.70 and 80%, 75%, and 0.65, respectively.

DISCUSSION

To the best of our knowledge there are no published studies that have leveraged the use of biexponential NG-IVIM modeling using multi-b-value DWI data sets to stratify PTCs into tumor groups with and without aggressive features. The quantitative imaging metrics ADC, D, and K exhibit promise as surrogate biomarkers for aggressiveness in patients with PTC, following appropriate validation.

Previously, Lu et al. using monoexponential modeling of quantitative DWI data stratified PTCs into tumor groups with and without ETE, one of the multiple aggressive features, thereby obtaining significantly lower mean ADC values for tumors with ETEs than without $(1.53 \pm 0.25) \times 10^{-3}$ [mm^2/s] vs $(2.4 \pm 0.7) \times 10^{-3}$ [mm^2/s] (20). Previously ETE was identified by surgery only (8, 45). Hao et al., also using DWI, stratified PTCs with and without ETEs, thereby showing significant lower median ADC values for tumor with ETE features $(1.41 \pm 0.29) \times 10^{-3}$ [mm^2/s] vs $(1.53 \pm 0.29) \times 10^{-3}$ [mm^2/s] (46). In the present study the cut-off value of ADC to discriminate PTCs with and without aggressive features was 1.79×10^{-3} mm^2/s and is consistent with the previous studies by Lu et al. and Hao et al., 1.85×10^{-3} mm^2/s and 1.89×10^{-3} mm^2/s, respectively (20, 46).

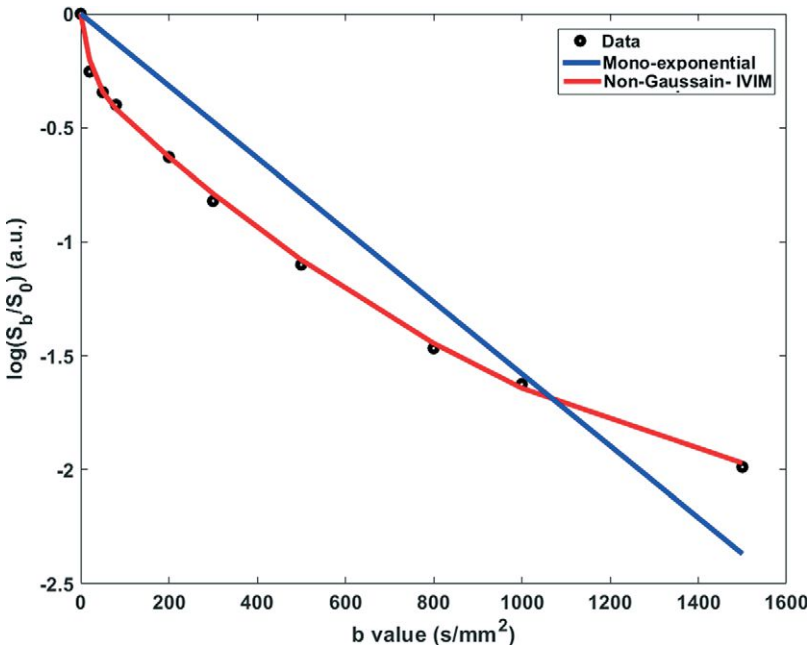


Figure 1. Logarithmic signal intensity (S_b/S_0) plotted as a function of b-value. The experimental data (black circle) obtained from a representative thyroid patient is fitted with a mono exponential model (blue line) and non-Gaussian intravoxel incoherent motion model (red line).

Recently biexponential modeling (IVIM) analysis using multi-b-value DWI data have shown clinical utility in several cancers, including prostate and head and neck (30, 47-49). Valerio et al. have shown that ADC and *D* values are significantly lower in prostate cancer tissue compared with healthy tissue ($0.76 \pm 0.27 \times 10^{-3}$ [mm^2/s] vs $0.99 \pm 0.38 \times 10^{-3}$ [mm^2/s]) (47). In addition, Barbieri et al. found that ADC and *D* differ significantly between high- and low-grade prostate cancer lesions ($0.76 \pm 0.27 \times 10^{-3}$ [mm^2/s] vs $0.99 \pm 0.38 \times 10^{-3}$ [mm^2/s]) (48). The clinical utility of multi-b-value DWI is being tested in cancers in the head and neck region, including the thyroid gland (30, 31, 49). Shen et al. investigated the feasibility of using IVIM to detect radiation changes of normal-appearing parotid glands in patients with differentiated thyroid cancer after radioiodine therapy (49). In a small study of 8 healthy volunteers, Becker et al. used an IVIM-derived imaging metric to

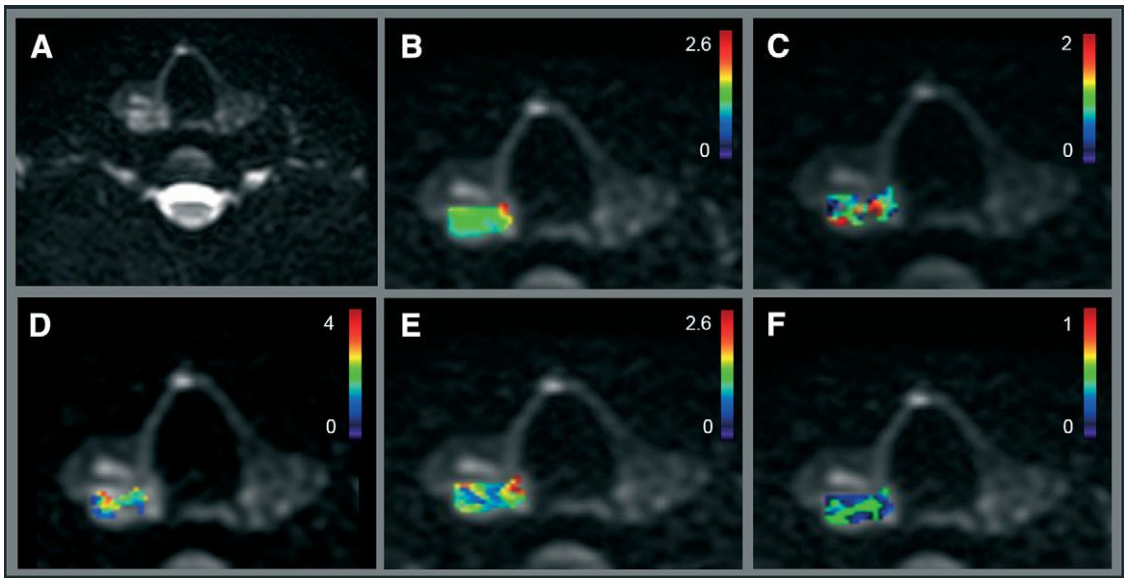


Figure 2. The representative patient with papillary thyroid carcinoma (PTC) with tumor aggressive features (female; 28 years; ultrasonography [US] maximum tumor diameter, 2.1 cm). Diffusion-weighted image (b = 0 s/mm²) (A). ADC map ($\times 10^{-3}$ mm^2/s) overlaid on diffusion-weighted image (b = 0 s/mm²) (B). *K* map overlaid on diffusion-weighted image (b = 0 s/mm²) (C). *D** ($\times 10^{-3}$ mm^2/s) map overlaid on diffusion-weighted image (b = 0 s/mm²) (D). *D* map ($\times 10^{-3}$ mm^2/s) overlaid on diffusion-weighted image (b = 0 s/mm²) (E). *f* map overlaid on diffusion-weighted image (b = 0 s/mm²) (F).

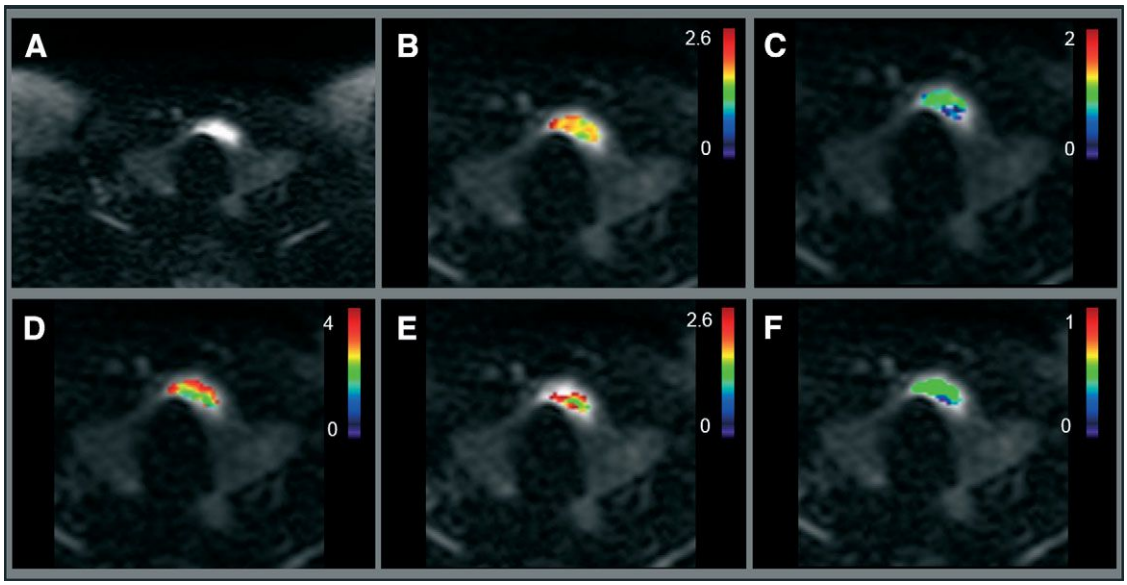


Figure 3. The representative patient with PTC without tumor-aggressive features (female; 48 years; US maximum tumor diameter, 2.1 cm). Diffusion-weighted image (b = 0 s/mm^2) (A). ADC map ($\times 10^{-3}$ mm^2/s) overlaid on diffusion-weighted image (b = 0 s/mm^2) (B). *K* map overlaid on diffusion-weighted image (b = 0 s/mm^2) (C). *D** ($\times 10^{-3}$ mm^2/s) map overlaid on diffusion-weighted image (b = 0 s/mm^2) (D). *D* map ($\times 10^{-3}$ mm^2/s) overlaid on diffusion-weighted image (b = 0 s/mm^2) (E). *f* map overlaid on diffusion-weighted image (b = 0 s/mm^2) (F).

establish a comprehensive description of tissue properties of healthy thyroid tissue (50). The IVIM imaging metric *D* was shown to be significantly different between complete responders (the change between pre- and intratreatment week 3 was from $0.67 \pm 0.17 \times 10^{-3}$ mm^2/s to $0.98 \pm 0.28 \times 10^{-3}$ mm^2/s) and noncomplete responders (the change between pre- and intratreatment week 3 was from $0.59 \pm 0.10 \times 10^{-3}$ mm^2/s to $0.72 \pm 0.03 \times 10^{-3}$ mm^2/s) in patients with head and neck squamous cell carcinoma treated with radiotherapy (31). For tumors with hindered and restricted diffusion, NG-IVIM modeling analysis from multi-b-value DWI, as developed by Lu et al., has shown to be a better-fitting model in head and neck region (36). This model is used in the present study for the first time in the thyroid region.

The findings from 14 patients with PTC with tumor diameter 1–2 cm (as measured by US) emphasize the role of NG-IVIM DWI in differentiating this sub group. Preoperative US could identify 6 out of 14 patients with aggressive features, while NG-IVIM DWI indicated 11 patients. As ground truth, there were 10 patients with aggressive tumor features determined by pathology, our reference standard (Table 2). Therefore, NG-IVIM could correctly identify all 10 patients with aggressive tumor features confirmed by pathology, whereas US correctly identified only 6 patients. US is the

Table 2. Statistical Analysis (mean ± SD) for Quantitative Imaging Metrics Using Tumor Size by US

US Tumor Size	<1 cm (n = 3)		1–2 cm (n = 14)		>2 cm (n = 7)	
Aggressive features on US	YES (n = 2)	NO (n = 1)	YES (n = 6)	NO (n = 8)	YES (n = 5)	NO (n = 2)
Aggressive features on pathology	YES (n = 3)	NO (n = 0)	YES (n = 10)	NO (n = 4)	YES (n = 5)	NO (n = 2)
ADC $\times 10^{-3}$ (mm^2/s)	(1.2 ± 0.7)	–	(1.32 ± 0.27)[a]	(1.9 ± 0.5)[a]	(1.7 ± 0.4)	(2.03 ± 0.06)
D $\times 10^{-3}$ (mm^2/s)	(1.4 ± 0.7)	–	(1.27 ± 0.25)[a]	(2.1 ± 0.6)[a]	(1.7± 0.6)	(2.20 ± 0.08)
*D** $\times 10^{-3}$ (mm^2/s)	(2.61 ± 0.62)	–	(2.84 ± 0.06)	(2.95 ± 0.06)	(2.7 ± 0.3)	(2.98 ± 0.02)
f	(0.17 ± 0.05)	–	(0.21 ± 0.06)	(0.16 ± 0.05)	(0.18 ± 0.05)	(0.10 ± 0.02)
K	(0.7 ± 0.6)	–	(0.70 ± 0.26)	(0.48 ± 0.29)	(0.71 ± 0.28)	(0.64 ± 0.15)

[a]statistical significance $P < 0.05$.

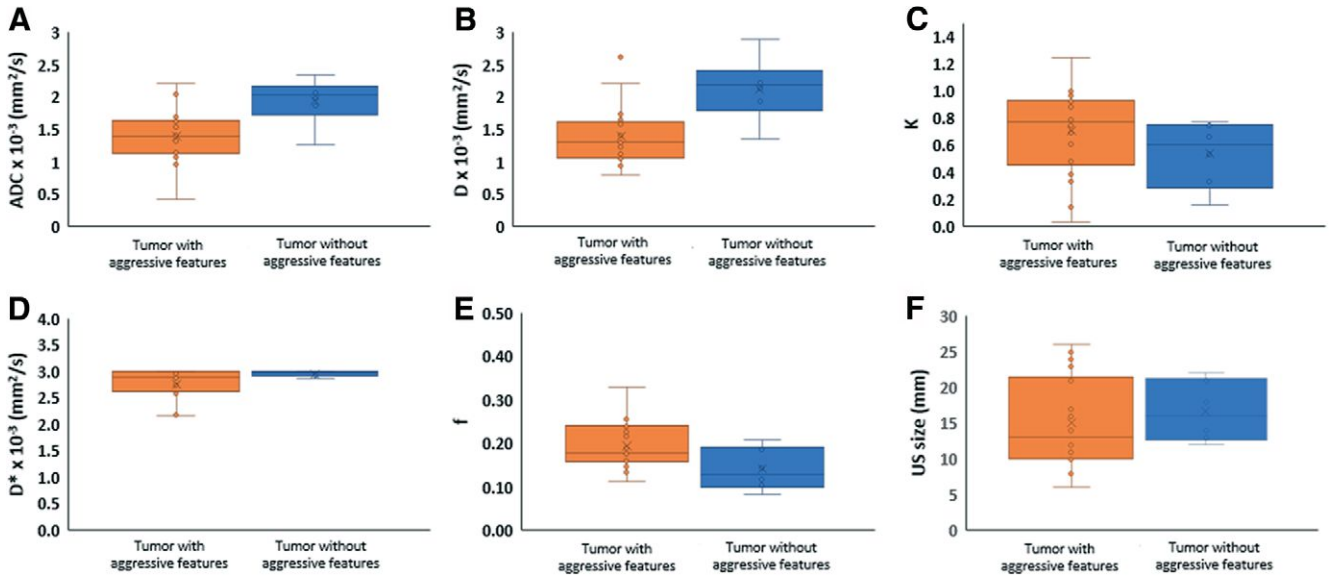


Figure 4. Box-and-whisker plots comparing the mean values for quantitative imaging metrics for all tumors in the 2 groups (tumor with and without aggressive features): ADC $\times 10^{-3}$ (mm^2/s) (A), $D \times 10^{-3}$ (mm^2/s) (B), f (C), $D^* \times 10^{-3}$ (mm^2/s) (D), K (E), and radiological information from US (mm) (F).

imaging modality most commonly used to identify and monitor locoregional disease progression and recurrence in thyroid cancer. However, US is unable to preoperatively identify features such as tall cell variant, necrosis, vascular, and/or tumor capsular invasion or distant metastases (5). NG-IVIM DWI were able to correlate nonaggressive tumor features in 3 out of 4 patients, whereas US overestimated nonaggressiveness in 8 patients. These data strongly suggest that US and MRI are complementary and should be used in combination for patients with tumor size in the range of 1–2 cm. This finding is of key clinical importance for treating physicians who are considering active surveillance for said patient population.

Quantitative NG-IVIM DWI and its derived diffusion and perfusion imaging biomarkers have shown promise in this study of patients with PTC when grouped on the basis of different tumors sizes from preoperative US measurements. In addition, for US-measured tumors sized in the range of 1–2 cm, substantial difference was observed in rc (%) in the NG-IVIM-derived metrics between the 2 groups. D is the true diffusion coefficient metric and a surrogate biomarker of tumor cellularity with 40% change, while metric K is considered as an index of tissue microstructure related to hindered and restricted diffusion with 46% change observed for tumors with aggressive features on comparison to tumors without aggressive feature. The rc (%) for imaging metrics f and D^* were 31% and 7%, and these imaging metrics remain exploratory in nature as their biological meaning has yet to be fully understood. In the present study, K was not necessarily independent of D for all tumors but a weak correlation coefficient between these 2 quantitative imaging

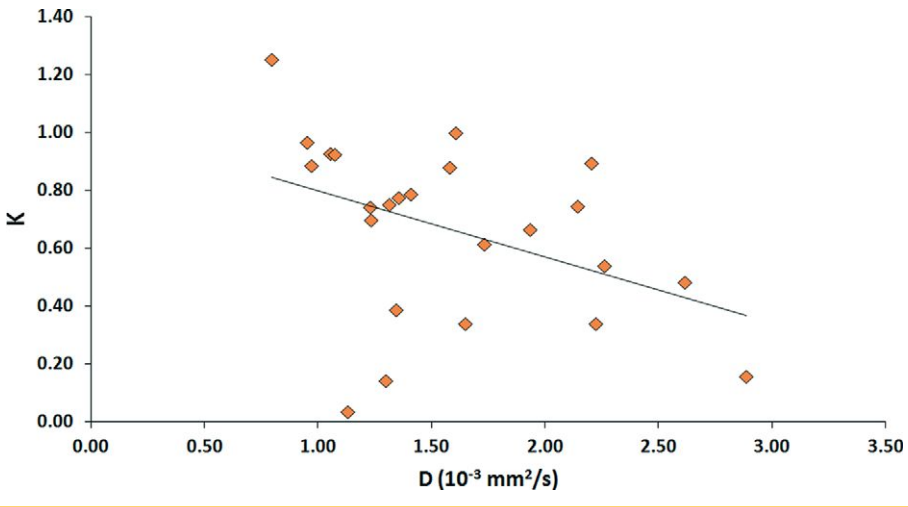


Figure 5. Scatter plot of the true diffusion coefficient (D) and the kurtosis value (K) obtained from all thyroid patients, showing a statistically significant negative correlation ($\rho = -0.46$; $P < 0.05$).

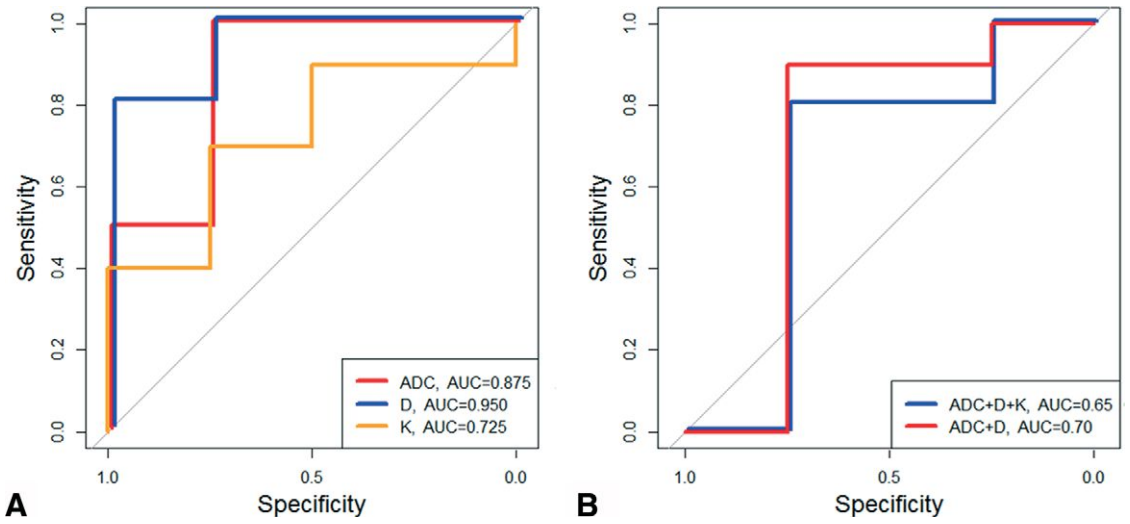


Figure 6. Receiver operating characteristic (ROC) curve to discriminate patients with PTC with and without aggressive features using apparent diffusion coefficient (ADC, black line), *D* (blue line), *K* (orange line) (A). ROC curve from logistic regression based on a leave-one-out cross validation method for the combination of ADC, *D*, and *K* (blue line) and ADC and D (red line) (B).

metrics suggested that *K* might provide additional information related to tissue microstructure. Similar results have been reported previously using NG analysis of DWI in head and neck squamous cell carcinoma (51).

In the present study, for all US-based tumor sizes, the univariate analysis showed the most favorable predictive power with *D* (AUC = 0.95). However, the AUC is lower for both combinations of the metrics in the cross-validation–multivariate analysis, implying cross-validation is necessary to build the predictive model for more realistic and unbiased assessment. The decrease in AUC between 2- and 3- metric models, is due to the metric *K* which may have discriminatory power for US-based tumor size, in the range of 1–2 cm only. For differentiation between the 2 PTC groups, the 2-metric model appears to be the model of choice.

These findings indicate that the quantitative imaging metrics derived from NG-IVIM modeling can provide important risk stratification information and additional insights into potential tumor behavior that cannot be gained from US evaluation alone. As consideration is being given to extend active surveillance to tumors larger than 1 cm, it is increasingly important to develop additional noninvasive tools to help clinicians risk-stratify these slightly larger tumors.

There are several known limitations in this study. First, further investigation is needed in those cohorts with tumors diameters >2 cm and <1 cm. Although no active surveillance is needed for tumors that are >2 cm, it is important to identify aggressive features in tumors that are <1 cm, as has been shown by Lu et al. for papillary microcarcinomas with ETE features (20). Second, a validation study with a larger cohort of patients with PTC is necessary to confirm our initial findings for use in clinical trials. Finally, DWI acquisition using SS EPI suffers from susceptibility artifacts owing to voluntary and involuntary bulk motion in the thyroid region (52). Modified sequences, such as reduced field of view, can help obtain images with fewer distortions (53).

In conclusion, quantitative imaging biomarkers (ADC, *D*, and *K*) derived from NG-IVIM DWI could be used to noninvasively identify tumors with aggressive histological features to preclude an active surveillance management approach in patients with PTC with primary tumor diameters ranging between 1–2 cm.

ACKNOWLEDGMENTS

We would like to thank Mr. Christian Czmielewski (MSc) for his kind contribution to data management and Ms. Eve LoCastro (MS) for carefully editing the manuscript. This research was supported by the National Cancer Institute/National Institutes of Health (grant numbers R21CA176660-01A1 and P50 CA172012-01A1) and in part through the NIH/NCI Cancer Center Support Grant (P30 CA008748).

Disclosures: No disclosures to report.

Conflict of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. American Thyroid Association (ATA) Guidelines Taskforce on Thyroid Nodules and Differentiated Thyroid Cancer; Cooper DS, Doherty GM, Haugen BR, Kloos RT, Lee SL, Mandel SJ, Mazzaferri EL, McIver B, Pacini F, Schlumberger M, Sherman SI, Steward DL, Tuttle RM. Revised American thyroid association management guidelines for patients with thyroid nodules and differentiated thyroid cancer. Thyroid. 2009;19:1167–1214.
2. Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, Pacini F, Randolph GW, Sawka AM, Schlumberger M, Schuff KG, Sherman SI, Sosa JA, Steward DL, Tuttle RM, Wartofsky L. 2015 American thyroid association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American thyroid association guidelines task force on thyroid nodules and differentiated thyroid cancer. Thyroid. 2016;26:1–133.
3. Eggener SE, Mueller A, Berglund RK, Ayyathurai R, Soloway C, Soloway MS, Abouassaly R, Klein EA, Jones SJ, Zappavigna C, Goldenberg L, Scardino PT, Eastham JA, Guillonneau B. A multi-institutional evaluation of active surveillance for low risk prostate cancer. J Urol. 2009;181:1635–1641; discussion 1641.
4. Machens A, Holzhausen HJ, Dralle H. The prognostic value of primary tumor size in papillary and follicular thyroid carcinoma—a comparative analysis. Cancer. 2005;103:2269–2273.
5. Ito Y, Amino N, Miyauchi A. Thyroid ultrasonography. World J Surg. 2010;34:1171–1180.
6. Ito Y, Masuoka H, Fukushima M, Inoue H, Kihara M, Tomoda C, Higashiyama T, Takamura Y, Kobayashi K, Miya A, Miyauchi A. Excellent prognosis of patients with solitary T1N0M0 papillary thyroid carcinoma who underwent thyroidectomy and elective lymph node dissection without radioiodine therapy. World J Surg. 2010;34:1285–1290.
7. Ito Y, Miyauchi A. Appropriate treatment for asymptomatic papillary microcarcinoma of the thyroid. Expert Opin Pharmacother. 2007;8:3205–3215.
8. Ito Y, Miyauchi A, Inoue H, Fukushima M, Kihara M, Higashiyama T, Tomoda C, Takamura Y, Kobayashi K, Miya A. An observational trial for papillary thyroid microcarcinoma in Japanese patients. World J Surg. 2010;34:28–35.
9. Lee CY, Kim SJ, Ko KR, Chung KW, Lee JH. Predictive factors for extrathyroidal extension of papillary thyroid carcinoma based on preoperative sonography. J Ultras Med. 2014;33:231–238.
10. Gweon HM, Son EJ, Youk JH, Kim JA, Park CS. Preoperative assessment of extrathyroidal extension of papillary thyroid carcinoma comparison of 2- and 3-dimensional sonography. J Ultras Med. 2014;33:819–825.
11. Basser PJ, Jones DK. Diffusion-tensor MRI: theory, experimental design and data analysis—a technical review. NMR Biomed. 2002;15:456–467.
12. Tezuka M, Murata Y, Ishida R, Ohashi I, Hirata Y, Shibuya H. MR imaging of the thyroid: correlation between apparent diffusion coefficient and thyroid gland scintigraphy. J Magn Reson Imaging. 2003;17:163–169.
13. Tunca F, Giles Y, Salmaslioglu A, Poyanli A, Yilmazbayhan D, Terzioglu T, Tezelman S. The preoperative exclusion of thyroid carcinoma in multinodular goiter: Dynamic contrast-enhanced magnetic resonance imaging versus ultrasonography-guided fine-needle aspiration biopsy. Surgery. 2007;142:992–1002; discussion e1–e2.
14. Razek AA, Sadek AG, Kombar OR, Elmahdy TE, Nada N. Role of apparent diffusion coefficient values in differentiation between malignant and benign solitary thyroid nodules. AJNR Am J Neuroradiol. 2008;29:563–568.
15. Schueller-Weidekamm C, Kaserer K, Schueller G, Scheuba C, Ringl H, Weber M, Czerny C, Herneth AM. Can quantitative diffusion-weighted MR imaging differentiate benign and malignant cold thyroid nodules? Initial results in 25 patients. AJNR Am J Neuroradiol. 2009;30:417–422.
16. Bozgeyik Z, Coskun S, Dagli AF, Ozkan Y, Sahpaz F, Ogur E. Diffusion-weighted MR imaging of thyroid nodules. Neuroradiology. 2009;51:193–198.
17. Erdem G, Erdem T, Muammer H, Mutlu DY, Firat AK, Sahin I, Alkan A. Diffusion-weighted images differentiate benign from malignant thyroid nodules. J Magn Reson Imaging. 2010;31:94–100.
18. Schueller-Weidekamm C, Schueller G, Kaserer K, Scheuba C, Ringl H, Weber M, Czerny C, Herneth AM. Diagnostic value of sonography, ultrasound-guided fine-needle aspiration cytology, and diffusion-weighted MRI in the characterization of cold thyroid nodules. Eur J Radiol. 2010;73:538–544.
19. Mutlu H, Sivrioglu AK, Sonmez G, Velioglu M, Sildiroglu HO, Basekim CC, Kizilkaya E. Role of apparent diffusion coefficient values and diffusion-weighted magnetic resonance imaging in differentiation between benign and malignant thyroid nodules. Clin Imaging. 2012;36:1–7.
20. Lu Y, Moreira AL, Hatzoglou V, Stambuk HE, Gonen M, Mazaheri Y, Deasy JO, Shaha AR, Tuttle RM, Shukla-Dave A. Using diffusion-weighted MRI to predict aggressive histological features in papillary thyroid carcinoma: a novel tool for preoperative risk stratification in thyroid cancer. Thyroid. 2015;25:672–680.
21. Le Bihan D. Intravoxel incoherent motion imaging using steady-state free precession. Magn Reson Med. 1988;7:346–351.
22. Le Bihan D, Breton E, Lallemand D, Aubin ML, Vignaud J, Laval-Jeantet M. Separation of diffusion and perfusion in intravoxel incoherent motion MR imaging. Radiology. 1988;168:497–505.
23. Lemke A, Laun FB, Klauss M, Re TJ, Simon D, Delorme S, Schad LR, Stieltjes B. Differentiation of pancreas carcinoma from healthy pancreatic tissue using multiple b-values: comparison of apparent diffusion coefficient and intravoxel incoherent motion derived parameters. Invest Radiol. 2009;44:769–775.
24. Lu Y, Jansen JF, Stambuk HE, Gupta G, Lee N, Gonen M, Moreira A, Mazaheri Y, Patel SG, Deasy JO, Shah JP, Shukla-Dave A. Comparing primary tumors and metastatic nodes in head and neck cancer using intravoxel incoherent motion imaging: a preliminary experience. J Comput Assist Tomogr. 2013;37:346–352.
25. Riches SF, Hawtin K, Charles-Edwards EM, de Souza NM. Diffusion-weighted imaging of the prostate and rectal wall: comparison of biexponential and monoexponential modelled diffusion and associated perfusion coefficients. NMR Biomed. 2009;22:318–325.
26. Sigmund EE, Cho GY, Kim S, Finn M, Moccaldi M, Jensen JH, Sodickson DK, Goldberg JD, Formenti S, Moy L. Intravoxel incoherent motion imaging of tumor microenvironment in locally advanced breast cancer. Magn Reson Med. 2011;65:1437–1447.
27. Zhu L, Wang H, Zhu L, Meng J, Xu Y, Liu B, Chen W, He J, Zhou Z, Yang X. Predictive and prognostic value of intravoxel incoherent motion (IVIM) MR imaging in patients with advanced cervical cancers undergoing concurrent chemo-radiotherapy. Sci Rep. 2017;7:11635.
28. Lu W, Jing H, Ju-Mei Z, Shao-Lin N, Fang C, Xiao-Ping Y, Qiang L, Su-Yu Z, Ying G. Intravoxel incoherent motion diffusion-weighted imaging for discriminating the pathological response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer. Sci Rep. 2017;7:8496.
29. Detsky JS, Keith J, Conklin J, Symons S, Myrehaug S, Sahgal A, Heyn CC, Soliman H. Differentiating radiation necrosis from tumor progression in brain metastases treated with stereotactic radiotherapy: utility of intravoxel incoherent motion perfusion MRI and correlation with histopathology. J Neurooncol. 2017;134:433–441.
30. Ding Y, Hazle JD, Mohamed AS, Frank SJ, Hobbs BP, Colen RR, Gunn GB, Wang J, Kalpathy-Cramer J, Garden AS, Lai SY, Rosenthal DI, Fuller CD. Intravoxel incoherent motion imaging kinetics during chemoradiotherapy for human papillomavirus-associated squamous cell carcinoma of the oropharynx: preliminary results from a prospective pilot study. NMR Biomed. 2015;28:1645–1654.
31. Paudyal R, OH JH, Riaz N, Venigalla P, Li J, Hatzoglou V, Leeman J, Nunez DA, Lu Y, Deasy JO, Lee N, Shukla-Dave A. Intravoxel incoherent motion diffusion-weighted MRI during chemoradiation therapy to characterize and monitor treatment response in human papillomavirus head and neck squamous cell carcinoma. J Magn Reson Imaging. 2017;45:1013–1023.
32. Le Bihan D. Intravoxel incoherent motion perfusion MR imaging: a wake-up call. Radiology. 2008;249:748–752.
33. Jensen JH, Helpern JA, Ramani A, Lu H, Kaczynski K. Diffusional kurtosis imaging: the quantification of non-gaussian water diffusion by means of magnetic resonance imaging. Magn Reson Med. 2005;53:1432–1440.
34. Jansen JF, Koutcher JA, Shukla-Dave A. Non-invasive imaging of angiogenesis in head and neck squamous cell carcinoma. Angiogenesis. 2010;13:149–160.
35. Jansen JF, Lu Y, Stambuk HE, Lee NY, Koutcher JA, Shukla-Dave A. Kurtosis analysis for DWI improves prediction of short-term response in head and neck cancer. Proc Int Soc Magn Reson Med 2011;1505.
36. Lu Y, Jansen JF, Mazaheri Y, Stambuk HE, Koutcher JA, Shukla-Dave A. Extension of the intravoxel incoherent motion model to non-gaussian diffusion in head and neck cancer. J Magn Reson Imaging. 2012;36:1088–1096.
37. Grieve SM, Blamire AM, Styles P. Elimination of Nyquist ghosting caused by read-out to phase-encode gradient cross-terms in EPI. Magnet Reson Med. 2002;47:337–343.
38. Rasband WS. ImageJ. Bethesda, MD: U. S. National Institutes of Health; 1997–2016.
39. Ghossein R, Ganly I, Biagini A, Robenshtok E, Rivera M, Tuttle RM. Prognostic factors in papillary microcarcinoma with emphasis on histologic subtyping: a clinicopathologic study of 148 cases. Thyroid. 2014;24:245–253.
40. Ganly I, Ibrahimpasic T, Rivera M, Nixon I, Palmer F, Patel SG, Tuttle RM, Shah JP, Ghossein R. Prognostic implications of papillary thyroid carcinoma with tall-cell features. Thyroid. 2014;24:662–670.
41. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. Biom J. 2005;47:458–472.
42. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3:32–35.
43. RstudioTeam. RStudio: Integrated Development for R. Boston, MA: RStudio, Inc; 2015.
44. Stata Statistical Software: Release 15. College Station, TX: StataCorp LLC; 2017.

45. Rivera M, Ricarte J, Tuttle RM, Ganly I, Shaha A, Knauf J, Fagin J, Ghossein R. Molecular, morphologic, and outcome analysis of thyroid carcinomas according to degree of extrathyroid extension. Thyroid. 2010;20:1085–1093.
46. Hao Y, Pan C, Chen W, Li T, Zhu W, Qi J. Differentiation between malignant and benign thyroid nodules and stratification of papillary thyroid cancer with aggressive histological features: whole-lesion diffusion-weighted imaging histogram analysis. J Magn Reson Imaging. 2016;44:1546–1555.
47. Valerio M, Zini C, Fierro D, Giura F, Colarieti A, Giuliani A, Laghi A, Catalano C, Panebianco V. 3T multiparametric MRI of the prostate: does intravoxel incoherent motion diffusion imaging have a role in the detection and stratification of prostate cancer in the peripheral zone? Eur J Radiol. 2016;85:790–794.
48. Barbieri S, Bronnimann M, Boxler S, Vermathen P, Thoeny HC. Differentiation of prostate cancer lesions with high and with low Gleason score by diffusion-weighted MRI. Eur Radiol. 2017;27:1547–1555.
49. Shen J, Xu XQ, Su GY, Hu H, Shi HB, Liu W, Wu FY. Intravoxel incoherent motion magnetic resonance imaging of the normal-appearing parotid glands in patients with differentiated thyroid cancer after radioiodine therapy. Acta Radiol. 2018;59:204–211.
50. Becker AS, Wurnig MC, Finkenstaedt T, Boss A. Non-parametric intravoxel incoherent motion analysis of the thyroid gland. Heliyon. 2017;3:e00239.
51. Jansen JF, Stambuk HE, Koutcher JA, Shukla-Dave A. Non-gaussian analysis of diffusion-weighted MR imaging in head and neck squamous cell carcinoma: a feasibility study. AJNR Am J Neuroradiol. 2010;31:741–748.
52. Skare S, Newbould RD, Clayton DB, Albers GW, Nagle S, Bammer R. Clinical multishot DW-EPI through parallel imaging with considerations of susceptibility, motion, and noise. Magnet Reson Med. 2007;57:881–890.
53. Lu Y, Hatzoglou V, Banerjee S, Stambuk HE, Gonen M, Shankaranarayanan A, Mazaheri Y, Deasy JO, Shaha AR, Tuttle RM, Shukla-Dave A. Repeatability investigation of reduced field-of-view diffusion-weighted magnetic resonance imaging on thyroid glands. J Comput Assist Tomogr. 2015;39:334–339.

TOMOGRAPHY® RESEARCH ARTICLE

Multicenter Repeatability Study of a Novel Quantitative Diffusion Kurtosis Imaging Phantom

Dariya I. Malyarenko[1], Scott D. Swanson[1], Amaresha S. Konar[2], Eve LoCastro[2], Ramesh Paudyal[2], Michael Z. Liu[3], Sachin R. Jambawalikar[3], Lawrence H. Schwartz[3], Amita Shukla-Dave[2,4], and Thomas L. Chenevert[1]

[1]Department of Radiology, University of Michigan Medical School, Ann Arbor, MI; [2]Departments of Medical Physics and [4]Radiology, Memorial Sloan Kettering Cancer Center, New York, NY; and [3]Department of Radiology, Columbia University Irving Medical Center, New York, NY

Corresponding Author:
Dariya I. Malyarenko, PhD
1500 E. Medical Center Dr.; UHB2 Room A205F,
University of Michigan Hospitals,
Ann Arbor, MI 48109-5030;
E-mail: dariya@umich.edu

Key Words: diffusion kurtosis, micro-scale lamellar vesicles, tunable parameters, repeatability, temporal stability

Abbreviations: Diffusion kurtosis imaging (DKI), diffusion-weighted imaging (DWI), apparent diffusion coefficient (ADC), signal-to-noise ratio (SNR), cetyltrimethylammonium bromide (CTAB), behentriammonium chloride (BTAC), cetearyl alcohol (CA), decyl alcohol (DEC), region of interest (ROI), Bland–Altman (BA), limits of agreement (LOAs), single shot echo planer imaging (SS EPI), field of view (FOV), quantitative imaging biomarker (QIB), digital image communication in medicine (DICOM)

ABSTRACT

Quantitative kurtosis phantoms are sought by multicenter clinical trials to establish accuracy and precision of quantitative imaging biomarkers on the basis of diffusion kurtosis imaging (DKI) parameters. We designed and evaluated precision, reproducibility, and long-term stability of a novel isotropic (i)DKI phantom fabricated using four families of chemicals based on vesicular and lamellar mesophases of liquid crystal materials. The constructed iDKI phantoms included negative control monoexponential diffusion materials to independently characterize noise and model-induced bias in quantitative kurtosis parameters. Ten test–retest DKI studies were performed on four scanners at three imaging centers over a six-month period. The tested prototype phantoms exhibited physiologically relevant apparent diffusion, D_{app}, and kurtosis, K_{app}, parameters ranging between 0.4 and 1.1 ($\times 10^{-3}$ mm^2/s) and 0.8 and 1.7 (unitless), respectively. Measured kurtosis phantom K_{app} exceeded maximum fit model bias (0.1) detected for negative control (zero kurtosis) materials. The material-specific parameter precision [95% CI for D_{app}: 0.013–0.022($\times 10^{-3}$ mm^2/s) and for K_{app}: 0.009–0.076] derived from the test–retest analysis was sufficient to characterize thermal and temporal stability of the prototype DKI phantom through correlation analysis of inter-scan variability. The present study confirms a promising chemical design for stable quantitative DKI phantom based on vesicular mesophase of liquid crystal materials. Improvements to phantom preparation and temperature monitoring procedures have potential to enhance precision and reproducibility for future multicenter iDKI phantom studies.

INTRODUCTION

Diffusion-weighted imaging (DWI) is extensively used in clinical radiology studies to monitor changes in water mobility that reflect altered tissue cellularity (1-3). These alterations often arise from malignancy (4-6) or in response to treatment (7-9). Quantitative parametric maps are derived on the basis of physical models for DWI signal dependence on diffusion gradient-weighting strength (denoted by *b*-value). A single-component diffusion model, most widely used by clinical oncology trials (7, 9, 10), assumes monoexponential DWI signal decay with increasing *b*-value, where the decay rate is quantified by apparent diffusion coefficient (ADC).

Diffusion kurtosis (11, 12) is a heuristic extension of the single-component model that introduces an additional quantitative parameter (apparent kurtosis coefficient, K_{app}) to describe the degree of non-Gaussian deviation from monoexponential signal decay in tissue observed for certain in vivo structures and malignancies with increasing *b*-values (5, 13-15). These deviations are typically caused by the presence of cellular structures that substantially impede water mobility, leading to sustained DWI signal at high *b*-values (1, 11). Because typical diffusion kurtosis imaging (DKI) parameter fit is performed over a limited range of *b*-values ($b_{max} < 3000$ s/mm^2), the derived diffusion and kurtosis values are "apparent" rather than absolute characteristics.

Recently there has been a surge of interest in the diffusion imaging community to evaluate K_{app} as a noninvasive, surrogate biomarker of tissue microstructure (5, 13, 15-17). Unlike

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2018.00030

classic diffusion kurtosis in anisotropic brain tissue (11, 12), for nominally isotropic cancerous parenchyma, observed relatively high apparent kurtosis (0.8–1.7, for example, in head and neck or prostate and bladder cancers [5, 13-15]) is typically associated with tumor potency. To use DKI parameters as quantitative imaging biomarkers (QIBs) of tumor response to therapy in multicenter oncology trials (16, 17), the precision (repeatability) and accuracy (bias) of the potential QIBs need to be evaluated (18, 19) across multiple scanner platforms using a common scan protocol (20, 21). Construction of a novel phantom, one that provides true parameter values in the physiologically relevant ranges (5, 13-15), is the first step for the development of a repeatable multisite study protocol and the only means for the absolute bias estimate (20, 21).

The search for a viable DKI phantom has been ongoing for over a decade. The "natural" phantoms based on cream and asparagus (12, 13, 22) provide single "untunable" kurtosis parameter value and perish quickly. Synthetic phantoms comprising the polyethylene particle suspensions (23) and most recently suggested microbead impregnated gels (24) are more stable, but still suffer from limited range of provided kurtosis parameters ($K_{app} < 0.7$) and limited precision owing to microscopic sample inhomogeneity, chemical shift (23), and/or low signal-to-noise ratio (SNR) (short T2) (24). Our recent pilot study (25) proposed the development of novel kurtosis phantoms based on lamellar (amorphous layers) and vesicular (fluid-filled microsacs) phases of liquid crystal systems. These molecular constructs are composed of hydrophobic long-chain fatty alcohols and surfactants that mimic tissue cellularity by forming regularly spaced membranous mesostructures that impede water diffusion. Altering relative concentrations of restricted and free water pools allows a broad range of tunable apparent kurtosis parameters (25) with sufficient SNR for easy quantitative DKI scan protocol testing.

The purpose of the present multi-site study was to evaluate precision, reproducibility, and long-term stability of a novel (prototype) isotropic (i)DKI phantom, fabricated using four families of chemicals based on select combinations of vesicular and lamellar mesophases of liquid crystal materials with adjustable restricted diffusion fraction. The desired iDKI phantom characteristics included long-term temporal stability and homogeneous iDKI model parameters, tunable over physiologically relevant ranges.

METHODOLOGY

To guide design of the next-generation phantom toward improved stability and reproducibility, this study included the following four steps: [1] development and fabrication of the prototype iDKI phantom using four families of liquid crystal materials and three negative controls, [2] implementation of a common quantitative iDKI test–retest scan protocol, [3] parametric map generation and intra-scan test–retest repeatability analysis to establish measurement precision, and [4] apparent (water ADC-based) temperature calibration for characterization of thermal versus temporal inter-scan variability.

Isotropic Diffusion Kurtosis Imaging (iDKI) Phantoms

Four quantitative iDKI phantom materials were chemically designed based on water solutions of paired long carbon-chain surfactants (cetyltrimethylammonium bromide [CTAB] or behentriammonium chloride [BTAC]) and alcohols (cetearyl [CA] or decyl [DEC]), as well as prolipid 161 [PL161] (see details in (25); online Supplemental Figure 1). These materials formed two uniformly distributed physical compartments with distinct (several orders of magnitude different) proton diffusion rates, resulting in apparent water diffusion (D_{app}) and apparent kurtosis (K_{app}) at high *b*-values. Major differences among the tested chemical designs were in the physical origin of restricted diffusion for lamellar structures versus vesicular phase materials (25) (see online Supplemental Figure 1). Three negative control, monoexponential diffusion samples, were included based on polyvinylpirrolidone (PVP) (26) solutions in water at 0%, 20%, and 40%. All seven phantom materials were individually housed in polypropylene vials (V1–V7) of 150 mm in length and 25 mm of diameter, in a circular arrangement, submerged in water bath in a 1L plastic jar. The chemical phantom sample assignments for V1–V7 vials are provided in Table 1. The example axial-plane $b = 0$ image of the phantom with vial (region of interest [ROI]) labels is shown in Figure 1A. Three identical phantom prototypes were prepared using the same material batch, labeled for consistent scan geometry (see online Supplemental Figure1), and shipped to each of the participating sites. The jars were filled with tap water on-site and scanned at ambient temperature.

Table 1. Sample-Specific Inter-scan (All-Site) Average Kurtosis Parameters With Test–Retest (Repeatability-Based) 95% Confidence Intervals (95% CI)

Vial#	Sample	D_{app} ± 95% CI	K_{app} ± 95% CI
V1	DEC-CTAB	0.71 ± 0.014	1.11 ± 0.017
V2	CA-BTAC	1.02 ± 0.022	1.69 ± 0.013
V3	PVP20%	1.27 ± 0.017	0.04 ± 0.013
V4	Water	2.16 ± 0.034	0.06 ± 0.021
V5	PVP40%	0.60 ± 0.012	0.08 ± 0.022
V6	PL161	1.11 ± 0.014	1.29 ± 0.009
V7	CA-CTAB	0.39 ± 0.013	0.84 ± 0.076

Multicenter iDKI Phantom Studies

The prototype quantitative iDKI phantoms were scanned at three Quantitative Imaging Network (27) centers on four MRI scanners (2 at 1.5 T and 3 T each) using shared scan protocol over a period of six months. Consistent with the clinical iDKI scan protocol, the phantom scan instructions prescribed single-shot echo-planar imaging (SS EPI) acquisition of 3 orthogonal axial DWI directions with 11 *b*-values ($b = 0$, 50, 100, 200, 500, 800, 1000, 1500, 2000, 2500, 3000 s/mm^2), using a 16-channel head-coil. Other nominal acquisition parameters included the following: field of view (FOV) = 220 × 220 mm^2, echo time/repetition time = shortest/10 000 ms. (Actual minimum echo time varied from 93 ms to 107 ms across system scans owing to differences in gradient settings). The acquired section of the phantom ranged between 3 and 8 slices (3–5 mm in thickness) for the sites. Minor deviations from nominal scan protocol parameters among the sites were allowed with no effect on repeatability

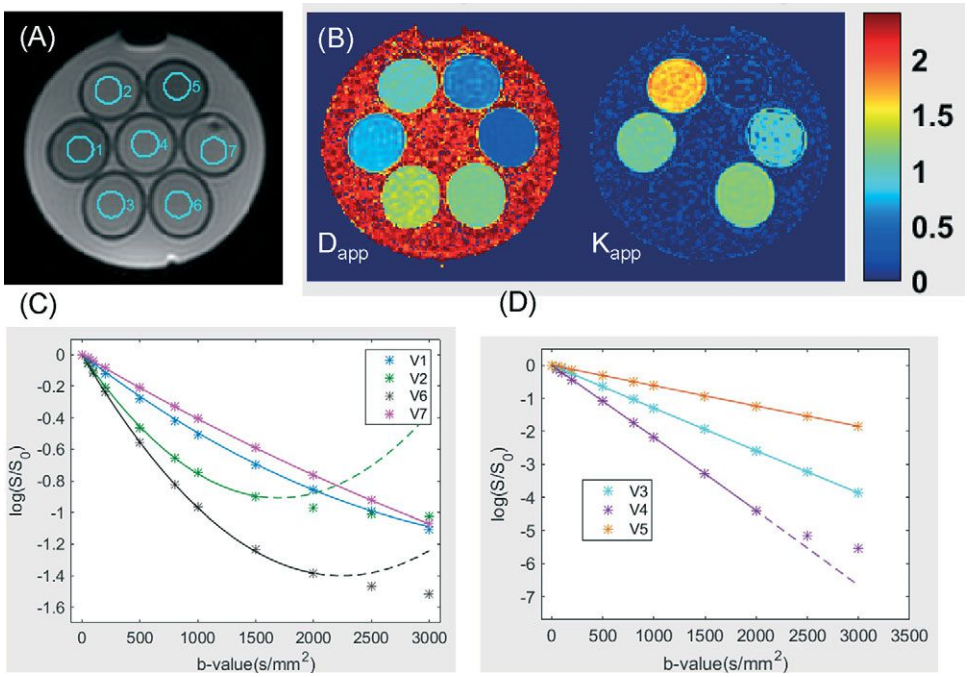


Figure 1. (A) Axial $b = 0$ image of the central slice of the kurtosis phantom acquired at 1.5T showing sample vial cross section with typical ROI placement. (B) example fit parametric diffusion kurtosis maps for (A) with the common color-bar indicating ($\times 10^{-3}$ mm^2/s) scale for D_{app} and dimensionless for K_{app}. (C) and (D) show b-value dependence of ROI-mean log-signal DWI (asterisks) and kurtosis model fit (traces) for kurtosis samples and negative (mono-exponential diffusion) controls, respectively. Sample vial data are color-coded in the legend corresponding to the ROI numbers in (A). Last DWI data point on the solid-fit curves indicates maximum *b*-value (b_{max}), allowed by the fit constraints of kurtosis model convergence in (C) and DWI noise floor in (D). Dashed segments for V2, V4, and V7 illustrate (unphysical) model extrapolation, prevented by constrained fit.

results. Test–retest acquisitions were performed with fixed scan protocol parameters with or without phantom repositioning, anywhere from several minutes to several days apart.

All acquired data were stored and distributed in Digital Image Communication in Medicine (DICOM) format (28), and centralized analysis of multi-*b* trace DWI DICOM data was performed using quality control routines developed in MATLAB 7 (MathWorks, Natick, MA) (20). Noncompliant scans from two dates that had large deviation in FOV (two scans) or had high EPI susceptibility artifacts (one scan), precluded uniform ROI definition and were excluded from the analysis. The remaining ten sets of test–retest data (three from each of the 3 T and two from each of the 1.5 T scanners) and four (early) single-run acquisitions (from one 3 T scanner) were analyzed. Test–retest studies were used for intra-scan repeatability assessment, while single runs were included for intra-scan reproducibility and sample stability evaluation. Phantom temperatures were not controlled and varied with the scanner room (ambient) environment. Reference scan room temperature was recorded for four (later) study scans. One site (that provided single-run acquisitions) stored the phantom in a scan room over the course of the study, while the other two allowed the phantom to thermally equilibrate in the scan room (for one 3 T and two 1.5 T systems) for <24 hours before each scan.

Parametric Map Generation and Repeatability Analysis

The parametric maps of apparent diffusion, D_{app}, and kurtosis, K_{app}, (Figure 1B) were calculated using linear least square fit of voxel DWI log-signal to a quadratic function of *b*-value, according to the iDKI model (11, 12), $Log(S_b/S_0) = -D_{app} \cdot b + K_{app}/6 \cdot (D_{app} \cdot b)^2$. Maximum *b*-value allowed in the fit was constrained by $b_{max} < 3/(K_{app} \cdot D_{app})$ to satisfy iDKI signal model convergence (11) and $S_{bmax}/S_0 > 0.01$ (to ensure $SNR_{bmax} > 2$). This yielded $b_{max} = 1500$ s/mm^2 for CA-BTAC (V2), and $b_{max} = 2000$ s/mm^2 for water (V4) and PL-161 (V6) vials (Figure 1, C and D). Absolute (residual) kurtosis bias of negative controls (Figure 1, A and D: V3, V4, and V5) was estimated as K_{app} fit parameter deviation from zero (29) for monoexponential (zero kurtosis) diffusion materials.

Uniform areas of the $b = 0$ image were used to define ROIs within phantom vials, for example, avoiding susceptibility and parallel imaging artifacts. Seven circular ROIs (12 mm diameter, 155 pixels) were defined on DWI ($b = 0$) for phantom tubes separately for the test–retest runs, using in-house MATLAB-based tools to generate ROI statistics for repeatability estimates of the D_{app} and K_{app} parameters. Uniform ROI definition was noted to be challenging for V7 owing to multiple small air bubbles (Figure 1A) apparently formed within the sample volume. These air bubbles were observed to "migrate" between test–retest runs. For all scans, the defined ROI pixel locations were within ±30 mm from the magnet isocenter that minimized potential contribution of gradient system and offset-dependent DWI bias (20, 21).

Sample-specific coefficient of variance (wCV) was calculated from available test–retest studies (18, 19): $wCV = \sqrt{2}/N \Sigma_{i=1}^{N} |X1 - X2|/(X1 + X2)$, where $X1$ and $X2$ were mean-

ROI test–retest (D_{app} or K_{app}) parameter values, respectively, for N repeatability studies. The 95% confidence interval (CI) for an average value of measured parameter (X), was estimated as $1.96 \cdot wCV \cdot ave(X)$, where the average was over all available (ten) test–retest DKI acquisitions (including less repeatable outliers) for each phantom vial. Single-acquisition 95% CI was also estimated for individual test–retest studies (N = 1) to assess systematic site and field dependencies. Bland–Altman (BA) repeatability analysis was performed for D_{app} and K_{app} across all test–retest samples (pool of 70). The overall BA limits of agreement (LOA) were calculated across all sample vials and test–retest scans excluding less repeatable scan "outliers." These "outliers" were identified on the basis of test–retest value differences >1.5 interquartile ranges above the upper quartile or below the lower quartile of the 70 sample test–retest parameter difference histogram, corresponding to $\pm 2.7 \times$ SD for the normal error distribution (defined according to MATLAB "*boxplot*" default outliers).

Pearson correlation, R, was evaluated for the derived mean parameter values and their corresponding 95% CI estimates versus scan time (days from phantom manufacturing), apparent (water ADC-based) phantom temperature, and system magnetic field, to characterize the sources of variation in the measured iDKI parameters and identify materials with desired properties. Among covariates, date was not correlated to temperature, allowing independent analysis, while magnetic field had significant negative correlation to temperature (−0.64; p_R = .02) as expected from dependence on scanner environment.

Water ADC-Based Apparent Phantom Temperature

Comprehensive characterization of thermal phantom properties was beyond the scope of this study; however, assessment of apparent phantom temperature (T_a) was deemed useful for discrimination between temporal and thermal origin of inter-scan variation in measured kurtosis parameters across multiple sites and dates. To this end, the T_a of each phantom scan was self-calibrated retrospectively using water diffusion coefficient based on Speedy–Angell relation (30): $T_a = 215.05 \cdot ([ADC/D_0]^{1/\gamma+1}) - 273.15$; $\gamma = 2.063$, $D_0 = 0.1635\ mm^2/s$; it ranged between 19.5°C and 25.5°C (±1°C) (Figure 2). For ADC-based T_a, water ADC was fit as a slope of log-signal DWI dependence on b-value up to b_{max} = 1000 s/mm^2 (to minimize SNR bias), and mean ADC value was measured from 15 × 15 mm^2 ROI defined on the central vial (V4, Figure 1A). ADC map vertical image "gradients" were observed for one system (online Supplemental Figure 2), with values increasing toward the posterior direction, indicative of phantom warming during the scan, possibly owing to contact from support pads or coil-induced heating. For this system, mean ADC values were used from three ROIs across the water-bath volume away from the posterior coil (see online Supplemental Figure 2).

Four independent, direct water temperature (T_m) measurements (with alcohol-based thermometer, CI = ±0.5°C) were recorded by the sites and indicated ~0.5°C positive bias of "apparent" T_a-values. (The ADC calculation using b-values up to 2000 s/mm^2 resulted in +1°C bias for the same independent T_m-measurements.) Notwithstanding the limited accuracy and precision of the utilized ADC-based T_a-calibration procedure (CI = ±1°C, owing to relatively imprecise water ADC values [$\pm 0.03 \times 10^{(-3)}$ mm^2/s]), the derived apparent temperature, T_a, was sufficient to differentiate thermal from temporal trends in the measured diffusion kurtosis parameters. Adequacy of the water ADC-based T_a-calibration procedure was confirmed by observation of (expected) linear temperature dependence for ADC of the negative control PVP samples (PVP20%: V3 and PVP40%: V5; Figure 2) not used for internal calibration. Minor excursions from linearity in Figure 2 for ADC values of PVP20%

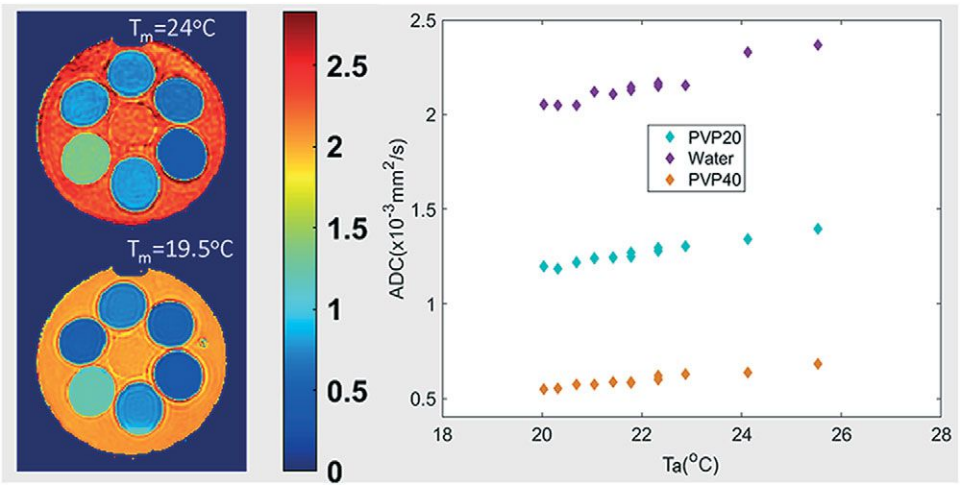


Figure 2. Left pane shows axial ADC maps (based on mono-exponential fit using b_{max} = 1000 s/mm^2) for phantoms scanned by two participating sites at different (measured) temperatures T_m = 24 and 19.5°C (recorded by sites). The common scale for the ADC maps is indicated by the color bar. Change in water ADC contrast between two maps illustrates sufficient thermal sensitivity for self-calibration of "apparent" phantom temperature, T_a (as described in Methodology). Linear ADC dependence on T_a is observed in the right pane for mean-ROI values of the negative control samples (color-coded in legend) from all analyzed scans (different sites and dates).

(V3) and PVP40% (V5) vials offset from the isocenter compared with the centrally positioned water (V4) (Figure 1), further confirmed the negligible effect of scanner gradient system bias (20, 21) on inter-scan variability for the measured diffusion parameters.

RESULTS

Four different chemical designs tested for iDKI phantom materials in V1, V2, V6, and V7 (Table 1) exhibited restricted diffusion at high *b*-values (>1000 s/mm^2), with DWI signals sustained above 20% of S_0 (Figure 1C), and apparent kurtosis coefficient exceeding negative control bias owing to background noise (Figure 1B, K_{app}). All materials allowed achievement of physiologically relevant apparent kurtosis parameter values (K_{app} ranges, 0.8–1.7; Table 1). Consistent qualitative observations across sites were that phantom samples apparently degassed after 3–4 weeks from preparation. Less viscous materials formed large air bubbles outside the sample volume, while more viscous materials formed small visible bubbles within the sample volume (Figure 1A; V7). The in-volume microbubbles tended to migrate between test–retest runs, potentially contributing fluctuating measurement errors owing to susceptibility artifact.

BA analysis across all test–retest acquisitions and samples summarized in Figure 3 showed generally good agreement for apparent diffusion kurtosis parameters of all phantoms across centers, compared with those of negative controls. Excluding outliers, BA 95% LOAs were ±0.025 (×10^{-3} mm^2/s) for D_{app} and ±0.035 for K_{app}. Negligible positive bias of 0.005 was observed for D_{app}. This bias and lower repeatability for several D_{app} (V4) and K_{app} (V7) "outliers" (well outside the LOA) was likely because of finite noise floor interference (V4, high water ADC) and "migrating" air bubble artifacts (V7) for the corresponding test–retest scans.

Finite spread of the mean parameter values of each sample observed along the horizontal axis in Figure 3 reported on cross-system and cross-scan variability, further detailed for individual sample vials in Figure 4A. The scan-to-scan differences in D_{app} of negative controls (V3, V4, V5, diamonds) were fully explained by the dependence on scanner ambient temperature (Figure 2; $R > 0.97$, $p_R < 1e\text{-}5$). Absolute bias for K_{app} of negative control materials (Figure 2, "x", right axis) did not exceed 0.1 (without significant temperature dependence). The highest bias, independent of system (magnetic field), was observed for V5 (40%PVP sample) consistent with contrast-to-noise limits for this (low ADC) control. For V4, the bias was inversely dependent on the field strength (higher for 1.5 T Sys2 and Sys3), indicating its SNR origin. All measured K_{app} for kurtosis samples (V1, V2, V6, and V7) exceeded negative control (zero kurtosis) bias. The estimated single test–retest 95% CIs (Figure 4B) for iDKI phantom materials ranged between 0.0003 and 0.15 (median 0.015), and (except for V7: K_{app} and V4: D_{app} outliers) these were not significantly different for D_{app} versus K_{app} and 1.5 T (Sys2, Sys3) versus 3 T (Sys1, Sys4) systems. CI(D_{app}) (Figure 4B, diamonds, left axis) for V1 and V2 has shown minor correlation to measured D_{app} values ($R = 0.59$, 0.57; $p_R = .07, .09$), suggesting negligible contribution of model fit error to test–retest repeatability. For V2 sample, CI(D_{app}) was significantly correlated to temperature ($R = 0.67$, $p_R = .033$), indicating thermal noise sensitivity of this material. No other significant correlations were observed for the material-specific test–retest measurement errors ($p_R > 0.1$).

The mean iDKI parameter values and derived 95% CIs observed across sites and scans are summarized in Table 1 for individual phantom components (including less repeatable "outliers"). Except for the V7 outlier K_{app} (95%CI: 0.076), the apparent measurement precision of iDKI phantom parameters (CI[D_{app}]: 0.013–0.022 (×10^{-3} mm^2/s) and CI[K_{app}]: 0.009–0.017) was as good (or better) than that of the negative controls (0.012–0.034 (×10^{-3} mm^2/s) and 0.013–0.022). The achieved measurement precision was sufficient for analysis of systematic scan-to-scan variability sources for kurtosis phantom parameters (Figure 4A, V1, V2, V6, V7).

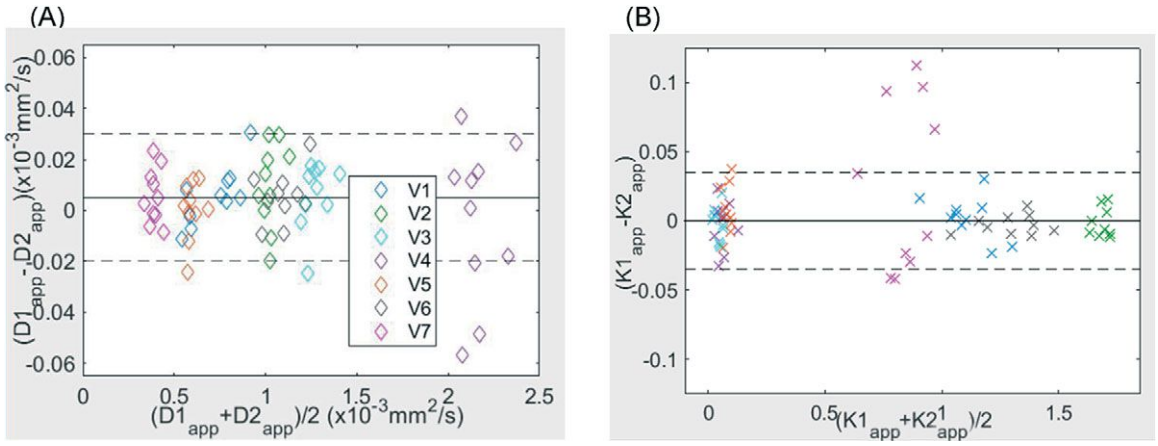


Figure 3. Bland–Altman (BA) plot for ROI-mean D_{app} (A) and K_{app} (B) fit parameters of all phantom samples (color-coded in the legend) from ten test–retest study scans. Solid and dashed horizontal lines mark mean bias and 95% LOA, respectively, across all samples (excluding outliers for V4 (A) and V7 (B)). Horizontal data spread for individual vials reflects inter-scan (temporal and thermal) variability of measured parameters.

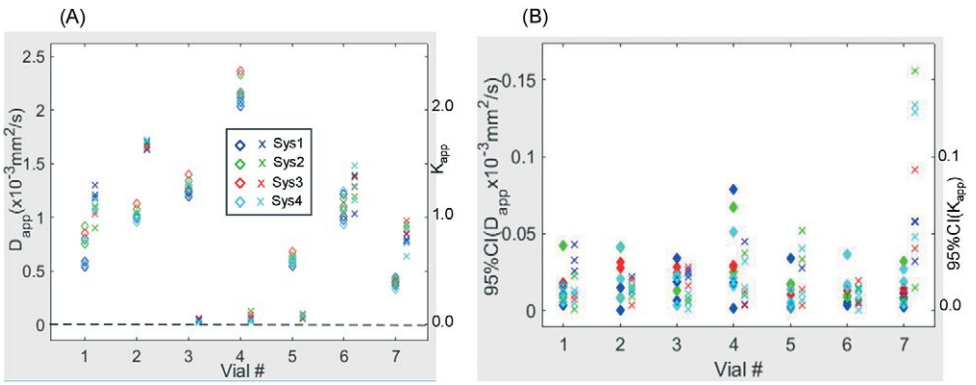


Figure 4. Vial-specific mean test–retest values for D_{app} (diamonds) and K_{app} ("x", horizontally offset for clarity) in (A) and corresponding (single test–retest) 95% CI in (B) are color-coded for system of origin, as indicated in the legend (Sys1, Sys4 are 3 T, and Sys2, Sys3 are 1.5 T). Left and right vertical axes are for D_{app} and K_{app},, respectively. The vertical spread of mean values reflects potential thermal, temporal, and field dependence of measured diffusion kurtosis parameters, while spread of CIs reports on test–retest measurement error.

Table 2 summarizes correlation between mean parameters and apparent scan room temperature (T_a), day and field variables. The bulk of the significant correlation to magnetic "field" observed for sample V1 (negative for D_{app}, and positive for K_{app}) apparently originated from the systematic kurtosis parameter differences observed for Sys1 phantom stored in the scanner room versus two other sites using prolonged storage outside of their scanners (Sys2, Sys3, Sys4). Unambiguous interpretation of significant correlation to magnetic field observed for K_{app} of V7 sample was not warranted owing to limited precision of the corresponding measurements (Table 1, CI[K_{app}] = 0.076).

For vials showing significant thermal and temporal correlations in Table 2, the corresponding parameter dependence is plotted in Figure 5. Temperature dependence was a significant contributor to 10%–15% variation in D_{app} (Figure 5A) of V2 and both "parallel" trends of V1 phantom materials. The deviations from linear trends were due to finite precision of self-calibrated T_a values and temporal stability. Marginally significant negative thermal correlation for K_{app} of V1 was evidently caused by 2 high-T_a > 24°C measurements for Sys2 and Sys3 (Figure 5B), when this viscous material might not have reached thermal equilibrium. Temporal D_{app} and K_{app} parameter value trends (Figure 5, C and D) for V1, V2, and V7 materials exhibit initial slope (D_{app} V1: 9%, V7: −22%; K_{app} V1: −18%, V2: 6%) which settled into relatively stable values after a 3- to 4-week stabilization period (coincidental with observed active sample degassing). In contrast, V6 (PL161 lamellar phantom) diffusion kurtosis parameter values continued to drift toward 50%-higher K_{app} and 20%-lower D_{app} parameter values over the whole study period (without significant T_a-dependence). Interestingly, D_{app} of V7 sample was also nominally independent of temperature. The site-dependent ~0.2 ($\times 10^{-3}$ mm^2/s) "fork" in D_{app} was observed consistently for thermal and temporal dependence of V1, strongly suggesting involvement of phantom storage conditions and/or low thermal conductivity of this viscous material.

DISCUSSION

All four of the different chemical designs evaluated for the prototype iDKI phantom (25) provided quantitative diffusion characteristics which could be tuned to a physiologically relevant range of parameters (K_{app} > 0.8) observed for in vivo tumor tissue, for example, for head and neck, prostate, and bladder cancers (5, 13-15). The study confirmed feasibility of quantitative iDKI phantoms based on vesicular and lamellar phases of liquid crystal

Table 2. D_{app} and K_{app} Percent-Correlation (%R) Summary for Kurtosis Phantom Materials (V1, V2, V6, V7)

Vial\%R (p_R)	(D_{app}, T_a)	(D_{app}, day)	(D_{app}, field)	(K_{app}, T_a)	(K_{app}, day)	(K_{app}, field)
V1	54.1 (0.046)	64.1 (0.013)	−76.9 (0.0013)	−53.8 (0.047)	−70.7 (0.0047)	71.2 (0.0043)
V2	86.5 (<0.001)	−3.9 (0.89)	−42.9 (0.13)	−23.4 (0.42)	−56.2 (0.036)	12.8 (0.66)
V6	25.1 (0.39)	−90.6 (<0.001)	6.8 (0.82)	9.5 (0.75)	98.3 (3E-10)	−30.5 (0.29)
V7	13.3 (0.65)	−84.5 (<0.001)	10.7 (0.72)	32.8 (0.25)	−5.1 (0.86)	-61.8 (0.019)

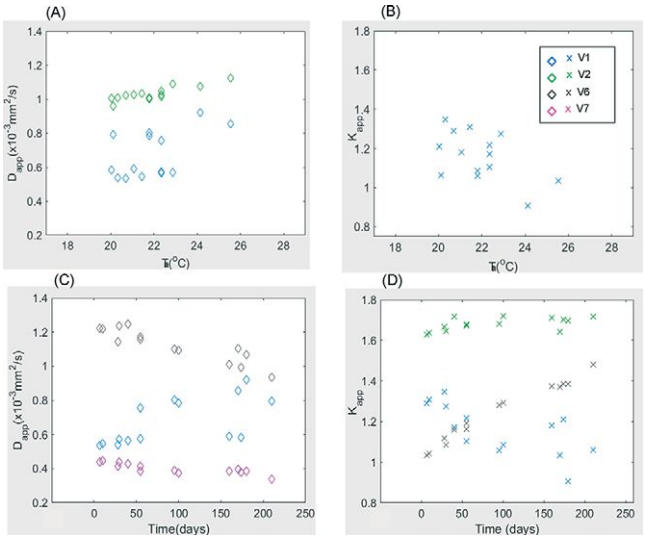


Figure 5. Thermal (A, B) and temporal (C, D) dependence of mean D_{app} (A, C, diamonds) and K_{app} (B, D, 'x") for phantom materials with significant ($p_R < 0.05$, Table 2) correlation to T_α/Time variables. Source vial number is color-coded in the legend (consistent with previous figures). The vertical data spread in each plot is indicative of cross-dependence on alternative variable (T_α in C, D or Time in A, B).

materials of different viscosity, and provided guidance toward a phantom product and multisite quality control protocol with improved precision and reproducibility. Included negative control samples allowed independent characterization of kurtosis bias and supplied internal standards for thermal diffusion-based calibration. Independent of chemical design, all kurtosis phantoms allowed sufficient SNR to avoid noise-bias or noise-limited precision in measured parameters. Phantom material-specific confidence intervals, derived from test–retest repeatability measurements, depended on sample preparation and handling more than on scan SNR, indicating possible improvement venue. The achieved model parameter precision (95% CI) of 1%–3.5% was sufficient to study sources of systematic inter-scan variability related to thermal and temporal stability of the prototype phantom materials. These results will be used in future studies to improve development of the next-generation quantitative iDKI phantom for utilization in multicenter clinical trials.

Among studied chemical designs, CA-BTAC (V2) phantom has shown the most promising characteristics and was least sensitive to sample preparation (6% D_{app} change during stabilization stage). Owing to thermal sensitivity of D_{app} (~3%/°C), typical of water-based phantoms (30), this iDKI phantom should be best used with temperature control or monitoring. In contrast, moderately viscous CA-CTAB (V7) phantom has shown thermally stable parameters, but large (22%) change in D_{app} during initial stabilization period, as well as limited K_{app} precision (9%, likely owing to migrating in-volume microbubbles). The observed field dependence of its K_{app} might be related to chemical properties of the material; however, this would require further investigation with improved precision. More viscous lamellar DEC-CTAB (V1) material exhibited moderate (9%–18%) kurtosis parameter changes during stabilization and moderate thermal sensitivity (10%), but was sensitive to prolonged storage and thermal equilibrium conditions, likely due to lower thermal conductivity. The stable kurtosis parameter values were not achieved for PL161 (V6) sample and continually changed over the course of the study, reflecting poor temporal stability of this material.

A limitation of this study was that all phantoms shared among sites were prepared from a single batch of (four families of) materials; repeatability of the batch preparation procedure itself was not evaluated. Temperature was not consistently monitored during scanning, which should be implemented for future multicenter studies, for example, by including in situ thermometer. The phantom T1 and T2 relaxation properties were not studied, and likely do not match in vivo tissue characteristics. However, having longer-than-tissue T2 relaxation times could be desirable for intended use of the kurtosis phantoms to increase the range of accessible *b*-values for DKI protocol optimization. Furthermore, adjustment of relaxation properties for vesicular phase (predominantly water) materials by adding relaxivity agents should be possible without substantial interference with diffusion characteristics.

Overall, observed apparent phantom diffusion sensitivity to temperature (2%–3%/°C) was similar to free water diffusion (30) and markedly higher than that of apparent kurtosis, consistent with the restricted diffusion origin of the latter. All phantom materials were noted to undergo initial parameter stabilization period of 3-4 weeks following preparation, coincidental with evident sample degassing. The parameter values for vesicular phase materials remained relatively stable after stabilization period. The candidate materials based on more viscous multilamellar vesicle phase, exhibited either poor temporal stability (PL161: V6) or notable dependence on site storage and thermal equilibrium conditions (DEC-CTAB: V1), and hence are not recommended for product iDKI phantom manufacturing. The kurtosis parameter values of CA-CTAB (V7) vesicular material had limited precision (9%) likely owing to formation of in-volume gas microbubbles, but warranted further evaluation after improved preparation due to offered thermal stability.

These observations suggested that sample degassing (eg, by centrifuging) during preparation should be attempted to improve precision and shorten stabilization period, preferably down to <1 week. The future studies should also monitor DKI parameter changes for up to a month for several material batches to evaluate reproducibility of phantom preparation and stabilization time. For use of temperature-sensitive phantoms, temperature monitoring (with DKI parameter calibration) is also recommended for multisite reproducibility studies at ambient temperature. Temperature monitoring could be implemented using an in situ thermometer or a calibrated internal standard, and it would be preferred to temperature control (eg, with ice-water bath) to avoid kurtosis phantom material phase transition (to gel) at lower temperatures.

CONCLUSION

The present multisite repeatability study has identified the liquid crystal materials based on vesicular phase as best candidates for quantitative iDKI phantom production. Independent of chemical design, the preparation procedure for iDKI phantoms could be improved by including degassing step to enhance repeatability and reduce stabilization period of diffusion kurtosis character-

istics. The most promising iDKI phantom design recommended for multisite trials is based on CA-BTAC (V2) vesicular suspension that allowed easy preparation, temporal stability, and independence of storage. Before utilization in multisite studies, this phantom would require temperature calibration and monitoring owing to observed thermal sensitivity of diffusion (similar to other water-based phantoms). Another iDKI phantom design (based on CA-CTAB: V7) with desirable thermal stability, needs to be studied after improved preparation to enhance precision and allowed longer thermal equilibration before scanning to ensure reproducibility for adaption in future longitudinal multicenter clinical trials.

Supplemental Materials

Supplemental Figure 1: http://dx.doi.org/10.18383/j.tom.2018.00030.sup.01

Supplemental Figure 2: http://dx.doi.org/10.18383/j.tom.2018.00030.sup.02

ACKNOWLEDGMENTS

This research was supported by National Institutes of Health Grants: U01CA166104, U01 CA211205, P01CA085878 and in part through P30 CA008748.

Disclosure: S.D. Swanson, D.I. Malyarenko, and T.L. Chenevert are co-inventors on intellectual property assigned to and managed by the University of Michigan for the technology underlying the manufacturing of the quantitative diffusion kurtosis phantoms utilized in this manuscript.

Conflict of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. Le Bihan D. Molecular diffusion, tissue microdynamics and microstructure. NMR Biomed. 1995;8:375–386.
2. Manenti G, Di Roma M, Mancino S, Bartolucci DA, Palmieri G, Mastrangeli R, Miano R, Squillaci E, Simonetti G. Malignant renal neoplasms: correlation between ADC values and cellularity in diffusion weighted magnetic resonance imaging at 3 T. Radiol Med. 2008;113:199–213.
3. Squillaci E, Manenti G, Cova M, Di Roma M, Miano R, Palmieri G, Simonetti G. Correlation of diffusion-weighted MR imaging with cellularity of renal tumours. Anticancer Res. 2004;24:4175–4179.
4. Koh DM, Collins DJ. Diffusion-weighted MRI in the body: applications and challenges in oncology. AJR Am J Roentgenol. 2007;188:1622–1635.
5. Lawrence EM, Gnanapragasam VJ, Priest AN, Sala E. The emerging role of diffusion-weighted MRI in prostate cancer management. Nat Rev Urol. 2012;9:94–101.
6. Levy A, Medjhoul A, Caramella C, Zareski E, Berges O, Chargari C, Boulet B, Bidault F, Dromain C, Balleyguier C. Interest of diffusion-weighted echo-planar MR imaging and apparent diffusion coefficient mapping in gynecological malignancies: a review. J Magn Reson Imaging. 2011;33:1020–1027.
7. Foltz WD, Wu A, Chung P, Catton C, Bayley A, Milosevic M, Bristow R, Warde P, Simeonov A, Jaffray DA, Haider MA, Ménard C. Changes in apparent diffusion coefficient and T2 relaxation during radiotherapy for prostate cancer. J Magn Reson Imaging. 2013;37:909–916.
8. Jensen LR, Garzon B, Heldahl MG, Bathen TF, Lundgren S, Gribbestad IS. Diffusion-weighted and dynamic contrast-enhanced MRI in evaluation of early treatment effects during neoadjuvant chemotherapy in breast cancer patients. J Magn Reson Imaging. 2011;34:1099–1109.
9. Lacognata C, Crimi F, Guolo A, Varin C, De March E, Vio S, Ponzoni A, Barilà G, Lico A, Branca A, De Biasi E, Gherlinzoni F, Scapin V, Bissoli E, Berno T, Zambello R. Diffusion-weighted whole-body MRI for evaluation of early response in multiple myeloma. Clin Radiol. 2017;72:850–857.
10. Kallehauge JF, Tanderup K, Haack S, Nielsen T, Muren LP, Fokdal L, Lindegaard JC, Pedersen EM. Apparent Diffusion coefficient (ADC) as a quantitative parameter in diffusion weighted MR imaging in gynecologic cancer: Dependence on b-values used. Acta Oncol. 2010;49:1017–1022.
11. Jensen JH, Helpern JA. MRI quantification of non-Gaussian water diffusion by kurtosis analysis. NMR Biomed. 2010;23:698–710.
12. Jensen JH, Helpern JA, Ramani A, Lu H, Kaczynski K. Diffusional kurtosis imaging: the quantification of non-gaussian water diffusion by means of magnetic resonance imaging. Magn Reson Med. 2005;53:1432–1440.
13. Jansen JF, Stambuk HE, Koutcher JA, Shukla-Dave A. Non-Gaussian analysis of diffusion-weighted MR imaging in head and neck squamous cell carcinoma: a feasibility study. AJNR Am J Neuroradiol. 2010;31:741–748.
14. Raab P, Hattingen E, Franz K, Zanella FE, Lanfermann H. Cerebral gliomas: diffusional kurtosis imaging analysis of microstructural differences. Radiology. 2010;254:876–881.
15. Rosenkrantz AB, Padhani AR, Chenevert TL, Koh DM, De Keyzer F, Taouli B, Le Bihan D. Body diffusion kurtosis imaging: Basic principles, applications, and considerations for clinical practice. J Magn Reson Imaging. 2015;42:1190–1202.
16. Grech-Sollars M, Hales PW, Miyazaki K, Raschke F, Rodriguez D, Wilson M, Gill SK, Banks T, Saunders DE, Clayden JD, Gwilliam MN, Barrick TR, Morgan PS, Davies NP, Rossiter J, Auer DP, Grundy R, Leach MO, Howe FA, Peet AC, Clark CA. Multi-centre reproducibility of diffusion MRI parameters for clinical sequences in the brain. NMR Biomed. 2015;28:468–485.
17. Jerome NP, Miyazaki K, Collins DJ, Orton MR, d'Arcy JA, Wallace T, Moreno L, Pearson AD, Marshall LV, Carceller F, Leach MO, Zacharoulis S, Koh DM. Repeatability of derived parameters from histograms following non-Gaussian diffusion modelling of diffusion-weighted imaging in a paediatric oncological cohort. Eur Radiol. 2017;27:345–353.
18. Barnhart HX, Barboriak DP. Applications of the repeatability of quantitative imaging biomarkers: a review of statistical analysis of repeat data sets. Trans Onc. 2009;2:231–235.
19. Raunig DL, McShane LM, Pennello G, Gatsonis C, Carson PL, Voyvodic JT, Wahl RL, Kurland BF, Schwarz AJ, Gönen M, Zahlmann G, Kondratovich MV, O'Donnell K, Petrick N, Cole PE, Garra B, Sullivan DC; QIBA Technical Performance Working Group. Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment. Stat Methods Med Res. 2015;24:27–67.
20. Malyarenko D, Galban CJ, Londy FJ, Meyer CR, Johnson TD, Rehemtulla A, Ross BD, Chenevert TL. Multi-system repeatability and reproducibility of apparent diffusion coefficient measurement using an ice-water phantom. J Magn Reson Imaging. 2013;37:1238–1246.
21. Mulkern RV, Ricci K, Vajapeyam S, Chenevert TL, Malyarenko DI, Kocak M, Poussaint TY. Pediatric Brain Tumor Consortium multi-site assessment of apparent diffusion coefficient z-Axis variation assessed with an ice water phantom. Acad Radiol. 2015;22:363–369.
22. Fieremans E, Pires A, Jensen JH. A simple isotropic phantom for diffusional kurtosis imaging. Magn Reson Med. 2012;68:537–542.
23. Phillips J, Charles-Edwards GD. A simple and robust test object for the assessment of isotropic diffusion kurtosis. Magn Reson Med. 2015;73:1844–1851.
24. Portakal ZG, Shermer S, Jenkins C, Spezi E, Perrett T, Tuncel N, Phillips J. Design and characterization of tissue-mimicking gel phantoms for diffusion kurtosis imaging. Med Phys. 2018;45:2476–2485.
25. Swanson S, Malyarenko D, Paudyal R, LoCastro E, Jambawalikar S, Schwartz L, et al. Design and Development of a Novel Phantom to Assess Quantitative Diffusion Kurtosis Imaging, a Multi-Site Initiative. Proceedings Quantitative Imaging Networ Annual Face-to-Face meeting; 2018, Bethesda, MD.
26. Pierpaoli C, Sarlls J, Nevo U, Basser PJ, Horkay F, editors. Polyvinylpyrrolidone (PVP) water solutions as isotropic phantoms for diffusion MRI studies. Proceedings Intl Soc Mag Reson Med; 2009; Honolulu HI.
27. NCI Quantitative Imaging Network (QIN). (last accessed 27 Sept 2018) https://imaging.cancer.gov/programs_resources/specialized_initiatives/qin/about/default.htm
28. Clunie DA. DICOM structured reporting and cancer clinical trials results. Cancer Inform. 2007;4:33–56.
29. Scheel M, Hubert AAB, editors. Diffusion kurtosis imaging and Pseudokurtosis in phantom studies. Eur Soc Radiol 2015.
30. Holz M, Heil SR, Sacco A. Temperature-dependent self-diffusion coefficients of water and 6 selected molecular liquids for calibration in accurate H-1 NMR PFG measurements. Phys Chem Chem Phys. 2000;2:4740–4742.

TOMOGRAPHY® RESEARCH ARTICLE

Magnetization Transfer MRI of Breast Cancer in the Community Setting: Reproducibility and Preliminary Results in Neoadjuvant Therapy

John Virostko[1,2,4], Anna G. Sorace[1,2,3,4], Chengyue Wu[3], David Ekrut[5], Angela M. Jarrett[2,5], Raghave M. Upadhyaya[3], Sarah Avery[6], Debra Patt[7], Boone Goodgame[8,9], and Thomas E. Yankeelov[1,2,3,4,5]

[1]Department of Diagnostic Medicine, [2]Livestrong Cancer Institutes, [3]Department of Biomedical Engineering, [4]Department of Oncology, [5]Institute for Computational Engineering and Sciences, University of Texas at Austin, Austin, TX; [6]Austin Radiological Association, Austin, TX; [7]Texas Oncology, Austin, TX; [8]Seton Hospital, Austin, TX; and [9]Department of Internal Medicine, University of Texas at Austin, Austin, TX

Corresponding Author:
John Virostko, PhD
Department of Diagnostic Medicine, University of Texas at Austin, Dell Medical School, 1701 Trinity St., Stop C0200, Austin, TX 78712.
E-mail: jack.virostko@austin.utexas.edu

Key Words: MT-MRI, MTR, NAT, reproducibility, repeatability

Abbreviations: Neoadjuvant therapy (NAT), magnetization transfer ratio (MTR), magnetization transfer magnetic resonance imaging (MT-MRI), magnetic resonance imaging (MRI), repetition time (TR), echo time (TE), magnetization transfer (MT), pathological complete response (pCR), quantitative magnetization transfer (qMT)

ABSTRACT

Repeatability and reproducibility of magnetization transfer magnetic resonance imaging of the breast, and the ability of this technique to assess the response of locally advanced breast cancer to neoadjuvant therapy (NAT), are determined. Reproducibility scans at 3 different 3 T scanners, including 2 scanners in community imaging centers, found a 16.3% difference (n = 3) in magnetization transfer ratio (MTR) in healthy breast fibroglandular tissue. Repeatability scans (n = 10) found a difference of ~8.1% in the MTR measurement of fibroglandular tissue between the 2 measurements. Thus, MTR is repeatable and reproducible in the breast and can be integrated into community imaging clinics. Serial magnetization transfer magnetic resonance imaging performed at longitudinal time points during NAT indicated no significant change in average tumoral MTR during treatment. However, histogram analysis indicated an increase in the dispersion of MTR values of the tumor during NAT, as quantified by higher standard deviation ($P = .005$), higher full width at half maximum ($P = .02$), and lower kurtosis ($P = .02$). Patients' stratification into those with pathological complete response (pCR; n = 6) at the conclusion of NAT and those with residual disease (n = 9) showed wider distribution of tumor MTR values in patients who achieved pCR after 2–4 cycles of NAT, as quantified by higher standard deviation ($P = .02$), higher full width at half maximum ($P = .03$), and lower kurtosis ($P = .03$). Thus, MTR can be used as an imaging metric to assess response to breast NAT.

INTRODUCTION

Magnetization transfer magnetic resonance imaging (MT-MRI) is sensitive to the macromolecular content of tissue, providing a contrast mechanism that differs from conventional magnetic resonance imaging (MRI) relaxation measurements such as T1 and T2. This macromolecular content includes contributions from a variety of biomolecules. For example, in white matter, the tissue for which MT-MRI has been best characterized, the macromolecular content is considered to include contributions from cholesterol, sphingomyelin, and galactocerebroside (1). The macromolecules that contribute to MT-MRI in cancer have not been fully described, although it has been postulated that increased proteolytic activity or decreased enzyme inhibition may play a role (2). The protons on these macromolecules are difficult to image directly owing to their fast transverse relaxation, but their effects can be observed indirectly by perturbing the macromolecular pool and imaging the effect on free water protons. More specifically, the image contrast in MT-MRI reflects the exchange of magnetization between protons in free water and protons bound to semisolid macromolecules through dipole–dipole interactions and/or chemical exchange. Since first reported by Wolff and Balaban (3), MT-MRI has been used extensively as a research tool in neuroimaging (1, 4), with notable success in assessing the demyelination process accompanying multiple sclerosis (5).

Compared with progress in neuroimaging, MT-MRI has been relatively underexplored in cancer imaging. A study of excised breast tissue found that magnetization transfer (MT) saturation improved the discrimination between healthy and malignant tissue (6). Further in vivo studies in the breast found that MT-MRI improved conspicuity of breast lesions (7) and distinguished malignant and benign lesions (2, 8). Although the

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2018.00019

biochemical basis for MT-MRI contrast in tumors has not been fully explored, these studies suggest that MT may reflect some aspects of malignant tissue, putatively the extracellular matrix, which has garnered increased recent attention for its role in tumor formation, growth, and metastasis (9). In breast cancer, in particular, the extracellular matrix has been implicated as a crucial driver of tumor progression and metastasis, as well as a potent mediator of treatment resistance (10).

MRI has shown the capability of characterizing changes in the tumor and tumor microenvironment that are associated with therapy. In breast cancer, MRI performed early in the course of neoadjuvant therapy (NAT) has proven capable of predicting the eventual tumor response before downstream changes in tumor size (11-13). The 2 MRI techniques that have been the most investigated for predicting therapeutic response to breast NAT are diffusion-weighted MRI (14) and dynamic contrast-enhanced MRI (15, 16), as well as their combination (17, 18). MT-MRI has not yet been investigated for evaluating (or predicting) response during therapy in locally advanced breast cancer.

In this study, we first characterize the repeatability and reproducibility of MT-MRI in healthy breast fibroglandular tissue (FGT). Then in a pilot cohort of women with locally advanced breast cancer, we investigate changes in MT-MRI in response to NAT and correlate these changes with surgical pathology results. Importantly, these studies are performed on MRI scanners sited in community imaging centers, showing that MT-MRI can be deployed beyond academic research centers and into routine clinical practice.

METHODOLOGY

MRI Protocol

MRI was performed using 3 T Siemens Skyra scanners (Erlangen, Germany) equipped with 8- or 16-channel receive double-breast coil (Sentinelle, Invivo, Gainesville, Florida). Three scanners were used in this study: 2 were located at community imaging facilities, while 1 was sited at an academic research facility. Repeatability studies were performed on the scanner sited at the academic research facility, while the normal subject reproducibility experiment was performed on all 3 scanners. The study in patients with breast cancer was performed at the 2 community imaging facilities.

All breast MRI data were acquired in the sagittal plane. To calculate the magnetization transfer ratio, 2 images were acquired, identical, save for the inclusion of an MT saturation pulse on 1 of the acquisitions. Both images consisted of spoiled gradient-echo sequences with repetition time (TR) = 48 milliseconds, echo time (TE) = 6.4 milliseconds, flip angle = 6°, receiver bandwidth = 260 Hz/pixel, acquisition matrix = 192 × 192, field of view = 256 × 256 mm, number of sections = 10, section thickness (with no section gap) = 5 mm. GRAPPA (GeneRalized Autocalibrating Partial Parallel Acquisition) acceleration factor of 2 and SPAIR (SPectral Attenuated Inversion Recovery) fat suppression were performed. The MT saturation pulse consisted of a 9.88-millisecond Gaussian-shaped MT prepulse performed within each repetition, with a flip angle of 500°, which was offset from the water frequency peak by 1.5 kHz. The acquisition time was 53 seconds for each acquisition, yielding a total imaging time of 1 minute, 46 seconds to acquire data both with and without the MT pulse.

The MRI protocol also included a high-resolution T1-weighted 3D gradient-echo FLASH (fast low angle shot) acquisition for identifying anatomy and segmentation of fibroglandular and adipose tissue. The following are parameters of this anatomical image: TR = 5.3 milliseconds, TE = 2.3 milliseconds, flip angle = 20°, acquisition matrix = 256 × 256, FOV = 256 × 256 mm, section thickness = 1 mm, GRAPPA acceleration = 2, and SPAIR fat suppression. Acquisition time for the anatomical image was 3 minutes and 11 seconds. For patients with breast cancer, a dynamic contrast-enhanced MRI protocol was performed after the MT acquisition to segment the tumor. The dynamic contrast-enhanced protocol consisted of a T1-weighted VIBE (volumetric interpolated breath-hold examination; no breath-holding was, however, used in these studies) acquisition with TR = 7.02 milliseconds, TE = 4.6 milliseconds, flip angle = 6°, acquisition matrix = 192 × 192, field of view = 256 × 256 mm, number of sections = 30, slice thickness = 5 mm, GRAPPA acceleration factor = 3. Imaging was performed at 7.27-seconds temporal resolution for 1 minute before and 6 minutes after administration of a gadolinium-based contrast agent (Multihance; Bracco, Monroe Township, NJ) or Gadovist (Bayer, Leverkusen, Germany) via a power injector followed by a saline flush.

Repeatability/Reproducibility Study

Volunteers consisted of healthy women (n = 10; median age = 39.5 years [range = 22–62]) with no history of breast disease who were neither pregnant nor breastfeeding. For the repeatability study, 2 MRI examinations were performed on the same day on a single MRI scanner, with subject removal from the examination table and repositioning between scans. To assess reproducibility of MTR, subjects were scanned at 3 different locations (which included 1 academic and 2 community radiology centers) on 3 different days.

Response to Breast Cancer NAT Study

Women (n = 15; median age = 41 years [range = 25–63 years]) with stage II or III locally advanced breast cancer undergoing NAT were recruited from community oncology practices (n = 15). Longitudinal MTR measurements were performed at 4 time points: before the start of NAT, after 1 cycle of NAT, after 2–4 cycles of NAT, and 1 cycle after MRI #3.

Image Analysis

MTR was calculated for each voxel via equation (1):

$$MTR = 100\% \times \frac{S_0 - S_{MT}}{S_0} \qquad (1)$$

where S_{MT} and S_0 are the measured signal intensities with and without the MT saturation pulse, respectively. MTR values for voxels that returned undefined values (for which $S_0 = 0$) or for which $S_{MT} > S_0$ were excluded from subsequent analyses. All pixels with MTR values of 0 were excluded to minimize the number of residual pixels, with partial volume averaging from adipose tissue.

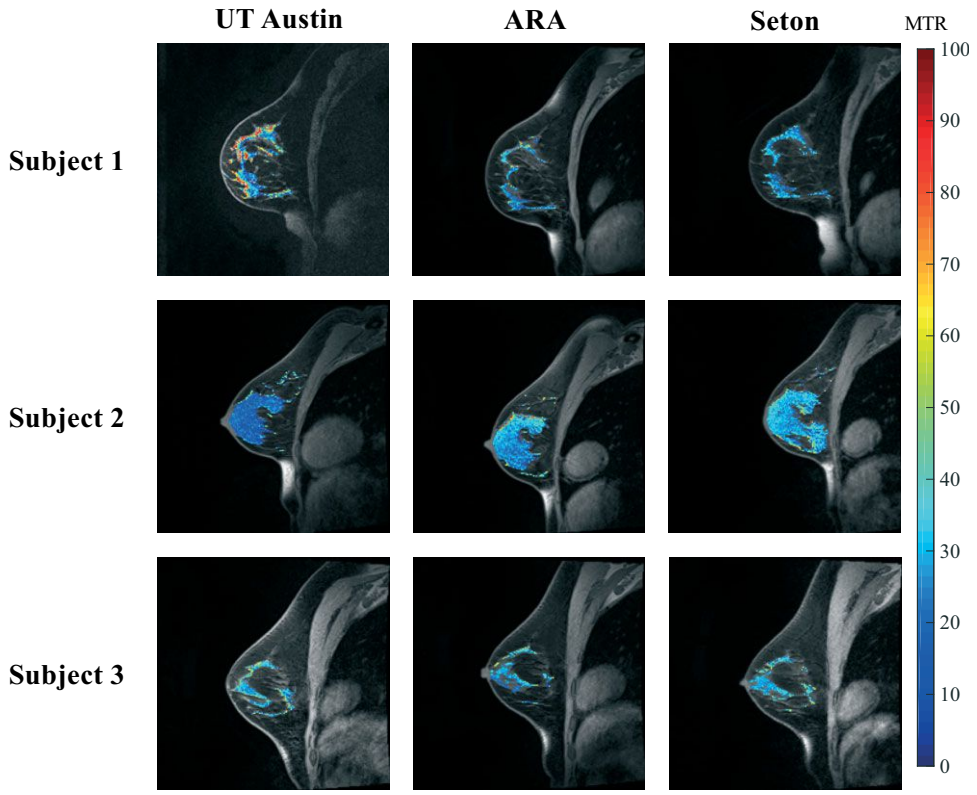


Figure 1. Multisite reproducibility of magnetization transfer ratio (MTR) in 3 normal subjects scanned at 3 sites. MTR values of fibroglandular tissue (FGT) are displayed in pseudo color and overlaid on top of an anatomical image.

Segmentation of FGT was performed on the anatomical images as previously described (20). Briefly, a k-means clustering algorithm was used to generate a mask specific to FGT based on image intensity. Segmentation of the tumor in the patient population with breast cancer was performed using dynamic contrast-enhanced images. A fuzzy c-means algorithm separated tumor from surrounding tissue based on the dynamics of signal intensity after contrast agent injection, similarly to a method previously described (21).

Statistical Analysis

Statistical analysis was performed using the statistical toolbox in MATLAB 2018a (The MathWorks, Natick, MA). Voxel values were averaged across the FGT, and repeatability/reproducibility statistics were calculated as previously described (20). Voxel values were both averaged across the tumor and binned into histograms for the therapeutic response study. Voxel distributions quantified the Kolmogorov–Smirnov test, and histogram analysis, similar to methods previously used to characterize MTR distributions in the brain (22). The following histogram parameters were derived from the histograms of MTR values: standard deviation; kurtosis; and the 25th, 75th, and 95th percentiles of the tumor MTR, where the nth percentile is the point at which n% of the voxel values that form the histogram are found to the left. The full width at half maximum, which was calculated using the "dfittool" in MATLAB to fit voxel values to a Gaussian distribution. To ensure that differences in voxel distributions were not due solely to tumor regression and thus a smaller number of voxels in the distribution, MTR distributions were truncated to be the same number of voxels and compared. Comparisons between 2 groups were made using a 2-tailed t test, with $P < .05$ considered significant. The voxel-wise distributions of MTR values were tested for equal variance using a 2-sample F-test to compare the pCR and non-pCR groups and Levene test to compare groups through time (multiple-sample test). Comparisons among >2 groups were made using 1-way ANOVA or, for measurements repeated in the same subject, repeated-measures ANOVA. Data are expressed as mean ± standard deviation.

RESULTS

Repeatability and Reproducibility

Reproducibility of MT-MRI in the breast was assessed by scanning normal breast FGT at the 3 different sites, yielding an average difference of 16.3% ± 14.4% in MTR between sites (Figure 1), with no statistical differences detected between the sites ($P = .1$). Repeatability was assessed by scanning the breast of the same woman twice with repositioning attempted between scans (Figure 2A). Repeatability scans of the same subject's FGT showed an average percent difference of 8.1% ± 7.9% in mean MTR measurement between the 2 scans. Repeatability of MTR in FGT was not significantly different between scans, suggesting lack of bias between the first and second scanning sessions (Figure 2B). In addition, the difference between repeated measurements was independent of the mean. The average MTR of FGT was 33% ± 5%. The difference between subject age and MTR was not significant ($P = .08$), but it did indicate a trend toward higher MTR values in younger subjects. The standard deviation of MTR values throughout the FGT was also assessed to determine how the voxel distribution varied across repeat scans (Figure 3B). The repeatability statistics for both mean and standard deviation of MTR are summarized in Table 1, which estab-

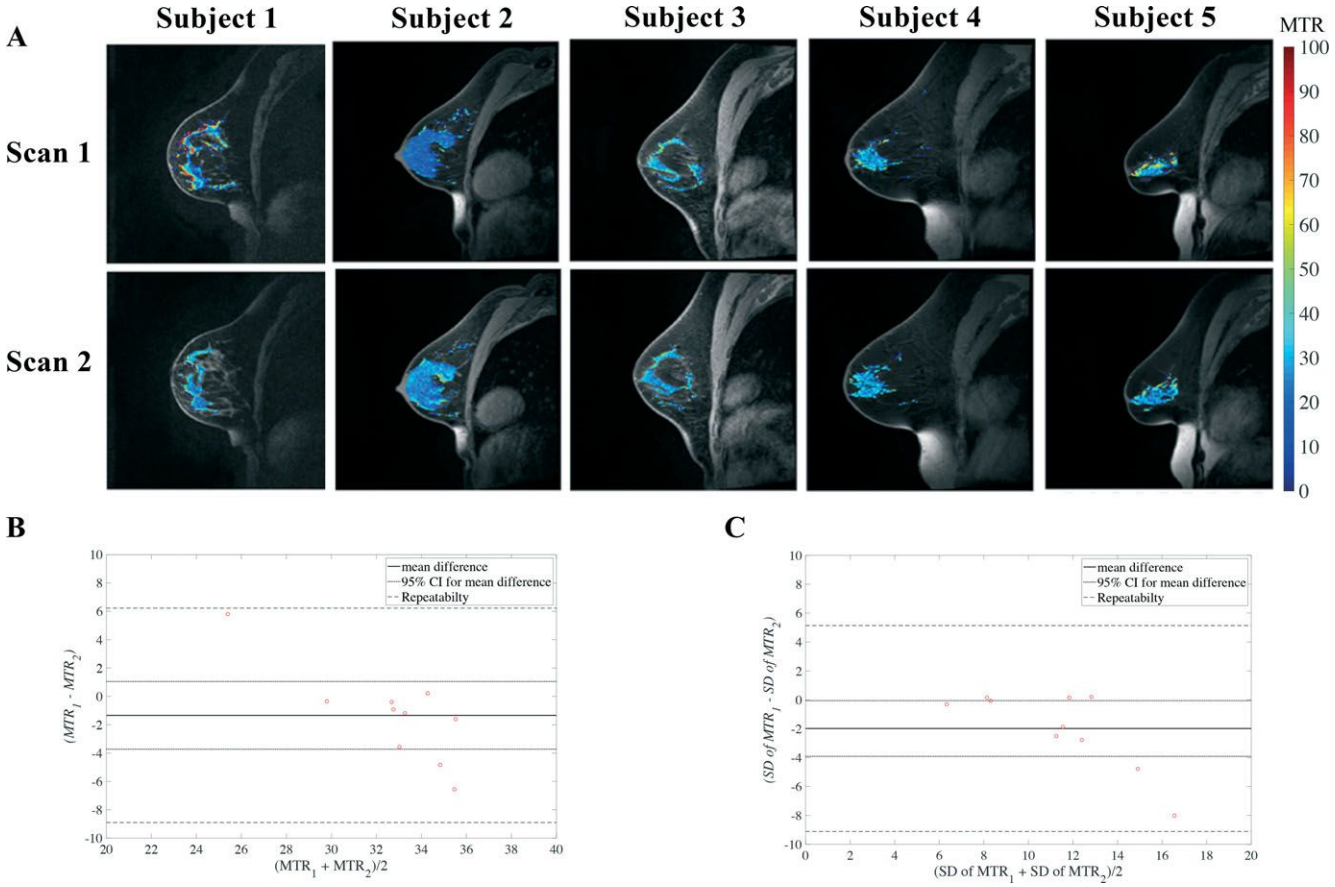


Figure 2. Repeatability of MTR maps of FGT in 5 women undergoing test–retest scanning with subject repositioning between scans (A). MTR maps of FGT are displayed in pseudo color and overlaid on top of an anatomical image. Bland–Altman plot of mean MTR repeatability in breast FGT (B). Bland–Altman plot of MTR standard deviation repeatability in breast FGT (C).

lish metrics for intraindividual variability of MTR in breast tissue.

Response to Neoadjuvant Therapy

To assess changes in MTR during treatment, longitudinal MT-MRI was performed on women before the start of NAT and at 3 time points during the course of NAT. Mean MTR values of the breast tumors before the start of therapy were higher than the MTR values of FGT in healthy controls (mean tumor MTR = 29% ± 1%; mean FGT MTR = 33% ± 5%; $P = .02$) and did not display any correlation with subject age ($P = .74$). Representative images from a subject who experienced partial response to NAT are shown in Figure 3A, with MTR values of the tumor overlaid in pseudo color. The volume of the tumor decreased from the first MRI, performed before the start of therapy, through the subsequent MRIs. Across all subjects receiving NAT, in comparison with the pretreatment MRI, the average tumor volume was 26% ± 37% smaller at the second scan session, 65% ± 28% smaller at the third scan session, and 68% ± 29% smaller at the fourth scan session ($P = .02$). Histograms of MTR values from all tumor voxels of all subjects are displayed in Figure 3B, which show increased dispersion through the course of therapy. Parameters characterizing tumor MTR distributions for all subjects are displayed in Table 2. There was no significant change in mean tumor MTR during therapy (Figure 4A; $P = .37$). However, the distribution of MTR values revealed an increase in the spread and relative distribution of extreme values as therapy progressed, with increases in standard deviation (Figure 4B; $P = .005$) and full width at half maximum (Figure 4C; $P = .02$) and a decrease in kurtosis (Figure 4D; $P = .02$). The Levene test indicated that the distribution of voxel-level MTR values had unequal variance during NAT ($P < .001$). There was no significant difference in the 25th ($P = .06$), 75th ($P = .06$), or 95th percentile ($P = .09$) of MTR distribution (Table 2).

To determine whether alterations in MTR in response to therapy were related with treatment efficacy, we separated the study participants into those who achieved a pCR (n = 6) and those who had residual disease at the conclusion of NAT (non-pCR, n = 9). Note that results from the fourth MRI time point are excluded from this analysis of results stratified by pathological response, as 5 of the subjects no longer had quantifiable residual tumor at the fourth MRI scan. Figure 5 presents example data sets for a patient who achieved pCR (Figure 5A) and 1 who had

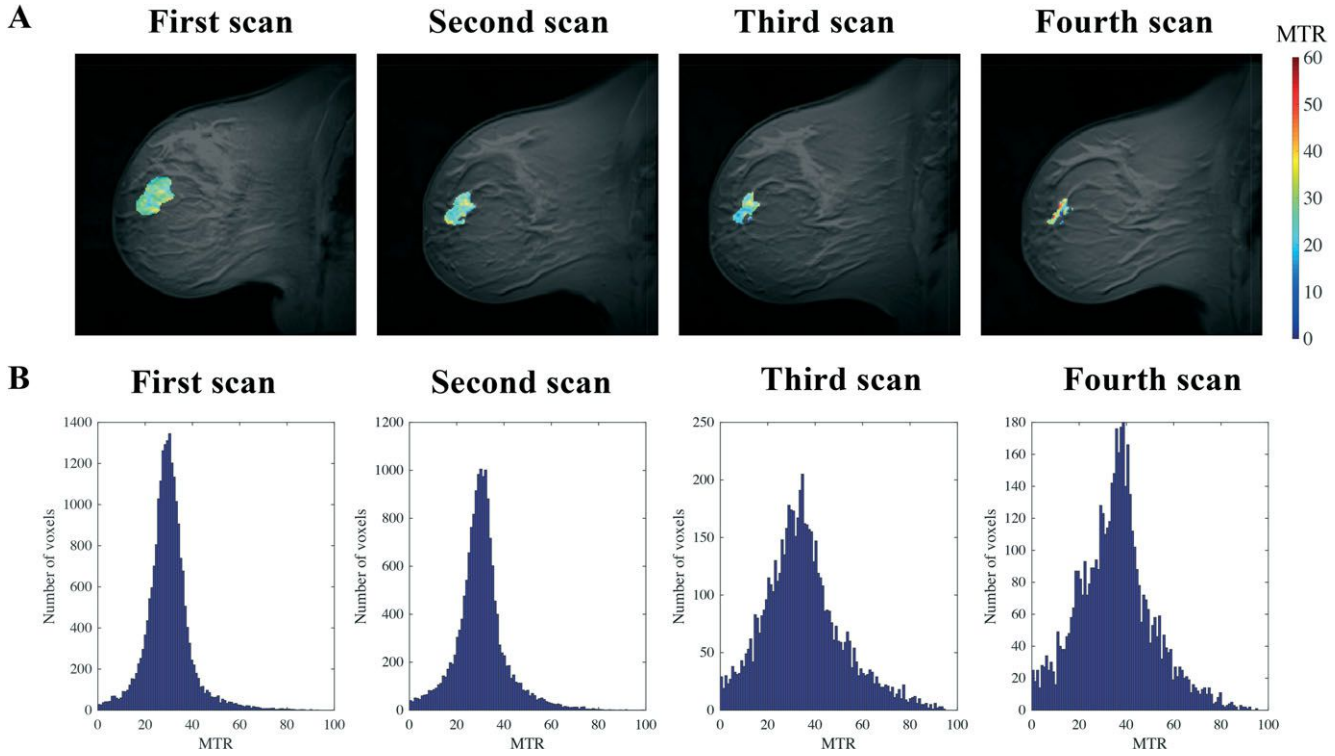


Figure 3. Representative MTR maps pseudo-colored and overlaid on an anatomical image of a subject who experienced a partial response to neoadjuvant therapy (NAT) (A). Histograms of voxel distributions of tumor MTR from all participants at the first, second, third, and fourth magnetic resonance imaging (MRI) sessions show higher dispersion at later time points after the start of treatment (B).

residual disease (Figure 5C). Of note, while both subjects display regression in tumor size, the patient who achieved pCR had a more heterogeneous MTR distribution throughout the tumor after NAT, as quantified by histogram analysis. This heterogeneity is shown for all patients achieving pCR and residual disease in Figure 5B and 5D, respectively; observe the increased spread in tumor MTR values in the patients achieving pCR after treatment compared with the tumor MTR values in non-pCR patients. The MTR value averaged over the entire tumor was similar in subjects achieving pCR and those with residual disease at the first (Figure 6A; $P = .44$), second (Figure 6B; $P = .07$), and third scan sessions (Figure 6C; $P = .22$). The heterogeneity in MTR values was quantified by standard deviation, showing similar standard deviation at the first (Figure 6D; $P = .19$) and second scan sessions (Figure 6E; $P = .98$), but larger standard deviation in patients ultimately achieving pCR who had viable tumor at the third scan session (Figure 6F; $P = .02$). In addition, the difference in standard deviation between the pCR and non-pCR cohorts (5.19) exceeds the 95% CI found in healthy FGT (1.92), indicating that the difference in standard deviation between the pCR and non-pCR cohorts is not due to intraindividual variation. Furthermore, 2 sample F tests comparing all voxel-level MTR values from subjects achieving pCR versus those who did not achieve pCR showed significantly different

Table 1. Repeatability Statistics for Normal Breast FGT (n = 10)

	Mean MTR	Standard Deviation of MTR
Kendall's tau, *P* (age vs mean)	0.08	0.06
Kendall's tau, *P* (difference vs mean)	0.236	0.04
95% CI (percentage of mean)	2.39, (7.30%)	1.92, (16.90%)
Root mean square deviation (percentage of mean)	3.44, (10.51%)	3.23, (28.36%)
Within-subject standard deviation (percentage of mean)	2.43, (7.43%)	2.29, (20.06%)
Repeatability value (*r*) (percentage of mean)	7.56, (23.13%)	7.12, (62.43%)

Table 2. Trends in MTR Parameters During NAT

	Average at Scan 1	Average at Scan 2	Average at Scan 3	Average at Scan 4	*P*-Value (from subjects with tumor at all 4 scans)
Mean	28 ± 5	27 ± 4	29 ± 5	31 ± 7	0.37
Standard deviation	11 ± 4	11 ± 4	13 ± 4	12 ± 4	0.005
FWHM	24 ± 9	26 ± 9	31 ± 10	27 ± 10	0.02
25th percentile	25 ± 6	22 ± 5	22 ± 9	27 ± 7	0.06
75th percentile	37 ± 9	36 ± 6	40 ± 11	41 ± 7	0.06
95th percentile	49 ± 11	48 ± 9	53 ± 14	53 ± 11	0.09
Kurtosis	5.59 ± 3.44	4.65 ± 2.88	3.17 ± 1.00	3.49 ± 1.26	0.02

The averages include all subject scans, while the *P*-value is from repeated-measures ANOVA, which includes only subjects with data at all 4 scans (n = 12).

variance at the first, second, and third scan sessions ($P < .05$, all scans). The kurtosis of the MTR distribution was similar at the first (Figure 6G; $P = .55$) and second scan sessions (Figure 6H; $P = .32$), but lower in the 4 patients who ultimately achieved pCR at the third scan session (Figure 6I; $P = .03$). The MTR values of the non-pCR group at the third MRI time point, when truncated to be the same number of voxels as the pCR group at the third MRI, showed a smaller standard deviation (0.09) than the pCR distribution (0.15), indicating that the increased spread in the pCR distribution is not solely due to a smaller sample size. Furthermore, the standard deviation of the truncated non-pCR distribution was similar to that of the full non-pCR distribution at the third MRI time point, suggesting that the smaller numbers of voxels in patients achieving pCR were not driving increased dispersion of MTR values.

DISCUSSION

To the best of our knowledge, this study is the first application of MT-MRI to assess the repeatability and reproducibility of the healthy breast tissue, as well as changes to breast tumors in response to therapy. Furthermore, this was accomplished in imaging clinics from the community setting (ie, not in the environment of an academic research center). In response to NAT, the distribution of tumor MTR values is more heterogeneous, with an increasing number of voxels exhibiting more extreme values of the MTR. Furthermore, the increased dispersion of MTR throughout the tumor is more pronounced in patients who display complete response to therapy, indicating that alterations in intratumoral MTR may reflect successful treatment response. Collectively, these findings indicate that MT-MRI may provide important information regarding tumor

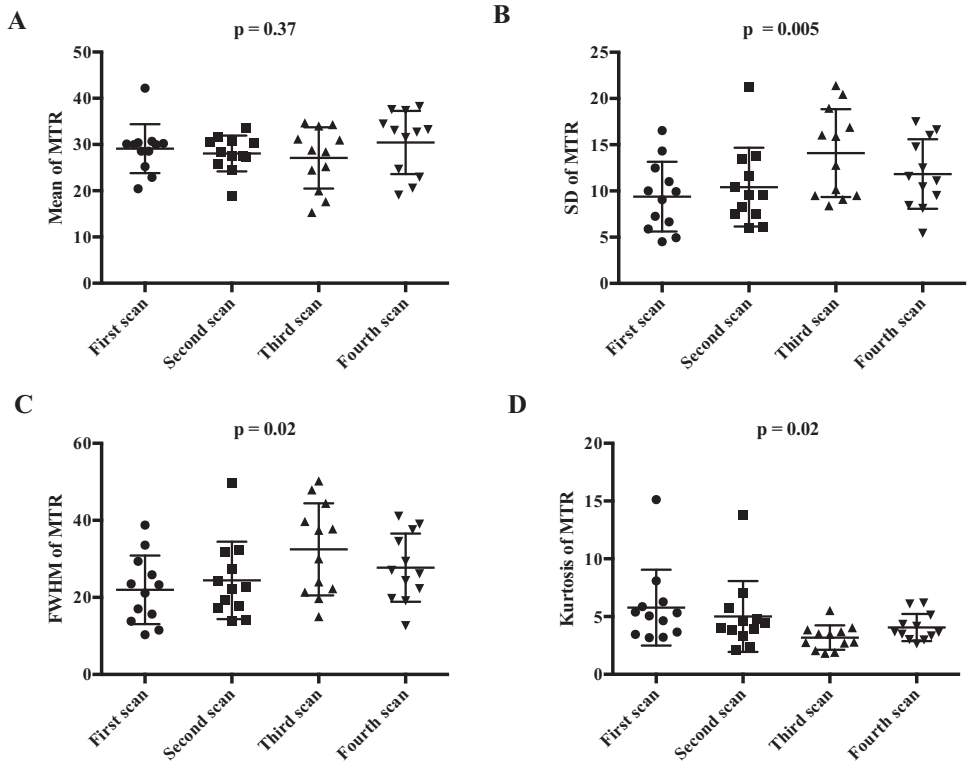


Figure 4. Mean tumoral MTR is similar across all subjects before therapy (first Scan) and at serial scans performed during the course of NAT (second, third, and fourth scans) (A). The standard deviation of tumor MTR values increases with longer duration of NAT (B). Full width half maximum (FWHM) of the distribution of tumor MTR values increases with longer duration of NAT (C). Kurtosis of the distribution of tumor MTR values decreases with longer duration of NAT (D).

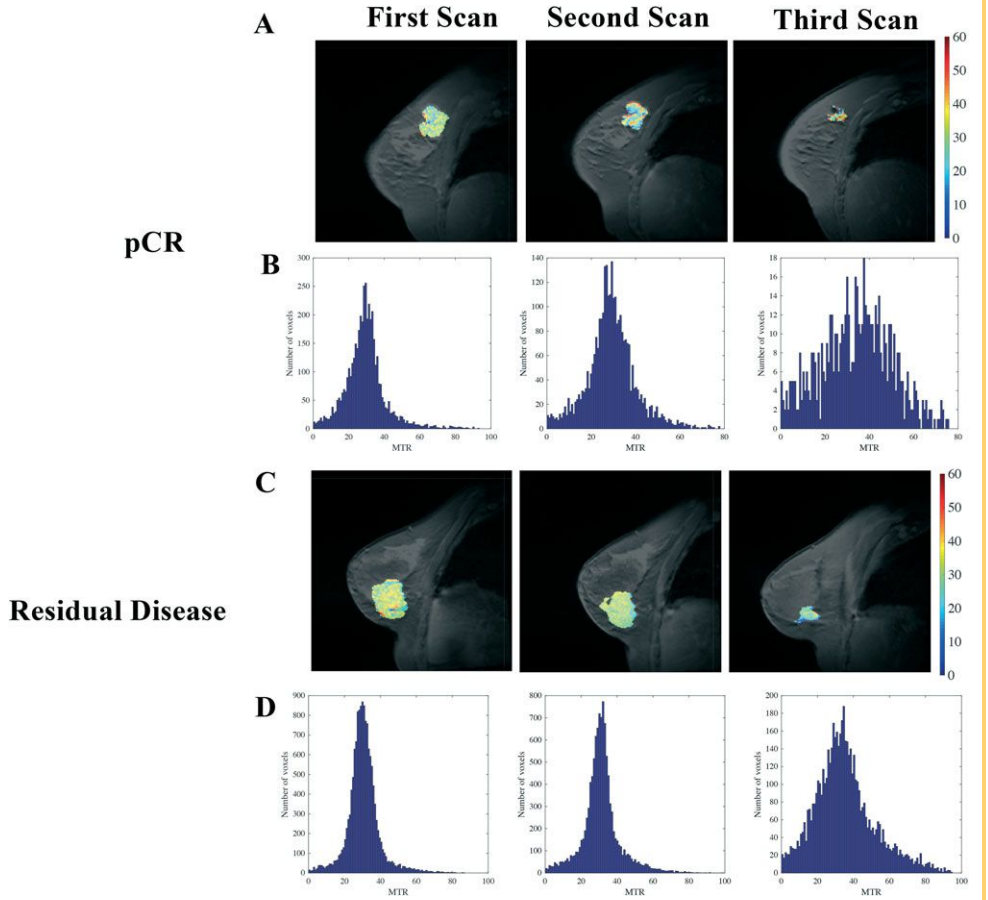


Figure 5. Example MTR maps from (A) a subject who achieved pathological complete response and (C) a subject who had residual disease at the conclusion of therapy at the first, second, and third scan sessions. Histograms of voxel distributions of MTR at the first, second, and third MRI scans for all patients who achieved pathological complete response (B) and all patients who had residual disease (bottom row) show increased heterogeneity in the patients who achieved pCR compared with histograms of those who had residual disease (D).

response to NAT and can be integrated into current standard-of-care therapeutic monitoring paradigms.

The primary finding of this study, that the tumor displays increased heterogeneity of MTR values during the course of NAT while the mean tumor MTR value is unchanged, suggests that MTR reflects a heterogeneous response throughout the tumor. The increased heterogeneity in MTR distribution found in patients achieving pCR compared with that in patients with residual disease at the end of NAT exceeds the heterogeneity found in repeatability studies of healthy breast, suggesting that the results found are not due solely to intraindividual variation. We hypothesize that areas with high MTR after therapy may reflect fibrotic areas, whereas areas with low MTR after therapy may reflect edematous areas remaining after tumor death. When performed at the conclusion of chemoradiation of rectal tumors, MTR has shown increased values in fibrotic tissue and lower values in edematous tissue (23). Breast cancer chemotherapy is also considered to induce fibrosis, owing to remodeling of the extracellular matrix by increasing expression of fibulin (24) and formation of cancer-associated fibroblasts that secrete fibronectin and collagen (25). Higher concentrations of these macromolecules would be expected to result in higher MTR values. Alternately, diffusion-weighted MRI has shown increased water diffusion in breast tumors after chemotherapy, reflecting cell necrosis (14). Correlative studies comparing results from diffusion-weighted MRI and MT-MRI on a voxel-wise basis are underway to further investigate mechanisms of altered MTR after therapy.

The MTR parameter quantified in this study is semiquantitative in nature, in contrast with quantitative magnetization transfer (qMT) techniques that model the magnetization transfer process to separate relaxation and exchange rates and ultimately derive the concentration of macromolecules relative to free water (26). In contrast, MTR depends on both frequency and power of the saturation pulse used during image acquisition, as well as the relaxation and exchange rates of the tissue (4). Thus, although the MTR values in this study can be compared across patients as they were performed on the same scanner with identical acquisition parameters, these MTR values calculated cannot be generalized across sites with different scanner hardware or protocols. MTR was used in this study owing to its fast acquisition time (<2 minutes) and the fact that this study was performed at community imaging centers, which do not typically have the capability to patch scanners with novel pulse programs necessary to perform qMT. Notably, a previous study of qMT in human FGT found repeatability metrics similar to those calculated for MTR in this experiment (27). Future studies using qMT to investigate changes in tumors in response to therapy are currently being performed to generalize these results across sites.

This study is subject to a number of limitations. The composition of breast tissue is known to change with age (28), as well as through the course of the menstrual cycle (29),

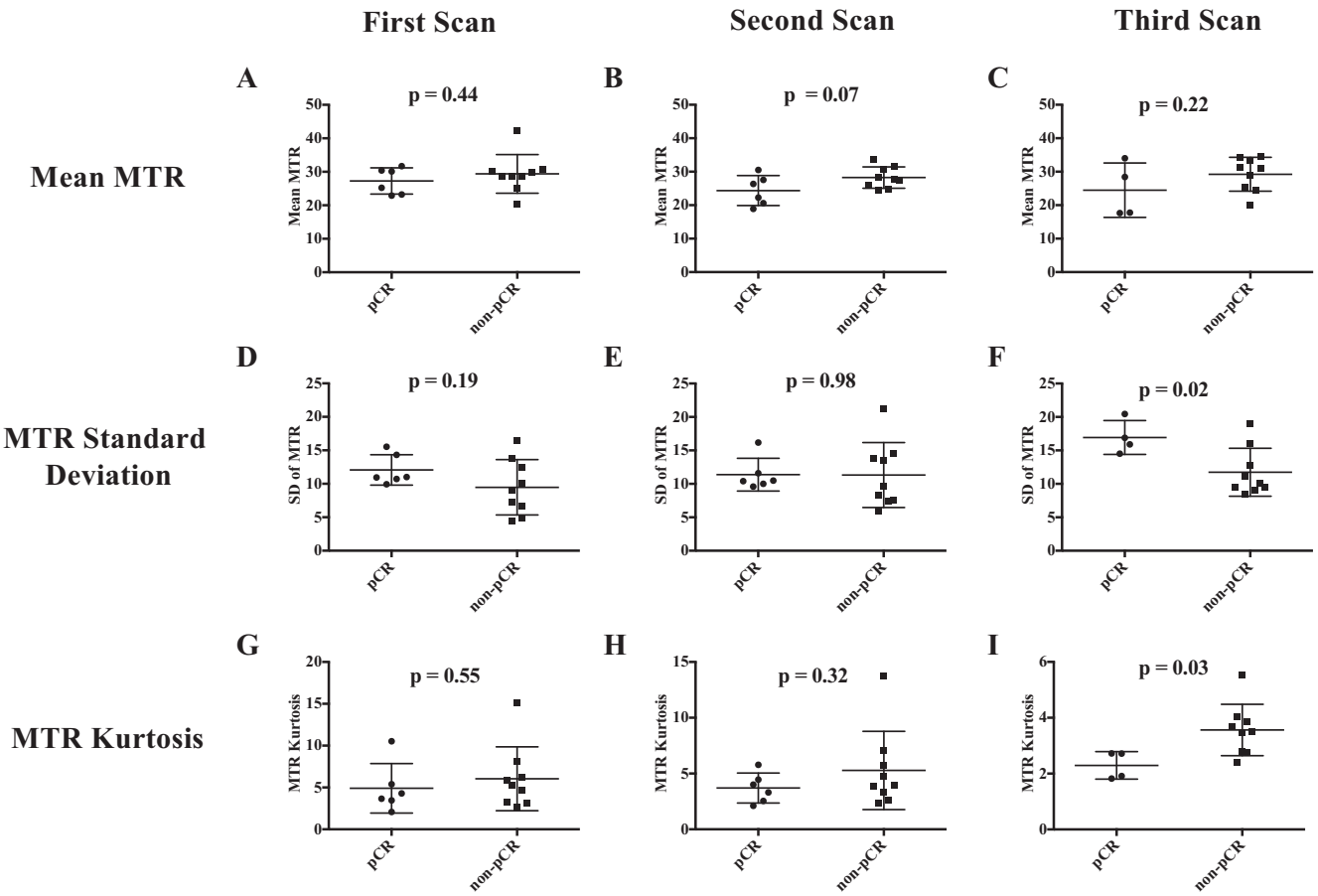


Figure 6. Average tumoral MTR values are similar at the first (A), second (B), and third scan (C) sessions in patients who achieved pathological complete response (pCR) and those who had residual disease at the conclusion of therapy. The standard deviation of the distribution of tumor MTR values at the first (D), second (E), and third (F) scan sessions show higher standard deviation at the third scan session in patients who achieved pCR. The kurtosis of the distribution of tumor MTR values at the first (G), second (H), and third (I) scan sessions show lower kurtosis at the third scan session in patients who achieved pCR.

which could influence the results in this study. We found a nonsignificant trend toward increasing MTR in FGT of younger patients; however, there was no relationship between MTR measurements in tumors and patient age. MTR reproducibility measurements were not made at fixed points in the menstrual cycle, which may affect the reproducibility of our measurements. However, previous studies indicated that the MTR of FGT does not vary across different phases of the menstrual cycle (30). Furthermore, measurements in the breast tumors of patients undergoing NAT are not expected to be affected by menstrual cycle fluctuations, as patients often experience amenorrhea owing to chemotherapy (31).

In summary, this study shows the potential of MT-MRI for assessing changes to breast tumors induced by chemotherapy and that these measurements are repeatable and reproducible across time and scanners. These measurements were made in community radiology clinics, showing the potential for widespread clinical dissemination.

ACKNOWLEDGMENTS

We thank the National Cancer Institute for support through U01CA174706 and U01CA142565. We thank the Cancer Prevention and Research Institute of Texas (CPRIT) for funding through RR160005. We thank Dr. Vibhas Deshpande with Siemens Healthineers for assistance with developing the imaging protocol and helpful discussions.

Disclosure: No disclosures to report.

Conflict of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. Sled JG. Modelling and interpretation of magnetization transfer imaging in the brain. Neuroimage. 2018;182:128–135.
2. Bonini RH, Zeotti D, Saraiva LA, Trad CS, Filho JM, Carrara HH, de Andrade JM, Santos AC, Muglia VF. Magnetization transfer ratio as a predictor of malignancy in breast lesions: preliminary results. Magn Reson Med. 2008;59:1030–1034.
3. Wolff SD, Balaban RS. Magnetization transfer contrast (MTC) and tissue water proton relaxation in vivo. Magn Reson Med. 1989;10:135–144.
4. Henkelman RM, Stanisz GJ, Graham SJ. Magnetization transfer in MRI: a review. NMR Biomed. 2001;14:57–64.
5. Filippi M, Rocca MA. Magnetization transfer magnetic resonance imaging in the assessment of neurological diseases. J Neuroimaging. 2004;14:303–313.
6. Callicott C, Thomas JM, Goode AW. The magnetization transfer characteristics of human breast tissues: an in vitro NMR study. Phys Med Biol. 1999;44:1147–1154.
7. Santyr GE, Kelcz F, Schneider E. Pulsed magnetization transfer contrast for MR imaging with application to breast. J Magn Reson Imaging. 1996;6:203–212.
8. Heller SL, Moy L, Lavianlivi S, Moccaldi M, Kim S. Differentiation of malignant and benign breast lesions using magnetization transfer imaging and dynamic contrast-enhanced MRI. J Magn Reson Imaging. 2013;37:138–145.
9. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011; 144:646–674.
10. Kaushik S, Pickup MW, Weaver VM. From transformation to metastasis: deconstructing the extracellular matrix in breast cancer. Cancer Metastasis Rev. 2016; 35:655–667.
11. Martincich L, Montemurro F, Cirillo S, Marra V, De Rosa G, Ponzone R, Aglietta M, Regge D. Role of magnetic resonance imaging in the prediction of tumor response in patients with locally advanced breast cancer receiving neoadjuvant chemo-therapy. Radiol Med. 2003;106:51–58.
12. Prevos R, Smidt ML, Tjan-Heijnen VC, van Goethem M, Beets-Tan RG, Wildberger JE, Lobbes MB. Pre-treatment differences and early response monitoring of neoadjuvant chemotherapy in breast cancer patients using magnetic resonance imaging: a systematic review. Eur Radiol. 2012;22:2607–2616.
13. Virostko J, Hainline A, Kang H, Arlinghaus LR, Abramson RG, Barnes SL, Blume JD, Avery S, Patt D, Goodgame B, Yankeelov TE, Sorace AG. Dynamic contrast-enhanced magnetic resonance imaging and diffusion-weighted magnetic resonance imaging for predicting the response of locally advanced breast cancer to neoadjuvant therapy: a meta-analysis. J Med Imaging (Bellingham). 2018;5:011011.
14. Pickles MD, Gibbs P, Lowry M, Turnbull LW. Diffusion changes precede size reduction in neoadjuvant treatment of breast cancer. Magn Reson Imaging. 2006; 24:843–847.
15. Delille JP, Slanetz PJ, Yeh ED, Halpern EF, Kopans DB, Garrido L. Invasive ductal breast carcinoma response to neoadjuvant chemotherapy: noninvasive monitoring with functional MR imaging pilot study. Radiology. 2003;228:63–69.
16. Padhani AR, Hayes C, Assersohn L, Powles T, Makris A, Suckling J, Leach MO, Husband JE. Prediction of clinicopathologic response of breast cancer to primary chemotherapy at contrast-enhanced MR imaging: initial clinical results. Radiology. 2006;239:361–374.
17. Li X, Abramson RG, Arlinghaus LR, Kang H, Chakravarthy AB, Abramson VG, Farley J, Mayer IA, Kelley MC, Meszoely IM, Means-Powell J, Grau AM, Sanders M, Yankeelov TE. Multiparametric magnetic resonance imaging for predicting pathological response after the first cycle of neoadjuvant chemotherapy in breast cancer. Invest Radiol. 2015;50:195–204.
18. Wu LA, Chang RF, Huang CS, Lu YS, Chen HH, Chen JY, Chang YC. Evaluation of the treatment response to neoadjuvant chemotherapy in locally advanced breast cancer using combined magnetic resonance vascular maps and apparent diffusion coefficient. J Magn Reson Imaging. 2015;42:1407–1420.
19. Koenig SH, Brown RD, 3rd, Ugolini R. Magnetization transfer in cross-linked bovine serum albumin solutions at 200 MHz: a model for tissue. Magn Reson Med. 1993;29:311–316.
20. Sorace AG, Wu C, Barnes SL, Jarrett AM, Avery S, Patt D, Goodgame B, Luci JJ, Kang H, Abramson RG, Yankeelov TE, Virostko J. Repeatability, reproducibility, and accuracy of quantitative MRI of the breast in the community radiology setting. J Magn Reson Imaging. 2018. [Epub ahead of print].
21. Chen W, Giger ML, Bick U. A fuzzy c-means (FCM)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced MR images. Acad Radiol. 2006;13:63–72.
22. Al-Radaideh A, Mougin OE, Lim SY, Chou IJ, Constantinescu CS, Gowland P. Histogram analysis of quantitative T1 and MT maps from ultrahigh field MRI in clinically isolated syndrome and relapsing-remitting multiple sclerosis. NMR Biomed. 2015;28:1374–1382.
23. Martens MH, Lambregts DM, Papanikolaou N, Alefantinou S, Maas M, Manikis GC, Marias K, Riedl RG, Beets GL, Beets-Tan RG. Magnetization transfer imaging to assess tumour response after chemoradiotherapy in rectal cancer. Eur Radiol. 2016;26:390–397.
24. Pupa SM, Giuffré S, Castiglioni F, Bertola L, Cantú M, Bongarzone I, Baldassari P, Mortarini R, Argraves WS, Anichini A, Menard S, Tagliabue E. Regulation of breast cancer response to chemotherapy by fibulin-1. Cancer Res. 2007;67: 4271–4277.
25. Peiris-Pages M, Sotgia F, Lisanti MP. Chemotherapy induces the cancer-associated fibroblast phenotype, activating paracrine Hedgehog-GLI signalling in breast cancer cells. Oncotarget. 2015;6:10728–10745.
26. Sled JG, Pike GB. Quantitative imaging of magnetization transfer exchange and relaxation properties in vivo using MRI. Magn Reson Med. 2001;46: 923–931.
27. Arlinghaus LR, Dortch RD, Whisenant JG, Kang H, Abramson RG, Yankeelov TE. Quantitative magnetization transfer imaging of the breast at 3.0 T: reproducibility in healthy volunteers. Tomography. 2016;2:260–266.
28. Hart BL, Steinbock RT, Mettler FA, Jr., Pathak DR, Bartow SA. Age and race related changes in mammographic parenchymal patterns. Cancer. 1989;63: 2537–2539.
29. White E, Velentgas P, Mandelson MT, Lehman CD, Elmore JG, Porter P, Yasui Y, Taplin SH. Variation in mammographic breast density by time in menstrual cycle among women aged 40-49 years. J Natl Cancer Inst. 1998;90:906–910.
30. Clendenen TV, Kim S, Moy L, Wan L, Rusinek H, Stanczyk FZ, Pike MC, Zeleniuch-Jacquotte A. Magnetic resonance imaging (MRI) of hormone-induced breast changes in young premenopausal women. Magnetic resonance imaging. 2013; 31:1–9.
31. Rose DP, Davis TE. Ovarian function in patients receiving adjuvant chemotherapy for breast cancer. Lancet. 1977;1:1174–1176.

Assessing Treatment Response of Glioblastoma to an HDAC Inhibitor Using Whole-Brain Spectroscopic MRI

Saumya S. Gurbani[1,2], Younghyoun Yoon[1], Brent D. Weinberg[3], Eric Salgado[1], Robert H. Press[1], J. Scott Cordova[1], Karthik K. Ramesh[1,2], Zhongxing Liang[1], Jose Velazquez Vega[4], Alfredo Voloschin[5], Jeffrey J. Olson[6], Eduard Schreibmann[1], Hyunsuk Shim[1,2,3], and Hui-Kuo G. Shu[1]

[1]Department of Radiation Oncology, Winship Cancer Institute of Emory University, Atlanta, GA; [2]Department of Biomedical Engineering, Emory University and Georgia Institute of Technology, Atlanta, GA; Departments of [3]Radiology and Imaging Sciences, [4]Pathology and Laboratory Medicine; [5]Hematology and Medical Oncology, and [6]Neurosurgery, Emory University School of Medicine, Atlanta, GA

Corresponding Authors:
Hyunsuk Shim, PhD
Department of Radiation Oncology, Winship Cancer Institute of Emory University, 1701 Uppergate Drive, Atlanta, GA 30322;
E-mail: hshim@emory.edu; and Hui-Kuo Shu,
E-mail: hshu@emory.edu

Key Words: glioblastoma, spectroscopic MRI, histone deacetylase inhibitor, belinostat, orthotopic rat glioma model

Abbreviations: Histone deacetylase inhibitors (HDACi), glioblastoma (GBM), progression-free survival (PFS), radiation therapy (RT), temozolomide (TMZ), histone deacetylases (HDACs), suberanilohydroxamic acid (SAHA), spectroscopic magnetic resonance imaging (sMRI), myo-inositol (MI), myo-inositol phosphatase (MIP), dimethyl sulfoxide (DMSO), Dulbecco's modified eagle medium (DMEM), lipopolysaccharide (LPS), echo planar spectroscopic imaging (EPSI), T1-weighted (T1w), Inventory of Depressive Symptomatology Self Report (IDS-SR)

ABSTRACT

Histone deacetylases regulate a wide variety of cellular functions and have been implicated in redifferentiation of various tumors. Histone deacetylase inhibitors (HDACi) are potential pharmacologic agents to improve outcomes for patients with gliomas. We assessed the therapeutic efficacy of belinostat (PXD-101), an HDACi with blood–brain barrier permeability. Belinostat was first tested in an orthotopic rat glioma model to assess in vivo tumoricidal effect. Our results showed that belinostat was effective in reducing tumor volume in the orthotopic rat glioma model in a dose-dependent manner. We also tested the antidepression activity of belinostat in 2 animal models of depression and found it to be effective. Furthermore, we confirmed that myo-inositol levels improved by belinostat treatment in vitro. In a human pilot study, it was observed that belinostat in combination with chemoradiation may delay initial recurrence of disease. Excitingly, belinostat significantly improved depressive symptoms in patients with glioblastoma compared with control subjects. Finally, spectroscopic magnetic resonance imaging of 2 patient cases from this pilot study are presented to indicate how spectroscopic magnetic resonance imaging can be used to monitor metabolite response and assess treatment effect on whole brain. This study highlights the potential of belinostat to be a synergistic therapeutic agent in the treatment of gliomas.

INTRODUCTION

Glioblastomas (GBMs; WHO grade IV glioma) are highly aggressive malignant primary adult brain tumors. Despite comprehensive treatment consisting of neurosurgical resection, high-dose radiation therapy (RT), and chemotherapy (temozolomide, TMZ), the median progression-free survival (PFS) remains 5–7 months (1). Given these poor results, there is an urgent need for improved therapy options. A potential therapeutic target is the family of histone deacetylases (HDACs) that comprises 18 different nuclear and cytoplasmic proteins primarily involved in modulating gene expression through epigenetic mechanisms but also having a broad impact on many additional pathways, including ones associated with cellular metabolism and cell cycle regulation (2-4). Several specific HDACs, particularly in class I and class II, show increased expression and are thought to contribute to oncogenesis in several types of cancer, including ones arising in breast, prostate, lung, and brain (5). As a result, the development of targeted HDAC inhibitors (HDACi) is an active research area for pharmacologic treatment of these diseases (6). In 2006, suberanilohydroxamic acid (SAHA), a first-generation HDACi which targets multiple class I and class II HDAC family members, became the first HDACi to receive FDA approval for advanced cutaneous T cell lymphomas (7). Preclinical investigations of SAHA have also shown antitumor effects in orthotopic glioma animal models (8, 9). This suggests that development of potent HDACis capable of penetrating the blood–brain barrier has the potential to improve therapeutic outcomes of patients with GBM. Research is ongoing into evaluating the synergistic effect of HDACi and chemoradiation for such patients (10).

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2018.00031

Belinostat (PXD101, Spectrum Pharmaceuticals Inc., Irvine, CA), a new pan-HDACi that is structurally similar to SAHA, improves upon the former by having greater blood–brain barrier uptake, which may potentiate its use in the treatment of CNS tumors (11, 12). Belinostat received FDA approval for patients with relapsed/refractory peripheral T cell lymphoma in 2014 (13). In this work, we seek to show a translational analysis of belinostat in the treatment of GBM and describe how a quantitative imaging technique, proton spectroscopic magnetic resonance imaging (sMRI), can serve as a reliable imaging biomarker for monitoring therapy response of belinostat when combined with standard chemoradiation. First, we tested the antitumor effect of belinostat in an orthotopic rat glioma model. Second, we assessed the antidepression effect of belinostat in 2 well-known depression animal models. We also quantified the increase of mRNA levels of bottleneck enzymes for the production of myo-inositol (MI), myo-inositol phosphatase (MIP), an sMRI-detectable metabolite known to be associated with depression. Finally, we assessed the impact of belinostat in combination with chemoradiation in a human pilot study (ClinicalTrials.gov ID: NCT02137759) and present interim results for PFS and a survey of depressive symptomatology. We present sMRI and clinical data from patients in this study to evaluate the modality's use in monitoring response to belinostat + chemoradiation. Our results show a statistically significant improvement of depressive symptoms with belinostat treatment, consistent with our animal data. These results support the utility of belinostat as an adjuvant therapy for GBM and sMRI as a quantitative imaging technique that can noninvasively monitor therapy response.

METHODOLOGY

Cell Culture and In Vitro HDACi Treatment

Belinostat and other HDACis were dissolved in dimethyl sulfoxide (DMSO) to obtain a 100mM stock solution. A 9L rat glioma cell line was maintained in Dulbecco's modified eagle medium (DMEM) (Mediatech Inc., Manassas, MA) supplemented with 10% fetal bovine serum and antibiotics at 37°C in 5% CO_2. 9L cells were plated in 100-mm cell culture petri dishes. Cells were then treated 2 days following seeding with fresh medium containing various HDACis at concentrations of 1 μM for 12 h and were collected to prepare total RNA.

RNA Isolation, RT-PCR, and Real-Time RT-PCR

Cells were collected 12-hour postincubation; these cells underwent RNA isolation and reverse transcription-polymerase chain reaction (RT-PCR) to assess the total mRNA expression levels of the key enzymes in the synthesis of MI (MIP) (8). Total RNA was extracted from cultured cells following the manufacturer's instructions as previously described (14). Primer sequences of MIP-1 were as follows: MIP-1 (GenBank accession number: NM_016368), 5′- AGCTGCATCGAGAACATCCT –3′ and 5′-GGGTACCGGTCCTTTCTTGT –3′; SYBR Green quantitative PCR reaction was carried out in a 15-μL reaction volume containing 2× PCR Master Mix (Applied Biosystems) per our previous reports (14).

Antitumor Effect in an In Vivo Rat Glioma Model

Using a previously described orthotopic rat glioma model (8), the tumoricidal and psychological effects of belinostat were tested. 9L rat glioma cells were stereotactically injected into the frontal lobes of male Fischer 344 rats (n = 9). At postinjection day 9, rats were treated with a daily intraperitoneal injection of either vehicle (10% DMSO, n = 1) or tiered doses of belinostat (n = 2 each of 25 mg/kg, 50 mg/kg, 75 mg/kg, and 100 mg/kg) for 4 days. Throughout the experiment, rats were monitored for mood behavior and activity levels using the volume of droppings as a surrogate measurement. Animals were sacrificed on postinjection day 12, and tumors were excised. This protocol was approved by the Institutional Animal Care and Use Committee (IACUC) at Emory University.

Antidepression Effect Assessment of Belinostat in 2 Animal Models

As described previously (15), the forced-swim test and tail suspension tests were used to assess the antidepression effect of belinostat. Five 6-week-old C57 black female mice were used in each group for forced-swim test and five 7-week-old NIH Swiss male mice were used in each group for tail suspension test. C57 black mice may not perform well in the tail suspension test owing to tail climbing behavior (https://www.research.psu.edu/arp/experimental-guidelines/rodent-behavioral-tests-1/rodent-behavioral-tests.html), whereas the NIH Swiss mice did not have similar issues. The forced-swim test was performed 6 h after belinostat treatment (75 mg/kg i.p.) in a 4-L beaker containing 3 L of tap water at a temperature of 25°C. Video tracking–based methods were used to record the duration of time spent "immobile" in the arena over 6 minutes (immobility measured between 2 and 8 minutes of a 10-minute trial; the extent of immobility correlates with levels of depression). Similarly, the tail suspension test is based on the fact that animals subjected to inescapable stress of being suspended by their tail for the short term, would develop an immobile posture (16). For the tail suspension test, the lipopolysaccharide (LPS)-induced depression model was used (17). Twenty-four hours after LPS administration (Sigma L3129; 0.85 mg/kg i.p.), the tail suspension test was performed. Video tracking–based methods were used to record the duration of time spent in immobility for 6 minutes.

Clinical Study

Patients with newly diagnosed GBM were enrolled in either the control or treatment arm of an Institutional Review Board (IRB)-approved clinical trial at Emory University (ClinicalTrials.gov ID NCT02137759), wherein the treatment arm received intravenous belinostat (Spectrum Pharmaceuticals, Irvine, CA) as an investigational therapeutic. The study was not randomized, with patients serially enrolling into the control arm (in 2014–2015) followed by the belinostat treatment arm (in 2015–2018). All patients underwent maximal safe tumor resection, if resection was feasible, before enrolling in the study. Patients in both arms of the trial received standard-of-care therapy consisting of daily TMZ (75 mg/m^2) × 42 days and focal radiation doses of 60 and 51 Gy to the resection cavity/residual contrast-enhancing tissue (per T1-weighted contrast-enhanced MRI postresection) and T2/FLAIR signal, respectively, in 30 fractions. Margins of 0.5–1.0 cm and 0.3–0.5 cm were added to the target volumes to generate the clinical treatment volume and planning treatment volume to account for microscopic disease spread and spatial uncertainty

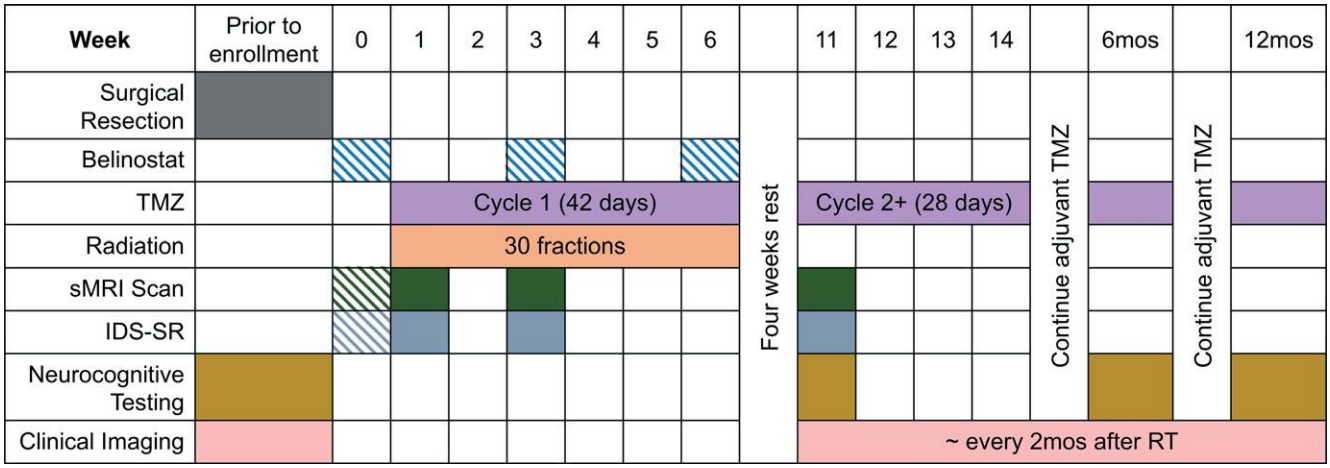


Figure 1. One-year timeline of chemotherapy, intravenous belinostat, radiation, spectroscopic magnetic resonance imaging (sMRI) scanning, Inventory of Depressive Symptomatology Self Report (IDS-SR) survey, and neurocognitive testing for patients in NCT02137759. Hashed boxes indicate items conducted for patients in only the treatment arm of the study.

in treatment delivery (18). In the treatment arm, patients received daily intravenous doses of belinostat at either 500 or 750 mg/m^2 for 5 consecutive days in 3 cycles, spaced 3 weeks apart beginning 1 week before the start of chemoradiation, as shown in Figure 1. The first 3 patients received a higher dose of belinostat. However, because 2 of the patients had serious adverse events with hematologic toxicity during the course of belinostat, TMZ, and radiation, the dose was lowered to 500 mg/m^2 for the remaining patients in the trial.

Each patient in the study underwent an sMRI scan prior to starting chemoradiation (week 1), after 2 weeks of chemoradiation (week 3), and 4 weeks after completing radiation (week 11). Patients in the treatment arm underwent an additional sMRI scan before starting the first week of belinostat (week 0). sMRI scans were conducted on a 3 T MR scanner (Siemens TimTrio or Siemens PRISMA with 32 channel head coil, Siemens Healthineers, Erlangen, Germany) using an echo planar spectroscopic imaging (EPSI) sequence combined with generalized autocalibrating partially parallel acquisition (GRAPPA), and metabolite maps were produced using the MIDAS software (University of Miami, Miami, FL) (19, 20). The metabolite maps were coregistered to a volumetric T1-weighted (T1w) MRI taken during the same scanning session with the patient in the same orientation. Longitudinal scans on the 4 patients were coregistered and brought into the first scan (week 0/1) imaging space using rigid registration. After each sMRI scan, the patient completed the Inventory of Depressive Symptomatology Self Report (IDS-SR), a validated 30-question survey designed to assess depressive symptoms (21, 22).

EPSI/GRAPPA sequence parameters were optimized to enhance the signal of choline (Cho, a metabolite involved in the synthesis of the phospholipid cell membrane and increased in tumors) and NAA (a healthy neuronal marker decreased as neoplasia invades into and destroys neuronal tissue). Patients were followed-up with standard-of-care imaging (contrast-enhanced T1-weighted MRI, CE-T1w MRI; fluid attenuation inversion recovery, FLAIR) for 12 months post-treatment or until progression of disease was confirmed by neuroradiologist. A total of 26 patients (13 control, 13 treatment) were enrolled at Emory University; of these, 3 did not complete the treatment protocol (1 in control arm, 2 in belinostat arm), and 2 did not undergo surgical resection of tumor (only underwent a biopsy for diagnosis). These 5 patients are excluded from analysis (see online Supplemental Figure 1). PFS is reported for patients based on time to radiologic confirmation of disease progression (per CE-T1w MRI) from the date of surgery. Data are right-censored for patients who have no known disease progression or were lost to follow-up. Sample cases from each arm of the study are shown to show the ability of sMRI to identify early response to treatment. Because follow-up data are continuing to be collected for patients in the treatment arm of the study, statistical analyses of the full data will be presented in future work.

RESULTS

Antitumor Effect of Belinostat in an Orthotopic Rat Model

In Figure 2, photographs of the 9 rats evaluated in this experiment are shown in their cage at pretreatment with belinostat and at day 13, 4 days after treatment, when the rats were sacrificed. The volume of animal droppings seen in the cage is used as a surrogate measure of activity and normal physiology. Rats showed decreased movement and grooming, measures of mood, before treatment with belinostat. The restoration of activity and improved mood was observed in a dose-dependent manner, with normal levels observed in rats treated with the highest 2 doses (75 mg/kg and 100 mg/kg). Photographs of the tumor in situ and excised are also shown in Figure 2, showing a similar dose-dependent decrease in tumor volume from untreated mice through the increasing doses of belinostat.

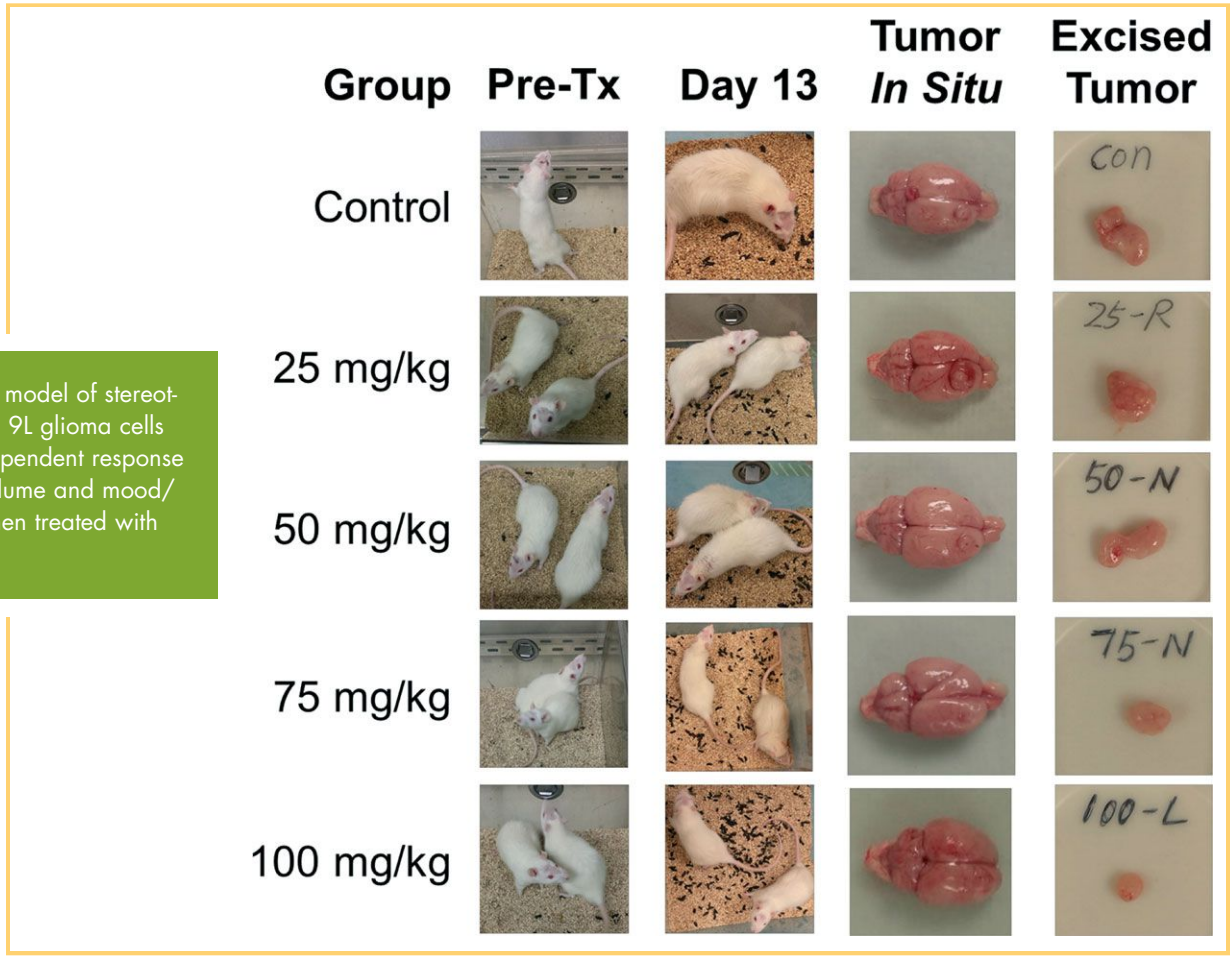


Figure 2. A rat model of stereotactically injected 9L glioma cells shows a dose-dependent response in both tumor volume and mood/activity levels when treated with belinostat.

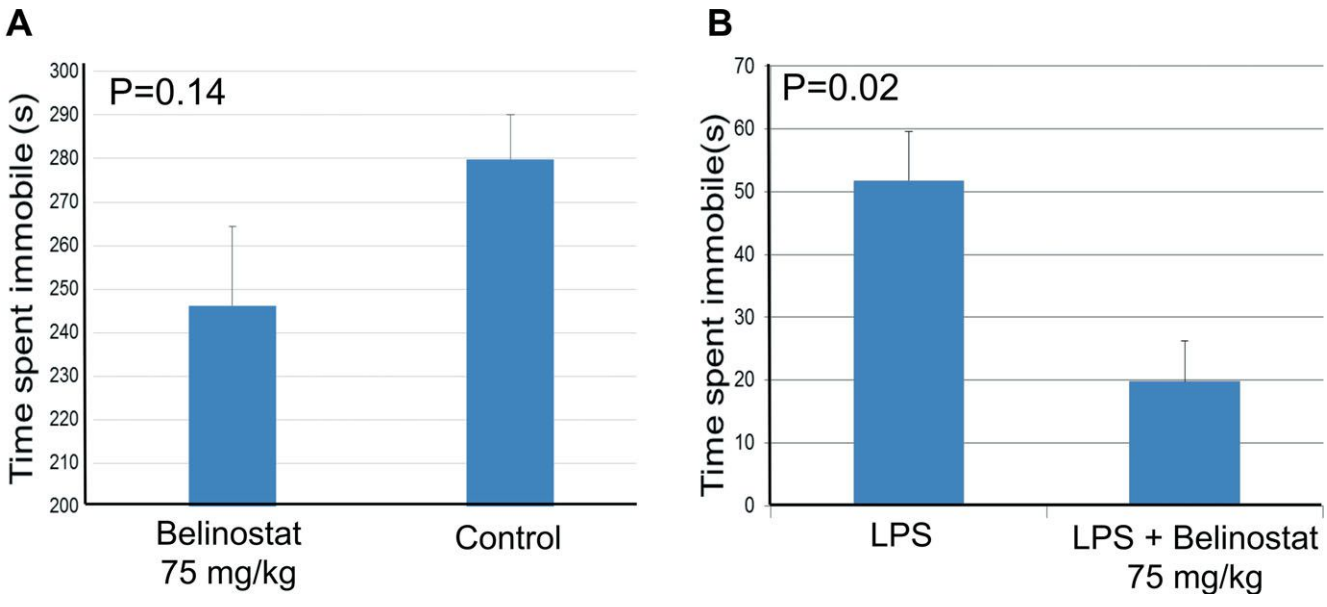


Figure 3. Two mouse models of depression to assess antidepressive effect of belinostat: Force-swim test measuring the time spent in immobility during 2–8 minutes (6 minutes) (A). Tail suspension test measuring the time spent in immobility in mice treated with lipopolysaccharide (LPS) for 6 minutes. Five mice were used in each group (B).

A

	Control	Belinostat
Number of patients	11	10
Age	57.5 ± 12.3	50.3 ± 13.6
IDH1 mutation	1 (9%)	1 (10%)
MGMT methylated	3 (27%)	3 (30%)
6 Month PFS	8 (73%)	10 (100%)

B

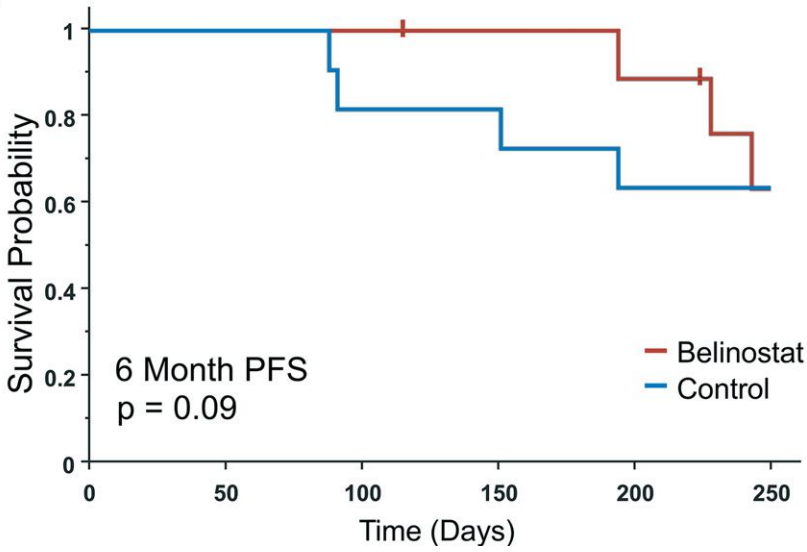


Figure 4. Progression-free survival (PFS) of patients in both arms of the study up to 1 year (A). Age indicates age at time of surgery, and PFS is right-censored from the time of surgery. Kaplan–Meier curves for the 2 arms of the study; tick marks indicate time of censoring (B).

Antidepression Effect of Belinostat in 2 Animal Models

Figure 3 shows the results of the forced-swim and tail suspension tests. In the forced-swim test, the mice who received belinostat spent less time immobile compared to the control mice ($P = .14$). In the tail suspension test, the mice who received LPS + belinostat had a statistically significant decrease in immobility compared with mice who received LPS alone ($P = .02$).

In Vitro Study of mRNA Expression

mRNA expression levels of MIP (a bottleneck enzyme in the production of MI) from HDACi-treated cells are shown in online Supplemental Figure 2 as fold-increases in log-scale compared with those of the untreated cells (DMSO vehicle control). Belinostat showed greater increases in restoration of mRNA levels at the same concentration as other HDACi, including SAHA. The only other HDACi which achieved greater efficacy is quisinostat (JNJ26481585, Janssen Pharmaceuticals, Beerse, Belgium), a second-generation pan-HDACi, which was being tested in phase II clinical trials for multiple myeloma (23). However, currently there are no active trials for quisinostat on ClinicalTrials.gov.

Clinical Study

In total, 21 patients who met inclusion criteria for analysis were assessed to determine differences in PFS between the 2 arms (see online Supplemental Figure 1). A table summarizing basic demographics of the 2 arms of the clinical study is shown in Figure 4A. Both arms showed similar distributions of known genetic targets that improve response to radiation—mutation of isocitrate dehydrogenase 1 (IDH1) and promoter methylation of the gene for O (6)-methylguanine-DNA methyltransferase (24). Figure 4B shows Kaplan–Meier curves for PFS from date of surgery (tick marks indicate time of censoring). Six-month PFS was 73% for the control arm and 100% for the belinostat arm. A log-rank test assessing PFS data up to 6 months trended toward statistical significance ($P = .09$). No statistically significant difference was observed on a log-rank test assessing PFS data up to 12 months ($P = .45$). Of these 21 patients, 17 completed an IDS-SR survey at both baseline (week 0 for belinostat arm, week

Table 1. IDS-SR Assessment

	Control	Belinostat	*P value*
Number of Patients	10	7	
Baseline Score	18.2 ± 9.1	22.0 ± 9.8	0.43
Week 11 Score	22.3 ± 10.9	16.1 ± 15.5	0.39
Change in Score	4.1 ± 9.7	−5.9 ± 8.7	0.04

The IDS-SR assessment of patients in both study arms between baseline and 1-month post-RT shows a statistically significant improvement in assessment scores for patients who received belinostat. *P* values indicate results of a two-tailed unpaired *t*-test.

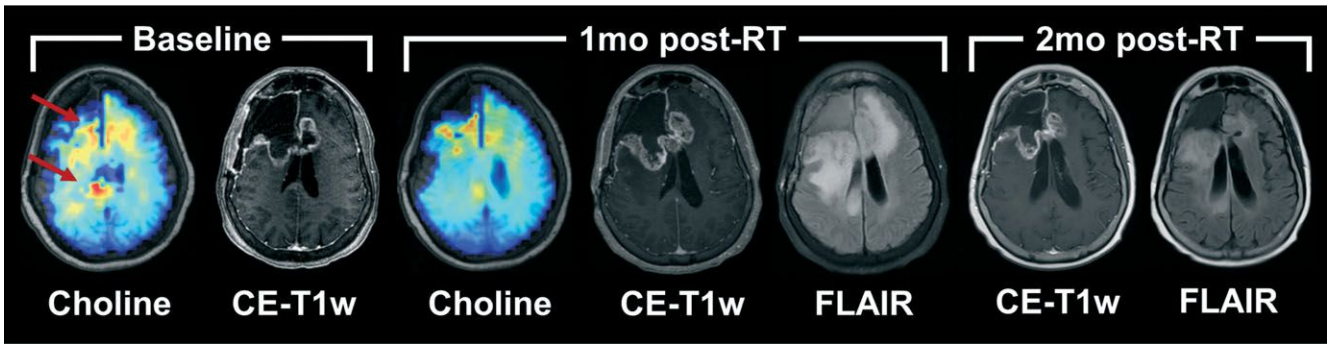


Figure 5. Longitudinal imaging of a patient in the control arm of the study. An sMRI map of choline indicates a response to chemoradiation between baseline and the first follow-up; however, standard clinical imaging indicates potential progression of disease. Further follow-up indicates that the imaging findings at 1 month were attributable to pseudoprogression.

1 for control arm) and at week 11 (see online Supplemental Figure 2; Table 1). While no significant difference in the scores was observed at either time point, the belinostat cohort had a statistically significant improvement in scores over the course of treatment using a 2-tailed unpaired T test ($P = .04$).

Figures 5 and 6 depict sMRI and clinical CE-T1w MRI scans for 2 representative patients, 1 from each of the study arms. At baseline, both patients showed elevated choline metabolism (red arrows) around the resection cavities, indicating the presence of increased cellular turnover associated with neoplasia. One-month post-RT (week 11), both patients showed decreased levels of choline compared with baseline, and low NAA levels owing to subsequent radiation damage to in-field neurons. Therefore, Cho/NAA did not reliably indicate early response (see online Supplemental Figure 3); however, we found that peritumoral MI was improved in the subject who received belinostat (see online Supplemental Figure 3) at 1-month post-RT. The control patient (Figure 5) was deemed to have potential progression of her disease because of the thickened contrast enhancement around the resection cavity on CE-T1w imaging; a month later, however, the thickened contrast rim was gone, and the patient was deemed to not yet have disease progression. The patient in the belinostat arm (Figure 6) had a similar course; only the increase in enhancement occurred 3 months after radiation was completed.

DISCUSSION

In this work, we sought to characterize the antitumor and antidepressant activity of belinostat, a new HDACi with improved brain penetration, in a translational manner: starting from in

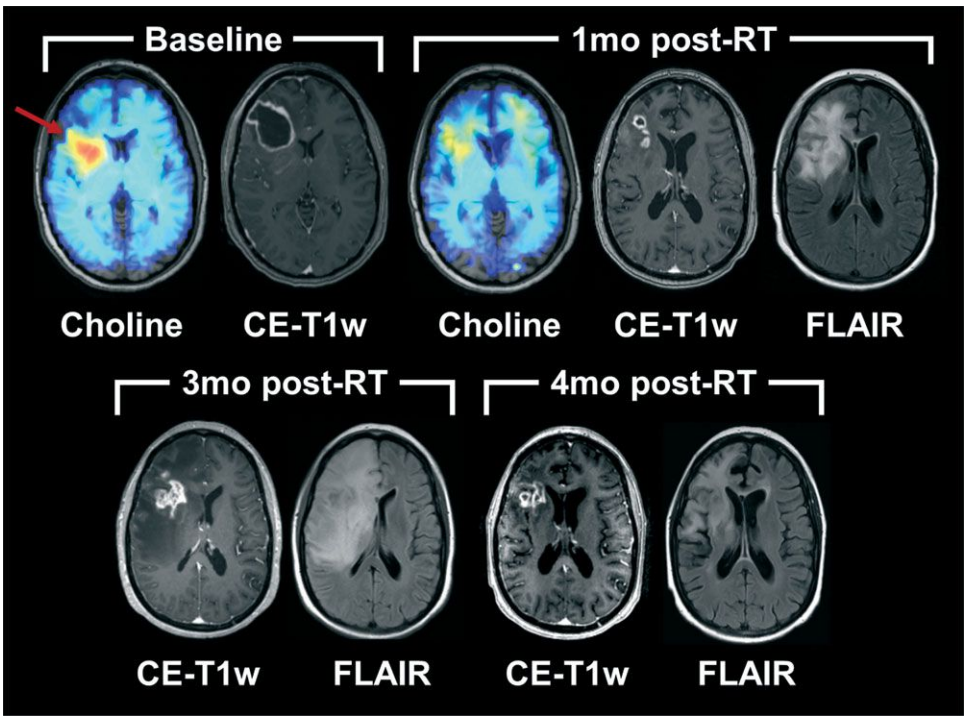


Figure 6. Longitudinal imaging of a patient in the belinostat (treatment) arm of the study. An sMRI map of choline indicates a response to chemoradiation as assessed at 1 month post-RT. Additional follow-up imaging indicates the pseudoprogression phenomenon occurring at 3 month post-RT, resolving by 4 month post-RT.

vivo animal glioma models to testing in patients with glioblastoma. First, we assessed the efficacy of belinostat in reducing tumor volume in an orthotopic rat glioma model. Here, a dose-dependent reduction in tumor size was observed, suggesting the antitumor properties of the drug are effective in crossing the blood–brain barrier in vivo (Figure 2). In addition, it was observed that the activity levels of rats, as measured by grooming activities and droppings, were higher in those treated with increased doses of belinostat. These improvements are consistent with previously reported literature by Covington et al. (25) that HDACis possess antidepressant properties. To further assess the antidepressive effect of belinostat, we subjected mice to 2 different models of induced depression. We found that belinostat reduced the immobility time in both the forced-swim and tail suspension tests (Figure 3), exhibiting the drug's antidepressant effect. We followed these tests with an in vitro assessment of belinostat's effect on MIP, the key enzyme in the production of MI that was implicated in depression. The in vitro cell study showed that belinostat had greater restorative activity for MIP than most other HDACi. An HDACi tested that had higher restoration than belinostat was quisinostat (JNJ26481585); however, there is no clinical trial currently enrolling patients testing quisinostat. As such, belinostat shows promise as a targeted HDACi for glioblastoma because of its increased uptake into the brain and its efficacy in restoring key metabolic activity for depression.

The belinostat clinical trial completed enrollment of patients in August 2018, and patients are continuing to be followed to assess long-term outcomes including progression-free and overall survival. Although statistical claims cannot yet be made regarding long-term survival and efficacy, a comparison of 6-month PFS and initial changes in mood are presented in this work. The cohort receiving belinostat showed a trend of improved 6-month PFS compared to the control cohort ($P = .09$); however, this difference was mitigated by 12 months ($P = .45$). Despite a limited sample size for this study, these results suggest that belinostat may improve response to chemoradiation therapy as hypothesized. A speculated reason for the improved PFS at 6 months but not at 12 months is that belinostat was given to subjects for only a short term during RT.

While 6-month PFS outcomes approached statistical significance, the belinostat cohort did achieve a statistically significant improvement in depression as measured by the IDS-SR ($P = 0.04$). This suggests that the mood improvement effect of belinostat, as shown in our animal data and with the claim by Covington et al. (25), may translate to humans. These preliminary data suggest future large cohorts to be evaluated.

Finally, this study showed the potential of sMRI as a noninvasive monitoring tool for investigational therapeutics. As shown in Figures 5 and 6, both patients appeared to have a reduced tumor burden when assessing choline metabolism at week 1, 1 month after the completion of RT. Owing to radiation-induced damage, NAA was reduced around the high dose area, which made Cho/NAA ineffective in assessing tumor response (see online Supplemental Figure 3). MI showed a slight improvement back towards normal level at the 1-month post-RT scan in the patient treated with belinostat, consistent with IDS-SR score improvement (see online Supplemental Figure 3). Further studies, including longitudinal scanning, are needed to fully elucidate the timeline of metabolite changes in these patients. Standard imaging, however, differed between the 2 patients and suggested that the control patient may have been experiencing disease progression, when eventually it turned out to be stable disease at that time. This is a phenomenon known as pseudoprogression, the ambiguity of CE-T1w findings in differentiating true progression of disease from normal tissue response to high-dose radiation. sMRI, however, is robust to the pseudoprogression phenomenon, as the modality is measuring endogenous intracellular metabolism rather than vasculature damages/changes or tissue phenomena such as edema. Both patients showed similar metabolic signatures, which turned out to be more accurate of the underlying pathology compared to clinical imaging.

CONCLUSION

In this work we described the therapeutic and antitumor, antidepression effects of belinostat, a potent pan-HDACi with blood–brain barrier permeability through: a rat glioma model, 2 mouse depression models, an in vitro cell study, and testing in a pilot clinical study in patients with glioblastoma. The results from this work suggest that belinostat may be an effective HDACi at delaying disease progression and improving depression. Furthermore, it shows that the treatment response can be monitored noninvasively using spectroscopic MRI during pseudoprogression period. Further studies and analysis of the ongoing clinical trial may yield a better understanding of the role that HDACis play in the metabolic profiles of GBM and motivate the development of better, targeted therapies for patients with this debilitating disease.

Additional testing of this drug in human subjects can help with separating this improved mood/activity effect due to a primary property of HDACis from improvement as a secondary effect to reduced tumor burden.

Supplemental Materials

Supplemental Figure 1-3: http://dx.doi.org/10.18383/j.tom.2018.00031.sup.01

ACKNOWLEDGMENTS

We would like to thank Spectrum Pharmaceuticals for providing the investigational drug used in this clinical study; the staff at the Clinical Trials Office at the Winship Cancer Institute for their support and assistance with recruiting and managing the patients enrolled in the clinical study; and Dr. Andrew Maudsley and Mr. Sulaiman Sheriff at the University of Miami for their support with the EPSI sequence and MIDAS software framework. Funding for this work came from National Institutes of Health grant U01CA172027.

Disclosures: No disclosures to report.

Conflict of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. Stupp R, Taillibert S, Kanner AA, Kesari S, Steinberg DM, Toms SA, Taylor LP, Lieberman F, Silvani A, Fink KL, Barnett GH, Zhu JJ, Henson JW, Engelhard HH, Chen TC, Tran DD, Sroubek J, Tran ND, Hottinger AF, Landolfi J, Desai R, Caroli M, Kew Y, Honnorat J, Idbaih A, Kirson ED, Weinberg U, Palti Y, Hegi ME, Ram Z. Maintenance therapy with tumor-treating fields plus temozolomide vs temozolomide alone for glioblastoma: a randomized clinical trial. JAMA. 2015;314:2535–2543.
2. Minucci S, Pelicci PG. Histone deacetylase inhibitors and the promise of epigenetic (and more) treatments for cancer. Nat Rev Cancer. 2006;6:38–51.
3. Cornago M, Garcia-Alberich C, Blasco-Angulo N, Vall-Llaura N, Nager M, Herreros J, Comella JX, Sanchis D, Llovera M. Histone deacetylase inhibitors promote glioma cell death by G2 checkpoint abrogation leading to mitotic catastrophe. Cell Death Dis. 2014;5:e1435.
4. Mottamal M, Zheng S, Huang TL, Wang G. Histone deacetylase inhibitors in clinical studies as templates for new anticancer agents. Molecules. 2015;20:3898–3941.
5. Bieliauskas AV, Pflum MKH. Isoform-selective histone deacetylase inhibitors. Chem Soc Rev. 2008;37:1402–1413.
6. Falkenberg KJ, Johnstone RW. Histone deacetylases and their inhibitors in cancer, neurological diseases and immune disorders. Nat Rev Drug Discov. 2014;13:673–691.
7. Mann BS, Johnson JR, Cohen MH, Justice R, Pazdur R. FDA approval summary: vorinostat for treatment of advanced primary cutaneous T-cell lymphoma. Oncologist. 2007;12:1247–1252.
8. Wei L, Hong S, Yoon Y, Hwang SN, Park JC, Zhang Z, Olson JJ, Hu XP, Shim H. Early prediction of response to Vorinostat in an orthotopic rat glioma model. NMR Biomed. 2012;25:1104–1111.
9. Eyüpoglu IY, Hahnen E, Buslei R, Siebzehnrübl FA, Savaskan NE, Lüders M, Tränkle C, Wick W, Weller M, Fahlbusch R, Blümcke I. Suberoylanilide hydroxamic acid (SAHA) has potent anti-glioma properties in vitro, ex vivo and in vivo. J Neurochem. 2005;93:992–999.
10. Krauze AV, Myrehaug SD, Chang MG, Holdford DJ, Smith S, Shih J, Tofilon PJ, Fine HA, Camphausen K. A phase 2 study of concurrent radiation therapy, temozolomide, and the histone deacetylase inhibitor valproic acid for patients with glioblastoma. Int J Radiat Oncol Biol Phys. 2015;92:986–992.
11. Wang C, Eessalu TE, Barth VN, Mitch CH, Wagner FF, Hong Y, Neelamegam R, Schroeder FA, Holson EB, Haggarty SJ, Hooker JM. Design, synthesis, and evaluation of hydroxamic acid-based molecular probes for in vivo imaging of histone deacetylase (HDAC) in brain. Am J Nucl Med Mol Imaging. 2013;4:29–38.
12. Hanson JE, LA H, Plise E, Chen Y-H, Ding X, Hanania T, Sabath EV, Alexandrov V, Brunner D, Leahy E, Steiner P, Liu L, Scearce-Levie K, Zhou Q. SAHA enhances synaptic function and plasticity in vitro but has limited brain availability in vivo and does not impact cognition. PLoS One. 2013;8:e69964.
13. Lee H-Z, Kwitkowski VE, Del Valle PL, Ricci MS, Saber H, Habtemariam BA, Bullock J, Bloomquist E, Li Shen Y, Chen XH, Brown J, Mehrotra N, Dorff S, Charlab R, Kane RC, Kaminskas E, Justice R, Farrell AT, Pazdur R. FDA Approval: Belinostat for the Treatment of Patients with Relapsed or Refractory Peripheral T-cell Lymphoma. Clin Cancer Res. 2015;21:2666–2670.
14. Liang Z, Wu H, Xia J, Li Y, Zhang Y, Huang K, Wagar N, Yoon Y, Cho HT, Scala S, Shim H. Involvement of miR-326 in chemotherapy resistance of breast cancer through modulating expression of multidrug resistance-associated protein 1. Biochem Pharmacol. 2010;79:817–824.
15. Krishnan R, Cella D, Leonardi C, Papp K, Gottlieb AB, Dunn M, Chiou CF, Patel V, Jahreis A. Effects of etanercept therapy on fatigue and symptoms of depression in subjects treated for moderate to severe plaque psoriasis for up to 96 weeks. Br J Dermatol. 2007;157:1275–1277.
16. Cryan JF, Mombereau C, Vassout A. The tail suspension test as a model for assessing antidepressant activity: review of pharmacological and genetic studies in mice. Neurosci Biobehav Rev. 2005;29:571–625.
17. O'Connor JC, Lawson MA, Andre C, Moreau M, Lestage J, Castanon N, Kelley KW, Dantzer R. Lipopolysaccharide-induced depressive-like behavior is mediated by indoleamine 2, 3-dioxygenase activation in mice. Mol Psychiatry. 2009;14:511–522.
18. Burnet NG, Thomas SJ, Burton KE, Jefferies SJ. Defining the tumour and target volumes for radiotherapy. Cancer Imaging. 2004;4:153–161.
19. Maudsley A, Domenig C. Signal normalization for MR spectroscopic imaging using an interleaved water-reference. Int Soc Magn Reson Med. 2009;61:548–559.
20. Maudsley AA, Domenig C, Govind V, Darkazanli A, Studholme C, Arheart K, Bloomer C. Mapping of brain metabolite distributions by volumetric proton MR spectroscopic imaging (MRSI). Magn Reson Med. 2009;61:548–559.
21. Rush AJ, Carmody T, Reimitz PE. The Inventory of Depressive Symptomatology (IDS): clinician (IDS-C) and self-report (IDS-SR) ratings of depressive symptoms. Int J Methods Psychiatr Res. 2000;9:45–59.
22. Rush AJ, Gullion CM, Basco MR, Jarrett RB, Trivedi MH. The inventory of depressive symptomatology (IDS): psychometric properties. Psychol Med. 1996;26:477–486.
23. Moreau P, Facon T, Touzeau C, Benboubker L, Delain M, Badamo-Dotzis J, Phelps C, Doty C, Smit H, Fourneau N, Forslund A, Hellemans P, Leleu X. Quisinostat, bortezomib, and dexamethasone combination therapy for relapsed multiple myeloma. Leuk Lymphoma. 2016;57:1546–1559.
24. Stupp R, Hegi ME, Mason WP, van den Bent MJ, Taphoorn MJB, Janzer RC, Ludwin SK, Allgeier A, Fisher B, Belanger K, Hau P, Brandes AA, Gijtenbeek J, Marosi C, Vecht CJ, Mokhtari K, Wesseling P, Villa S, Eisenhauer E, Gorlia T, Weller M, Lacombe D, Cairncross JG, Mirimanoff RO; European Organisation for Research and Treatment of Cancer Brain Tumour and Radiation Oncology Groups; National Cancer Institute of Canada Clinical Trials Group. Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. Lancet Oncol. 2009;10:459–466.
25. Covington HE, 3rd, Maze I, LaPlant QC, Vialou VF, Ohnishi YN, Berton O, Fass DM, Renthal W, Rush AJ 3rd, Wu EY, Ghose S, Krishnan V, Russo SJ, Tamminga C, Haggarty SJ, Nestler EJ. Antidepressant actions of histone deacetylase inhibitors. J Neurosci. 2009;29:11451–1160.

Real-Time Quantitative Assessment of Accuracy and Precision of Blood Volume Derived from DCE-MRI in Individual Patients During a Clinical Trial

Madhava P. Aryal[1], Choonik Lee[1], Peter G. Hawkins[1], Christina Chapman[1], Avraham Eisbruch[1], Michelle Mierzwa[1], and Yue Cao[1,2,3]

Departments of [1]Radiation Oncology; [2]Radiology; and [3]Biomedical Engineering, University of Michigan, Ann Arbor, MI

Corresponding Author:
Madhava P. Aryal, PhD
Department of Radiation Oncology, University of Michigan,
519 W William St, Ann Arbor, MI 48103;
E-mail: mparyal@med.umich.edu

Key Words: quantitative imaging, repeatability, real-time assessment, dynamic contrast enhanced MRI, blood volume, head and neck cancer

Abbreviations: Quantitative imaging (QI), blood volume (BV), dynamic contrast-enhanced (DCE), magnetic resonance imaging (MRI), magnetic resonance (MR), arterial input function (AIF), quality assurance (QA), head and neck (HN), radiation therapy (RT), echo time (TE), repetition time (TR), 3-dimensional (3D), field of view (FOV), volumes of interest (VOIs), repeatability coefficient (RC), within-subject mean squares (WMS), confidence interval (CI), sternocleidomastoid muscle (SCM)

ABSTRACT

Accuracy and precision of quantitative imaging (QI) metrics should be assessed in real time in each patient during a clinical trial to support QI-based decision-making. We developed a framework for real-time quantitative assessment of QI metrics and evaluated accuracy and precision of dynamic contrast-enhanced (DCE)-magnetic resonance imaging (MRI)–derived blood volume (BV) in a clinical trial for head and neck cancers. Patients underwent DCE-MRI before and after 2 weeks of radiation therapy (2wkRT). A mean as a reference value and a repeatability coefficient (RC) of BV values established from *n* patients in cerebellum volumes of interest (VOIs), which were normal and affected little by therapy, served as accuracy and precision measurements. The BV maps of a new patient were called accurate and precise if the values in cerebellum VOIs and the difference between the 2 scans agreed with the respective mean and RC with 95% confidence. The new data could be used to update reference values. Otherwise, the data were flagged for further evaluation before use in the trial. BV maps from 62 patients enrolled on the trial were evaluated. Mean BV values were 2.21 (±0.14) mL/100 g pre-RT and 2.22 (±0.17) mL/100 g at 2wkRT; relative RC was 15.9%. The BV maps from 3 patients were identified to be inaccurate and imprecise before use in the clinical trial. Our framework of real-time quantitative assessment of QI metrics during a clinical trial can be translated to different QI metrics and organ-sites for supporting QI-based decision-making that warrants success of a clinical trial.

INTRODUCTION

Quantitative imaging (QI) metrics are emerging as a tool for therapeutic response assessment in cancer treatment (1). As QI tools have been technically validated, clinical trials start to make decisions based upon these imaging metrics, for example, quantitative parameters derived from dynamic contrast-enhanced (DCE)-magnetic resonance imaging (MRI) (1, 2).

The DCE-MRI-derived QI metrics can be affected by differences in MRI platforms, pulse sequences, acquisition parameters, image reconstruction schemes, pharmacokinetic models, and quantification software packages (3-8), which limits deployment of DCE-MRI in clinical trials and practice. MRI scanners from each vendor have unique hardware configuration, vendor-specific pulse sequences, and reconstruction schemes, which can cause a systematic bias in estimated QI metrics (4). In addition, selection of magnetic resonance (MR) acquisition parameters can influence quantification of these metrics (5, 9). Furthermore, QI metrics derived from different image-processing software packages can lead to substantial variations in the metrics, even when using the same pharmacokinetic model, T1 map, arterial input function (AIF), and region of interest (6, 7). To address these challenges, collaborative efforts under the initiatives of professional societies and government agencies have been made for development of DCE-MRI profiles, T1 phantoms, digital reference object, and statistical methods to harmonize imaging acquisition across different platforms, to validate imaging hardware and software, to test computer algorithms, and to assess technical performance (4-6, 10-16). All these efforts are absolutely necessary but not sufficient to warrant the accuracy and precision of QI metrics obtained in each individual

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2018.00029

patient during a clinical trial, which could affect decision-making and even clinical outcomes (1). Therefore, it is necessary to develop and implement a quantitative quality assurance (QA) procedure to measure QI metrics acquired in the patients who are on the trial (17).

Accuracy, in general, refers to closeness of a measured QI metrics to a true or known value, while precision is an agreement between repeated measurements of a metrics (17). For any QI metrics that does not have its true value available, its deviation from a reference value, obtained as a group mean from a large sample study in any standard reference region, can serve as its measurement accuracy (17). Precision, more commonly known as repeatability, can be easily evaluated from repeated measurements, often called as test–retest studies, in a normal reference region that is not expected to have any changes during a time interval of test–retest studies (17). Under these principles, a reference value and repeatability coefficient (RC) of a QI metrics in a reference region under certain conditions or constraints of image acquisition and process can be determined from a sample of population with 95% confidence and used to assess accuracy and precision of the metrics measured from an individual patient.

DCE-MRI-derived blood volume (BV) is emerging as a promising QI metrics in assessing therapeutic response in head and neck (HN) cancers (18, 19). Tumor subvolumes characterized by low BV have been reported to be high-risk imaging biomarkers for tumor progression (19-22). Boosting those poorly perfused subvolumes with high radiation doses could improve local and regional control (23, 24). To test this clinical hypothesis, a randomized phase-II adaptive radiation therapy (RT) trial that targets persisting poorly perfused subvolumes of the tumor with high radiation doses in patients with poor prognosis HN cancers has been initiated (21, 22, 25). The persisting poorly perfused tumor subvolumes are defined on the basis of BV measurements pre-RT and 2 weeks after starting RT. Inaccurate and unrepeatable estimates of BV maps could generate false, poorly perfused subvolumes. Subsequently, intensifying radiation doses to these falsely classified subvolumes can lead to either tumor overdose or underdose, which could increase radiation toxicity or cause failure of disease control, respectively. To achieve the goal of the clinical trial, it is critical to ensure accuracy and precision of BV maps in each individual patient and thereby warrant proper segmentation of low BV tumor subvolumes.

The present study developed and evaluated a framework for real-time quantitative assessment of accuracy and precision of a QI metrics in individual patients during a clinical trial. The method was applied to DCE-MRI-derived BV maps acquired during an ongoing clinical trial for poor prognosis HN cancers. As the repeatability analysis cannot be done in treated tumor volume owing to expected therapy-caused changes, a normal tissue region in the cerebellum that has little therapy-induced change was used as a reference region for BV measurements and hence to assess the accuracy and precision of BV maps. Our study showed that inaccurate and imprecise BV maps could be detected in real time before clinical decision was made. This method can be extended to other QI metrics and body sites. This process should be a part of the workflow of a clinical trial.

MATERIALS AND METHODS

Human Subjects

Patients with advanced HN cancers were enrolled in an IRB-approved randomized phase-II clinical trial. The patients who have advanced human papillomavirus (HPV)-HN cancers (stage IV) or HPV+ T4/N3 HN cancers (stage III) were eligible for the trial. All patients gave their study-specific informed consent to participate in the trial. Patients underwent MRI scans before RT and after receiving 10 fractions (Fx) of 2 Gy per fraction of radiation.

MR Acquisition

All MRI scans were acquired on a 3 T MR scanner (Magnetom Skyra, Siemens Healthineers, Erlangen, Germany). Each patient underwent scanning in the radiation treatment position on a flat table top using the patient-specific immobilization face mask, head support, and bite bar. MRI series included 2-dimensional multislice pre- and postcontrast T1-weighted images with fat saturation (voxel size: $0.88 \times 0.88 \times 3.3$ mm^3; echo time [TE]/repetition time [TR] = 8.4/1040 milliseconds), 2-dimensional T2-weighted images (voxel size: $0.78 \times 0.78 \times 3.3$ mm^3; TE/TR = 89/11000 milliseconds), and 3-dimensional (3D) volumetric T1-weighted DCE images. The DCE image volumes were acquired using a 3D gradient-echo sequence in the sagittal orientation with a large field of view (FOV) in the superior and inferior directions to cover primary and nodal cancers, carotid artery, and cerebellum. The sagittal orientation allows us to achieve higher temporal resolution and avoid time-of-flight effects of blood-flow spins (Figure 1). Other acquisition parameters included flip angle/TE/TR = 10°/0.97/2.73 milliseconds, FOV = $300 \times 300 \times 150$ mm^3, and voxel size ≈ $1.6 \times 1.6 \times 2.5$ mm^3. Sixty dynamic scans were collected at 3 minute, with a temporal resolution of 3 second.

Extended Tofts Model for BV Quantification

Plasma volume maps (v_p) were generated from the T1-weighted DCE image series using the extended Tofts model (26);

$$C_{tiss}(t) = K^{trans}\int_0^t e^{-k_{ep}(t-\tau)}C_p(\tau)d\tau + V_pC_p(t), \qquad (1)$$

where $C_{tiss}(t)$ and $C_p(t)$ were respective tissue and plasma concentrations of the contrast agent, and K^{trans} and k_{ep} were respective transfer constant and rate. An assumption of $\Delta S/S_0 \propto C$ was used to fit equation (1). In-house software package of functional image analysis tool (FIAT) was used for image analysis and processing to generate parametric maps (20, 21), in which the implemented extended Tofts model has been validated using digital reference object (DRO) (5). To convert the plasma volume maps to the BV maps, a Hematocrit value of 0.45 was applied (27). A protocol-specific procedure of DCE analysis was established before initiation of the clinical trial, particularly regarding how to create an AIF. To obtain the AIF, a dynamic phase in which contrast just entered the carotid artery was chosen by visually inspecting the temporal profile of the dynamic image volumes. Then, an AIF was generated by thresholding 20 voxels with the largest intensity changes on the selected phase com-

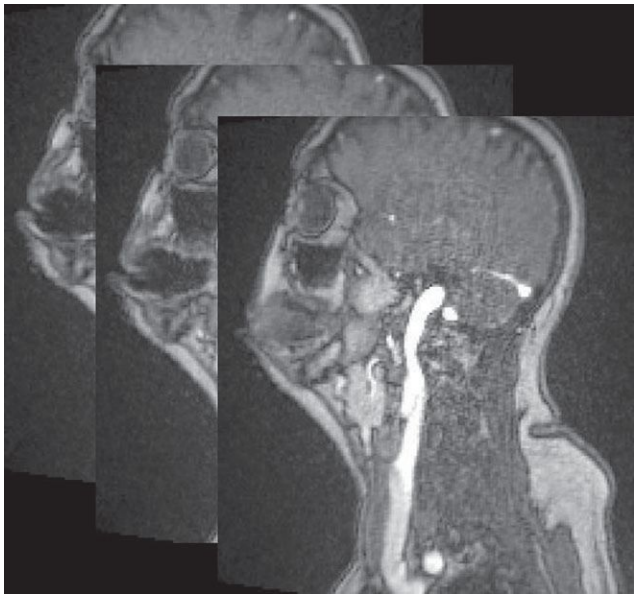


Figure 1. T1-weighted dynamic contrast-enhanced (DCE) images acquired using a 3-dimensional gradient-echo sequence in the sagittal orientation. As shown in the figure, these images were collected with a large field-of-view (FOV) in the superior and inferior directions to cover the primary and nodal tumors, carotid artery, and normal tissue region in cerebellum. The latter region, that is, the normal cerebellum region, was used as a reference region for quality assessment of blood volume (BV) measurement in each individual examination.

pared with the average baseline image intensities. Finally, the AIF was visually inspected to make sure that its voxels were located within the carotid artery and had the expected dynamic profile. BV maps were derived from the extended Tofts model using the patient-specific AIF, and then coregistered to the postcontrast T1-weighted images at pre-RT using rigid-body transformation (20).

System-Level QA

To ensure quality of quantitative parametric maps, QA of hardware and software at system-level was performed routinely. System-level QA of the MRI scanner was performed daily, weekly, and yearly using an ACR water phantom following the ACR protocol. Daily signal-to-noise ratio variations were recorded and were stable. Also, in an NCI Quantitative Imaging Network (QIN) multicenter collaborative project, we evaluated accuracy, repeatability, and interplatform reproducibility of T1 quantification from variable flip angles using an NIST T1 water phantom on our scanner, compared to others (4). For software QA, performance of our implementation of the extended Tofts model was evaluated using a digital reference object, that is, synthesized DCE phantoms with and without noise, which was fully reported previously (5). Also, we participated in an NCI QIN multicenter AIF challenge to validate and compare our AIF delineation procedure with others' (15). Based upon these evaluation and validation, *imFIAT* has been granted a level-2 benchmark by NCI QIN (28).

Individual-Level Assessment of Accuracy and Precision of BV Maps

Our pilot study indicates that repeatability of BV values in the cerebellum is stable and ~18% (unpublished data). Also, cerebellums in our patients received a mean radiation dose <3Gy after 10 Fx of 2 Gy treatment. Therefore, we chose cerebellum as a reference region and manually drew bilateral volumes of interest (VOIs) across 2–3 slices having a volume of ~4 cc

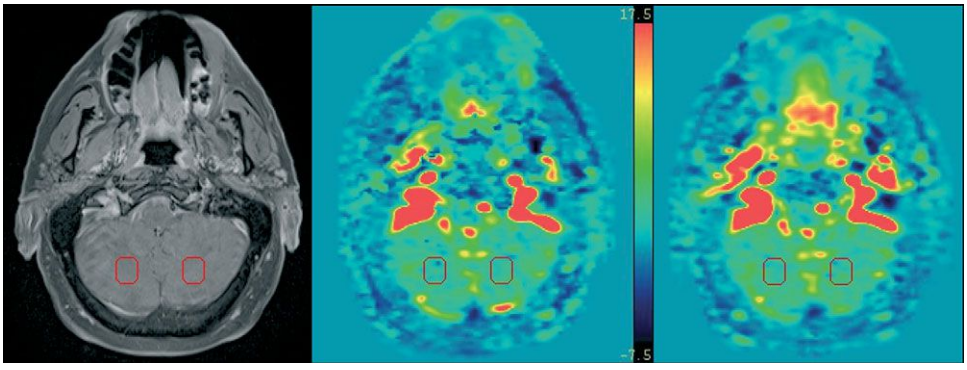


Figure 2. The coregistered postcontrast T1-weighted image (left), and the BV maps pre-radiation therapy (RT) (middle) and after 10 fractions of radiation therapy (right) from a sample study. The postcontrast T1-weighted images was used to delineate the tumor volumes and to locate the normal cerebellum region as a reference region. Red contours (~4cc in volume) represent the volumes of interest (VOIs) in the normal cerebellum, which was used as the reference region for the accuracy and precision analysis.

(number of voxels, ~1600) to extract mean BV values (Figure 2).

For each patient, MRI scanning was performed pre-RT and repeated after 10 Fx of radiation (2wkRT), which were considered as test and retest studies. An RC of BV values in the cerebellum VOIs was estimated using 1-way analysis of variance (ANOVA) model (29). First, within-subject mean squares (WMS) was estimated from n patients. Then RC and relative RC were estimated by $RC = 2.77 \times \sqrt{WMS}$ and $rRC = 100 \times RC/\hat{X}$, respectively, where $\hat{X}$ was the grand mean of overall observations from n patients. Because the WMS for 2 repeated measurements was distributed as $\chi^2_n wSD^2/n$, the 95% confidence interval (CI) of the estimated RC was given by $RC_L = RC \times \sqrt{n/\chi^2_n(0.975)}$ and $RC_U = RC \times \sqrt{n/\chi^2_n(0.025)}$, where $\chi^2_n(a)$ was the a^{th} percentile of the χ^2 distribution with n degrees of freedom.

To assess accuracy and precision of BV values in each individual patient, a group mean (M_n) of BV in cerebellum VOIs as a reference value with a 95% CI defined by standard deviation (SD_n), and an RC_n with a 95% CI defined by RC_L and RC_U were computed from n patients. For the next new scan, it was determined whether the mean BV value in the cerebellum VOI was between $M_n - 2SD_n$ and $M_n + 2SD_n$. If yes, the BV map was deemed accurate with 95% confidence. For each new patient, a difference of BV between the 2 scans (test and retest) was determined whether it was within $-RC_n$ and RC_n. If yes, the BV maps of this new patient were considered repeatable with 95% confidence. When the new patient's data passed both tests, the BV maps could be used to update the reference value and RC. Otherwise, the BV maps from this individual patient were flagged for further evaluation or correction before used in the clinical trial.

Other Statistical Analysis

A paired t test was performed to examine whether there was any difference between mean BVs measured at test and a retest with P-value <0.05 as statistically significant. The distribution of differences in mean BV values between the 2 scans was tested for normality using the Shapiro–Wilk test. Similarly, to detect a potential relationship between the measurement error and the magnitude of the combined mean BV values between 2 scans, a rank correlation coefficient (Kendall's tau) test between absolute differences against their combined means was performed.

Association Between Repeatability of AIF Peak and BV

As noted, we used a fixed imaging protocol to minimize variations in acquisition. However, it was unknown how repeatability of AIF was associated with repeatability of BV values. To examine this association, we measured the AIF peak value for each scan and calculated the RC from the 2 scans. We compared percentage differences of AIF peaks between the 2 scans with those of BV values measured in the cerebellum VOIs.

RESULTS

At the time of this report, 62 consecutive patients (median age, 62 years; male, 52; female, 10) were enrolled in the clinical trial. For the first 10 patients, the mean (±SD) BV values from test and retest were 2.22 (±0.13) mL/100 g and 2.21 (±0.19) mL/100 g, respectively, and not significantly different (P-value = 0.79: paired t test), yielding the overall group mean (±SD) of 2.21 (±0.16) mL/100 g (see Table 1). The difference in the BV values between test–retest studies was independent to the combined mean (P-value = 0.21: Kendall tau test), indicating that the measurement error was independent to the magnitude of measured BV values. Also, the Shapiro–Wilk test showed that the differences in BV values between the 2 examinations were normally distributed. An RC of BV values between the 2 tests was estimated to be 0.37, yielding a relative RC (rRC) of 16.7% with a 95% CI of (11.7%, 29.4%). Using the leave-1-out cross-validation, we did not find any outlier from the first 10 patients. Therefore, we used M_{10} and RC_{10} as starting reference values to evaluate the next patient (Table 1).

Table 1. Summary Statistics for BV Measurement at Normal Cerebellum Region

Statistical Parameters	Preliminary Statistics (n = 10)	Updated Statistics (n = 58)
Mean BV (±SD) (mL/100 g)		
Test study	2.22 (±0.13)	2.21 (±0.14)
Retest study	2.21 (±0.19)	2.22 (±0.17)
Overall	2.21 (±0.16)	2.22 (±0.15)
Paired t test (P-value)	0.79	0.73
Kendall's tau test (P-value)	0.21	0.67
WMS	0.02	0.02
RC (rRC%)	0.37 (16.7)	0.35 (15.9)
95%CI on rRC (%): rRC_L, rRC_U	11.7, 29.4	13.5, 19.5

BV measurements from 62 patients were evaluated in real time, and 3 patients were identified to have inaccurate BV values in 1 of the 2 scans (Figure 3). Mean BVs measured from these 3 patients were in the range of 3.05–3.95 mL/100 g, which were much higher than those measured from the group mean + 2 × SD value (2.52 mL/100 g). The repeatability tests found that the percentage differences of BV values between the 2 scans of the 3 patients were much greater than the uncertainty range defined by −RC and RC. Note that our procedure detected large variations of BV values in 3 scans in real time, but not in retrospective analysis. The consequences of the BV maps for decision-making with and without correction were evaluated and discussed with the physicians during the clinical trial.

As the patients were enrolled into the clinical trial, the data from the 3 patients were excluded from the updated reference values for accuracy and precision measurements. One additional patient who had BV values within the normal range for both test and retest was excluded owing to partial coverage of cerebellum in 1 scan and mismatched slices in cerebellum between the 2 scans. As a result, the data from 58 patients were included to update the reference values. A group mean (±SD) of BV values was of 2.21 (±0.14) mL/100 g at test, and 2.22 (±0.17) mL/100 g at retest, which were not significantly different (P-value = 0.

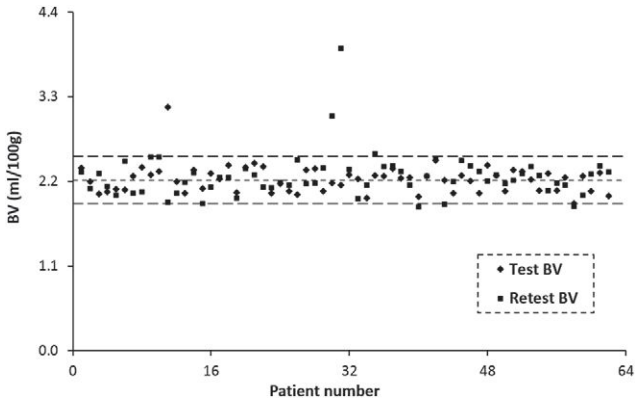


Figure 3. Mean BV values obtained in the cerebellum VOIs in each study plotted against the patient number. A center dotted line represents the overall group mean of BV, while 2 dashed lines depict the 95% confident interval (±1.96 × SD from the group mean). Note that 3 BV values are far away from the confident range, and are identified as inaccurate BV measurements.

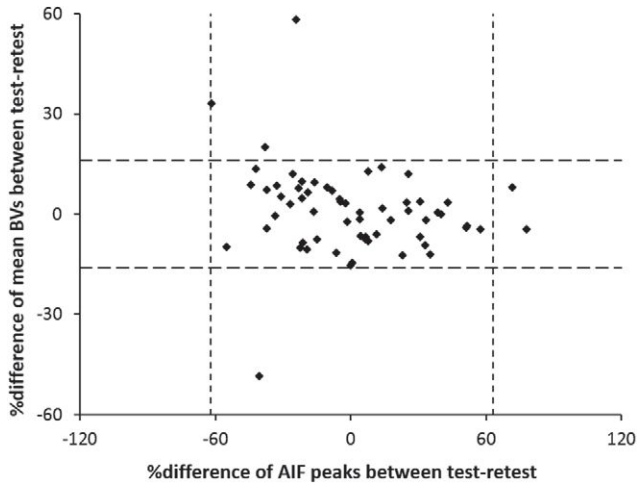


Figure 5. Scatter plot of percentage differences of mean BVs versus percentage differences of arterial input function (AIF) peaks between the 2 scans, with their corresponding RC ranges (horizontal dashes lines for BV and vertical ones for AIF peak).

73: paired *t* test; see Table 1), suggesting stability of the quantified BV maps. Also, the absolute difference was independent of their combined means (*P*-value = 0.67: Kendall tau test). ANOVA led to an RC of 0.35, and an rRC of 15.9% with a 95% CI of (13.5%, 19.5%). Note that the 95% CI (uncertainty) of estimated RC decreased with an increase in the number of patients. Figure 4 shows a plot of percentage differences of BV values between test–retest studies versus their combined means. As shown in the plot, percentage differences of mean BVs from the 3 patients, who had inaccurate mean BVs, were much large than the RC interval (% difference > 33% at the lowest), indicating the imprecision in the repeated measures.

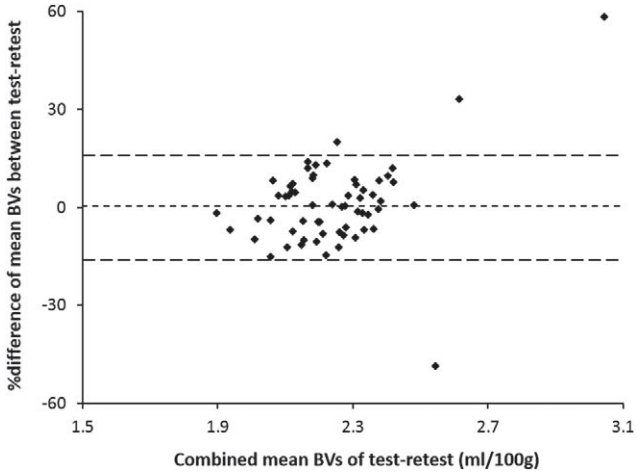


Figure 4. Bland–Altman plots of the percentage difference in mean BV between 2 studies plotted against their combined mean. A center dotted line shows the mean percentage difference of BV between the 2 scans, while 2 dashed lines represent the estimated repeatability coefficient (RC) interval (−RC, RC).

Finally, the relative RC of the AIF peak values was of 61.8%. Figure 5 shows a scatter plot of percentage differences of BV values in the cerebellum VOIs versus those of AIF peak values between the 2 scans. Note that there was no association or even a trend between the 2 differences, suggesting the variation of AIF peaks could not explain the variation in BV measurements.

DISCUSSION

In this study, we developed and evaluated a methodology and metrics for real-time quantitative assessment of accuracy and precision on DCE-MRI derived metrics using reference values in a normal reference tissue region. It is critical to establish such a real-time QA test in the workflow of a clinical trial to identify unreliable estimates of QI metrics before used in a trial. A subsequent action should be planned in the design of a clinical trial. A real-time QA procedure of QI metrics in individual patients would enhance the ability of the trial to achieve its objectives and increase reliability of scientific findings. Our method can be extended to other QI metrics and body-sites to support individualized therapy and improve therapeutic outcomes.

It would be worth noting that accuracy and precision of BV values investigated in this study do not represent how accurate the QI metrics measure a true physiological BV. As discussed in the Introduction, they are measures of bias and variation of BV values as a QI metrics quantified from HN DCE-MRI using the extended Tofts model to reference values. Our data show that the

group mean and RC of BV values in the cerebellum are stable, suggesting that it is a great candidate used as a reference region. As anticipated, the 95% CI of estimated RC decreases with an increase in the sample size. Using these reference values, we are able to detect unreliable QI measures of individual patients in real time during the clinical trial. Our test is different from test and retest analysis performed before therapy. The latter helps us understand the general technical behavior of a QI metrics in a sample of population, but it does not tell us whether the metrics acquired in each patient in a clinical trial is reliable or not. Finally, the impact of uncertainty of a QI metrics in a decision-making process needs to be investigated in future.

As shown in this study, reference values have to be established in a reference tissue region to perform the proposed QA test. The reference tissue region chosen may depend upon the image type and body site of interest. However, the QI metrics in a reference region has to be stable, less affected by therapy, and within the FOV of the scan. In our preliminary investigation, we tested sternocleidomastoid muscle (SCM) contralateral to tumor as a possible tissue reference region. We found that the BV values in SCM were not as stable as those in the cerebellum, possibly owing to low BV in SCM. Also, in some cases, tumors are distributed bilaterally, in which there is no noninvolved SCM that can be used as a reference region. On the other hand, the cerebellum tissue receives few Gy radiation doses (<3 Gy) for HN cancer treatment, and BV changes in cerebellum VOIs after 10 Fx of RT do not show any positive or negative trend (Figure 4), suggesting that the treatment effect within the cerebellum is minimum and can be ignored. Reference values of BV in the cerebellum VOIs are adequate for evaluation of the overall quality of BV maps, as MRI data are acquired in the k-space and BV maps are determined by a single AIF. However, local motion, e.g., swallowing, can cause local degradation in DCE-MRI, which cannot be captured by the analysis performed in the normal reference region. However, it still needs to be cautious to use QI metrics during a therapeutic trial.

In our study, patient positioning, scanner, image protocol, acquisition procedure, and analysis software and process are controlled carefully to maintain consistency of QI metrics delineation during the clinical trial. The factors that can influence repeatability of DCE-MRI-derived QI metrics include patient positioning, image registration, AIF delineation, image noise, image process, treatment effect, and unknown physiological fluctuation. We further investigated repeatability of AIF peaks, as well as its influence on repeatability of BV maps, but found no relationship among differences in the BV values and the AIF peaks between the 2 scans (Figure 5). These findings indicate that the AIF peak variation cannot solely explain one in the BV measures.

In conclusion, the present study developed and evaluated a methodology for quantitative assessment of accuracy and precision of DCE-MRI derived BV maps in a phase-II randomized clinical trial for poor prognosis HN cancers. The outlined framework was able to detect outliers, that is, identify the individual patients who had unreliable BV values in real time during the clinical trial. Because accuracy and precision of QI metrics influence decision-making in the individualized and adaptive cancer therapy, individual QA testing of such QI metrics needs to be integrated into a clinical trial workflow to warrant success of the trial.

ACKNOWLEDGMENTS

This work is supported in part by NIH/NCI grants of 1U01CA183848 and RO1CA184153.

Disclosures: No disclosures to report.

Conflict of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. Abramson RG, Arlinghaus LR, Dula AN, Quarles CC, Stokes AM, Weis JA, Whisenant JG, Chekmenev EY, Zhukov I, Williams JM, Yankeelov TE. MR imaging biomarkers in oncology clinical trials. Magn Reson Imaging Clin N Am. 2016;24:11–29.
2. NCT02031250, NCT00581906, NCT02070705, and NCT02878109. Available on National Institute Health US. National Library of Medicine Clinical Trials. 2018.
3. Kim H. Variability in quantitative DCE-MRI: sources and solutions. J Nat Sci. 2018;4. pii: e484.
4. Bane O, Hectors SJ, Wagner M, Arlinghaus LL, Aryal MP, Cao Y, Chenevert TL, Fennessy F, Huang W, Hylton NM, Kalpathy-Cramer J, Keenan KE, Malyarenko D, Mulkern RV, Newitt DC, Russek SE, Stupic KF, Tudorica A, Wilmes LJ, Yankeelov TE, Yen YF, Boss MA, Taouli B. Accuracy, repeatability, and interplatform reproducibility of T1 quantification methods used for DCE-MRI: Results from a multicenter phantom study. Magn Reson Med. 2018;79:2564–2575.
5. Cao Y, Li D, Shen Z, Normolle D. Sensitivity of quantitative metrics derived from DCE MRI and a pharmacokinetic model to image quality and acquisition parameters. Acad Radiol. 2010;17:468–478.
6. Huang W, Li X, Chen Y, Li X, Chang MC, Oborski MJ, Malyarenko DI, Muzi M, Jajamovich GH, Fedorov A, Tudorica A, Gupta SN, Laymon CM, Marro KI, Dyvorne HA, Miller JV, Barbodiak DP, Chenevert TL, Yankeelov TE, Mountz JM, Kinahan PE, Kikinis R, Taouli B, Fennessy F, Kalpathy-Cramer J. Variations of dynamic contrast-enhanced magnetic resonance imaging in evaluation of breast cancer therapy response: a multicenter data analysis challenge. Transl Oncol. 2014;7:153–166.
7. Heye T, Davenport MS, Horvath JJ, Feuerlein S, Breault SR, Bashir MR, Merkle EM, Boll DT. Reproducibility of dynamic contrast-enhanced MR imaging. Part I. Perfusion characteristics in the female pelvis by using multiple computer-aided diagnosis perfusion analysis solutions. Radiology. 2013;266:801–811.
8. Sourbron SP, Buckley DL. Tracer kinetic modelling in MRI: estimating perfusion and capillary permeability. Phys Med Biol. 2012;57:R1–R33.
9. Heisen M, Fan X, Buurman J, van Riel NA, Karczmar GS, ter Haar Romeny BM. The influence of temporal resolution in determining pharmacokinetic parameters from DCE-MRI data. Magn Reson Med. 2010;63:811–816.
10. Kurland BF, Gerstner ER, Mountz JM, Schwartz LH, Ryan CW, Graham MM, Buatti JM, Fennessy FM, Eikman EA, Kumar V, Forster KM, Wahl RL, Lieberman FS. Promise and pitfalls of quantitative imaging in oncology clinical trials. Magn Reson Imaging. 2012;30:1301–1312.
11. Obuchowski NA, Reeves AP, Huang EP, Wang XF, Buckler AJ, Kim HJ, Barnhart HX, Jackson EF, Giger ML, Pennello G, Toledano AY, Kalpathy-Cramer J, Apanasovich TV, Kinahan PE, Myers KJ, Goldgof DB, Barboriak DP, Gillies RJ, Schwartz LH, Sullivan DC, Algorithm Comparison Working Group. Quantitative imaging biomarkers: a review of statistical methods for computer algorithm comparisons. Stat Methods Med Res. 2015;24:68–106.
12. Raunig DL, McShane LM, Pennello G, Gatsonis C, Carson PL, Voyvodic JT, Wahl RL, Kurland BF, Schwarz AJ, Gönen M Zahlmann G, Kondratovich MV, O'Donnell K, Petrick N, Cole PE, Garra B, Sullivan DC; QIBA Technical Performance Working Group. Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment. Stat Methods Med Res. 2015;24:27–67.

13. Keenan KE, Ainslie M, Barker AJ, Boss MA, Cecil KM, Charles C, Chenevert TL, Clarke L, Evelhoch JL, Finn P, Gembris D, Gunter JL, Hill DLG, Jack CR, Jr., Jackson EF, Liu G, Russek SE, Sharma SD, Steckner M, Stupic KF, Trzasko JD, Yuan C, Zheng J. Quantitative magnetic resonance imaging phantoms: A review and the need for a system phantom. Magn Reson Med. 2018;79:48–61.
14. QIBA. Profile: DCE MRI quantification version 1.0; Accessed June 28, 2011. Available from: http://www.rsna.org/uploadedFiles/RSNA/Content/Science_and_Education/QIBA/DCE-MRI_Quantification_Profile_v1%200-Reviewed-Draft%208-8-12.pdf.
15. Huang W, Chen Y, Fedorov A, Li X, Jajamovich GH, Malyarenko DI, Aryal MP, LaViolette PS, Oborski MJ, O'Sullivan F, Abramson RG, Jafari-Khouzani K, Afzal A, Tudorica A, Moloney B, Gupta SN, Besa C, Kalpathy-Cramer J, Mountz JM, Laymon CM, Muzi M, Schmainda K, Cao Y, Chenevert TL, Taouli B, Yankeelov TE, Fennessy F, Li X. The impact of arterial input function determination variations on prostate dynamic contrast-enhanced magnetic resonance imaging pharmacokinetic modeling: a multicenter data analysis challenge. Tomography. 2016;2:56.
16. QIBA. Synthetic DCE-MRI Data. Available from: https://qibawikirsnaorg/index-php/Synthetic_DCE-MRI_Data. 2009.
17. Cercignani M, Dowell NG, Tofts PS. Quantitative MRI of the Brain; Quality Assurance: Accuracy, Precision, Controls and Phantoms 1. 2nd ed. CRC Press: Boca Raton, Florida; 2018.
18. Cao Y, Popovtzer A, Li D, Chepeha DB, Moyer JS, Prince ME, Worden F, Teknos T, Bradford C, Mukherji SK, Eisbruch A. Early prediction of outcome in advanced head-and-neck cancer based on tumor blood volume alterations during therapy: a prospective study. Int J Radiat Oncol Biol Phys. 2008;72:1287–1290.
19. Agrawal S, Awasthi R, Singh A, Haris M, Gupta R, Rathore R. An exploratory study into the role of dynamic contrast-enhanced (DCE) MRI metrics as predictors of response in head and neck cancers. Clin Radiol. 2012;67:e1–e5.
20. Cao Y. WE-D-T-6C-03: development of image software tools for radiation therapy assessment. Med Phys. 2005;32:2136–2136.
21. Wang P, Popovtzer A, Eisbruch A, Cao Y. An approach to identify, from DCE MRI, significant subvolumes of tumors related to outcomes in advanced head-and-neck cancer. Med Phys. 2012;39:5277–5285.
22. Hawkins PG, Lee JY, Lee C, Green M, Mierzwa ML, Aryal MP, Arnould GS, Worden F, Swiecicki PL, Spector ME, Schipper M, Cao Y, Eisbruch A. Adaptive chemoradiation therapy for head and neck cancer based on multiparametric MRI: interim results of a prospective randomized trial. Int J Radiat Oncol Biol Phys. 2017;99:E339.
23. Wang J, Zheng J, Tang T, Zhu F, Yao Y, Xu J, Wang AZ, Zhang L. A randomized pilot trial comparing position emission tomography (PET)-guided dose escalation radiotherapy to conventional radiotherapy in chemoradiotherapy treatment of locally advanced nasopharyngeal carcinoma. PloS One. 2015;10:e0124018.
24. Overgaard J. Hypoxic modification of radiotherapy in squamous cell carcinoma of the head and neck–a systematic review and meta-analysis. Radiother Oncol. 2011;100:22–32.
25. Teng F, Aryal M, Lee J, Lee C, Shen X, Hawkins P, Mierzwa M, Eisbruch A, Cao Y. Adaptive boost target definition in high-risk head and neck cancer based on multi-imaging risk biomarkers. Int J Radiat Oncol Biol Phys. 2018;102:969–977.
26. Tofts PS, Brix G, Buckley DL, Evelhoch JL, Henderson E, Knopp MV, Larsson HB, Lee TY, Mayr NA, Parker GJ, Port RE, Taylor J, Weisskoff RM. Estimating kinetic parameters from dynamic contrast-enhanced T(1)-weighted MRI of a diffusable tracer: standardized quantities and symbols. J Magn Reson Imaging. 1999;10:223–232.
27. Sourbron SP, Buckley DL. Classic models for dynamic contrast-enhanced MRI. NMR Biomed. 2013;26:1004–1027.
28. Farahani K, Kalpathy-Cramer J, Chenevert TL, Rubin DL, Sunderland JJ, Nordstrom RJ, Buatti J, Hylton N. Computational challenges and collaborative projects in the NCI Quantitative Imaging Network. Tomography. 2016;2:242–249.
29. Barnhart HX, Barboriak DP. Applications of the repeatability of quantitative imaging biomarkers: a review of statistical analysis of repeat data sets. Transl Oncol. 2009;2:231–235.

TOMOGRAPHY® RESEARCH ARTICLE

Habitats in DCE-MRI to Predict Clinically Significant Prostate Cancers

Nestor Andres Parra[1], Hong Lu[1,2], Jung Choi[3], Kenneth Gage[3], Julio Pow-Sang[4], Robert J. Gillies[1,3], and Yoganand Balagurunathan[1]

Departments of [1]Cancer Physiology, [3]Radiology, and [4]Urology, H.L. Moffitt Cancer Center, Tampa, FL; and [2]Department of Radiology, Tianjin Medical University Cancer Institute and Hospital, Tianjin, China

Corresponding Author:
Yoganand Balagurunathan, PhD
Cancer Physiology, H.L. Moffitt Cancer Center,
12902 USF Magnolia Ave, Tampa, FL 33612;
E-mail: Yoganand.Balagurunathan@moffitt.org.

Key Words: MRI, prostate cancer, machine learning, radiomics, habitats, DCE

Abbreviations: Dynamic contrast-enhanced imaging (DCE), prostatic-specific antigen, multiparametric magnetic resonance imaging (mpMRI), transrectal ultrasonography (TRUS), T2-weighted (T2W), diffusion-weighted imaging (DWI), peripheral zone (PZ), repetition time (TR), echo time (TE), endorectal coil (ERC), support vector machines (SVMs), regions of interest (ROIs), Area Under the Receiver Operating Characteristic curve (AUC), Non-negative matrix factorization (NMF)

ABSTRACT

Prostate cancer identification and assessment of clinical significance continues to be a challenge. Routine multiparametric magnetic resonance imaging has shown to be useful in assessing disease progression. Although dynamic contrast-enhanced imaging (DCE) has the ability to characterize perfusion across time and has shown enormous utility, radiological assessment (Prostate Imaging-Reporting and Data System or PIRADS version 2) has limited its use owing to lack of consistency and nonquantitative nature. In our work, we propose a systematic methodology to quantify perfusion dynamics for the DCE imaging. Using these metrics, 7 different subregions or *perfusion habitats* of the targeted lesions are localized and related to clinical significance. We found that quantitative features describing the habitat based on the late area under the DCE time-activity curve was a good predictor of clinical significance disease. The best predictive feature in the habitat had an AUC of 0.82, CI [0.81–0.83].

INTRODUCTION

Prostate cancer is the second leading cause of cancer deaths among men in the United States and accounts to be the third largest among newly diagnosed cancer cases (19%) (1). Rising prostatic-specific antigen and abnormal digital-rectal examination have been traditionally used in the diagnosis of prostate cancer. Advent of improved imaging resulted in the inclusion of multiparametric magnetic resonance imaging (mpMRI) in the clinical workflow (2). Recently, the United States Preventive Services Task Force (USPSTF) has recommended against the routine use of prostatic-specific antigen testing for diagnosis of prostate cancer, owing to the risk of overdiagnosis and overtreatment (3, 4). Advancements in image acquisition and resolution of mpMRI coupled with the use of fusion-based transrectal ultrasonography (TRUS)– guided biopsy has improved disease detection and shown promise in improving diagnosis and treatment (5). Routine MP-MRI includes T2-weighted (T2W) imaging that describes the prostate anatomy, diffusion-weighted imaging (DWI) that measures the density of cellular space by quantifying the diffusion of water molecules. DCE image data shows the dynamics of the administered contrast agent, which characterizes the blood flow into prostate tissue and allows the identification of suspicious lesions by localizing abnormal contrast absorption.

DCE analysis can be quantitative or semiquantitative. The first approach is based on a contrast concentration model used to determine the rate of contrast transfer from the blood plasma into the tissue's extravascular extracellular space (6, 7). The second approach describes different contrast absorption patterns based on the characteristics of time–activity curves (8-10). The Prostate Imaging-Reporting and Data System (PI-RADSv2) currently includes DCE along with T2W and DWI, but their added value in diagnosis seems to be limited (11). PI-RADSv2 limits the use of DCE to the peripheral zone (PZ) when DWI is not conclusive. The standard limits the use of DCE to a single binary observation: presence or absence of uptake. This could be attributed to the lack of consensus in the community to use better metrics. Traditionally, these DCE curves are qualitatively characterized, which includes wash-in and wash-out slopes and time-to-peak (12). These have been related to tumor aggressiveness (13). The difficulty in establishing consistent features from the DCE curves, as well as the high interobserver variability, has limited the use of DCE in a quantitative fashion. Nonetheless, there have been successful efforts to semiquantitatively characterize DCE and use these parameters for classification of prostate cancer aggressiveness (14, 15).

Recently, radiomic analysis of habitats defined by textural kinetic features has been used to predict recurrence-free survival in patients with breast cancer (16). DCE-based habitats have shown to correlate with estrogen receptor and nodal metastatic status in breast cancer. Habitats in MRI imaging have also been useful in identifying disease progression in glioblastoma (17).

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2018.00037

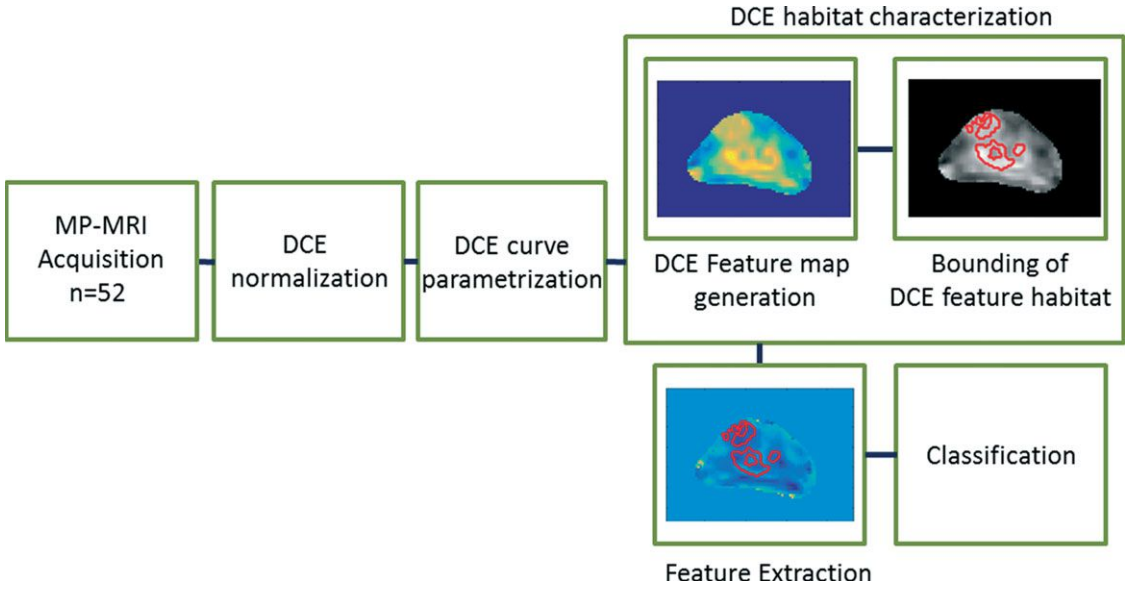


Figure 1. Block diagram shows the DCE habitat identification and processing. A perfusion tumor habitat was localized for each DCE feature map and these regions were characterized (by DCE features). Classification models were applied to identify features that can discriminate clinically significant prostate cancers.

MRI-defined features have been used to define radiotherapy treatment planning in prostate cancer (18).

In this study, we obtained the tumor region based on radiologist delineation on a T2W sequence. The region was centered on the TRUS biopsy location that was imported directly from the fused TRUS/MRI system. DCE characteristics at voxel level, across time, were quantified. Each feature map was used to form a habitat or localization of voxels. These new habitat regions were limited by a boundary around the known biopsy location that was quantified. The ability of the features to discriminate clinically significant cancers was evaluated for these specific habitats. Figure 1 shows the methodology followed.

METHODS

Patients and MRI Acquisition

Patient imaging and histopathology records were collected at H. Lee Moffitt Cancer Center, retrospective investigatory protocol approved by the University of South Florida IRB. Informed consent was waived for retrospective access of deidentified patient records. The study included patients that had MRI-guided targeted biopsy acquired between November 2015 and February 2018. Suspicious lesions were marked by a clinical radiologist on MRI. The patients in the study cohort had at least one biopsy with an assigned Gleason Score (GS) sum ≥6. The data set consisted of 72 biopsies from 54 patients. The average interval between imaging and biopsy sampling was 27 days. In this study, patients were grouped in 2 categories: *clinically insignificant cancer* (GS = 6) and *clinically significant cancer* (GS ≥ 7). All statistics were performed using this grouping. The data set consisted of 25 clinically insignificant and 47 clinically significant biopsies.

MRI Acquisition and DCE Normalization

Routine clinical MP-MRI acquisition includes T2-weighted imaging (T2W), DCE, and DWI. The DWI includes an apparent diffusion coefficient (ADC) map generated at the time of acquisition. Patients were injected with contrast agent Gadavist (Bayer HealthCare, Whippany, NJ) with a dose of 0.1 mL/kg before MRI-DCE acquisition. In total, 27 patients were imaged using a Siemens –SymphonyTim (Siemens, Munich, Germany) scanner at 1.5 T and endorectal coil (ERC) (eCoil, Medrad, Pittsburgh, PA) with median repetition time (TR) of 7.7 seconds (range, 6.4–9.5 seconds) and median echo time (TE) of 95 milliseconds (range, 94–95 milliseconds) for DWI. For DCE, TR was 4.72 milliseconds, TE was 1.34 milliseconds, flip angle was 12°, and temporal resolution was 11.45 seconds. Twenty-two patients were imaged using a Siemens-Skyra (Siemens, Munich, Germany) scanner at 3 T and a pelvic phased-array coil with a median TR of 4.6 seconds (range, 4.5–5.8 seconds) and a median TE of 77 milliseconds (range, 67–84 milliseconds) for DWI. For DCE, the median TR was 4.5 milliseconds (range, 4.5–5.08 milliseconds), the median TE was 1.71 milliseconds (range 1.71-1.87 milliseconds), flip angle was 12° (n = 20), and 15° (n = 2); temporal resolution was 11.45 seconds (n = 20) and 13.75 seconds (n = 2). Three patients were imaged on a Philips-Ingenia scanner at 3 T and a pelvic phased-array coil with a median TR of 6.0 seconds (range, 4.5–6.2 seconds) and a median TE of 114 milliseconds (range, 91–114 milliseconds) for DWI. For DCE, the median TR was 4.21 milliseconds (range, 3.56–4.28 milliseconds), the median TE was 2.02 milliseconds (range, 1.62–2.08 milliseconds), flip angle was 10°, and temporal resolution was 13.75 seconds. In summary, 27 patients (38 biopsies, 14 clinically insignificant and 24 clinically signifi-

Table 1. Patients Enrolled in the Study With Their Biopsies Clinical Status and Scanner Differences

	Patients	Biopsies	Clinically Insignificant	Clinically Significant
1.5 T/ERC	27	38	14	24
3 T	25	34	11	23
Total	52	72	25	47

cant) were imaged at 1.5 T with ERC and 25 patients (34 biopsies, 11 clinically insignificant and 23 clinically significant) at 3 T with a phased-array pelvic coil (Table 1).

Image registration against the T2W image was performed for all modalities using gradient descent of mutual information on the space spanned by 3D affine transformations. Manual contours of the prostate, PZ, and the radiologist finding in the prebiopsy MRI were stored as RT-DICOM structures. The peak-absorption time point S_{peak} was identified in DCE using the AIF (arterial input function) signal as reference. All other time points were registered to S_{peak}. DCE data were normalized using an automatically segmented arterial contour as described in the literature (19), which makes the signal proportional to the change in relaxation rate caused by the contrast agent weighted by the initial spin-lattice relaxation time (20).

DCE-Feature Maps

Seven features were extracted from the DCE time–activity curves, which describe both early and late enhancement (Table 2). DCE time–activity curves were represented using a biexponential semiquantitative model (12) that has the following 5 parameters: initial static intensity s_0, plateau s_m, start of enhancement t_0, time-to-peak *tau*, and wash-out slope, *wo*. The online Supplemental Figure 1 shows an example with the parameters used to characterize the DCE time activity curve. Peak enhancement $s_p = s_m - s_0$, wash-in slope $wi = s_p/tau$. In addition, we computed 2 features that describe the area under the DCE curve between a time interval, namely: $AUC_{t1\text{-}t2}$ is the area under the biexponential fitted DCE curve between time t1 and t2. $AUCi = AUC_{t0\text{-}t0+60}$ measures the early wash-in uptake curve and $AUCf = AUC_{t0+240-t0+270}$ measures the late wash-out curve. The seventh feature computes the multiplicative effect of wash-in and wash-out slopes and was computed as $m_{io} = wi^* wo$. Each one of these parameters was used to generate a 3D DCE-feature map that was used to obtain a habitat (Figure 2).

Habitat Representation

We localized the regions of interest (ROIs) based on each of the 7 DCE feature maps, which includes intra and peritumoral regions around the biopsy location, referred to as *DCE based Habitats*. A sphere (radius $r = 15$ mm) around the biopsy location was placed on each DCE feature map used to bound the tumor habitat. This region was additionally bounded by the prostate zones (PZ or peripheral zone, TZ or transitional zone) allowing convergence of largest lesion volume. The values for each feature map within the localized sphere were used to obtain the region defined by either the lower or upper quartile depending on the feature. The converged habitats were labeled as *H-DCE feature*. The mean DCE signal at the converged habitat region at each sampling time was used as a representative perfusion curve for the patient biopsy. DICE index between each habitat and the radiologist's lesion ROI were computed to assess the volume of intratumoral habitat.

Statistical Analysis

Univariate analysis of the 7 DCE features was performed to evaluate the overall discrimination using support vector machines (SVMs) to discriminate clinically significant cancers. Sensitivity, specificity, and AUC were computed on the habitats (Table 3). Pair-wise multivariable analysis was performed by exhaustive comparison of all possible DCE features. The underrepresented GS class was oversampled using SMOTE (21), cali-

Table 2. List of DCE Features

Number	Feature ID	Feature Description	Dice
1	s_p	Peak enhancement, s_m-s_0	0.22
2	*tau*	Time-to-peak	0.42
3	*wi*	Wash-in slope	0.21
4	*wo*	Wash-out slope	0.25
5	*AUCi*	Initial AUC, $AUC_{t0\text{-}t0+60}$	0.33
6	*AUCf*	Final (late) AUC, $AUC_{t0+240\text{-}t0+270}$	0.22
7	m_{io}	Slope product, $wi \times wo$	0.17

The DCE features were used in this paper to converge a *habitat* from the associated feature map and to characterize the average time activity curve in each *habitat*.

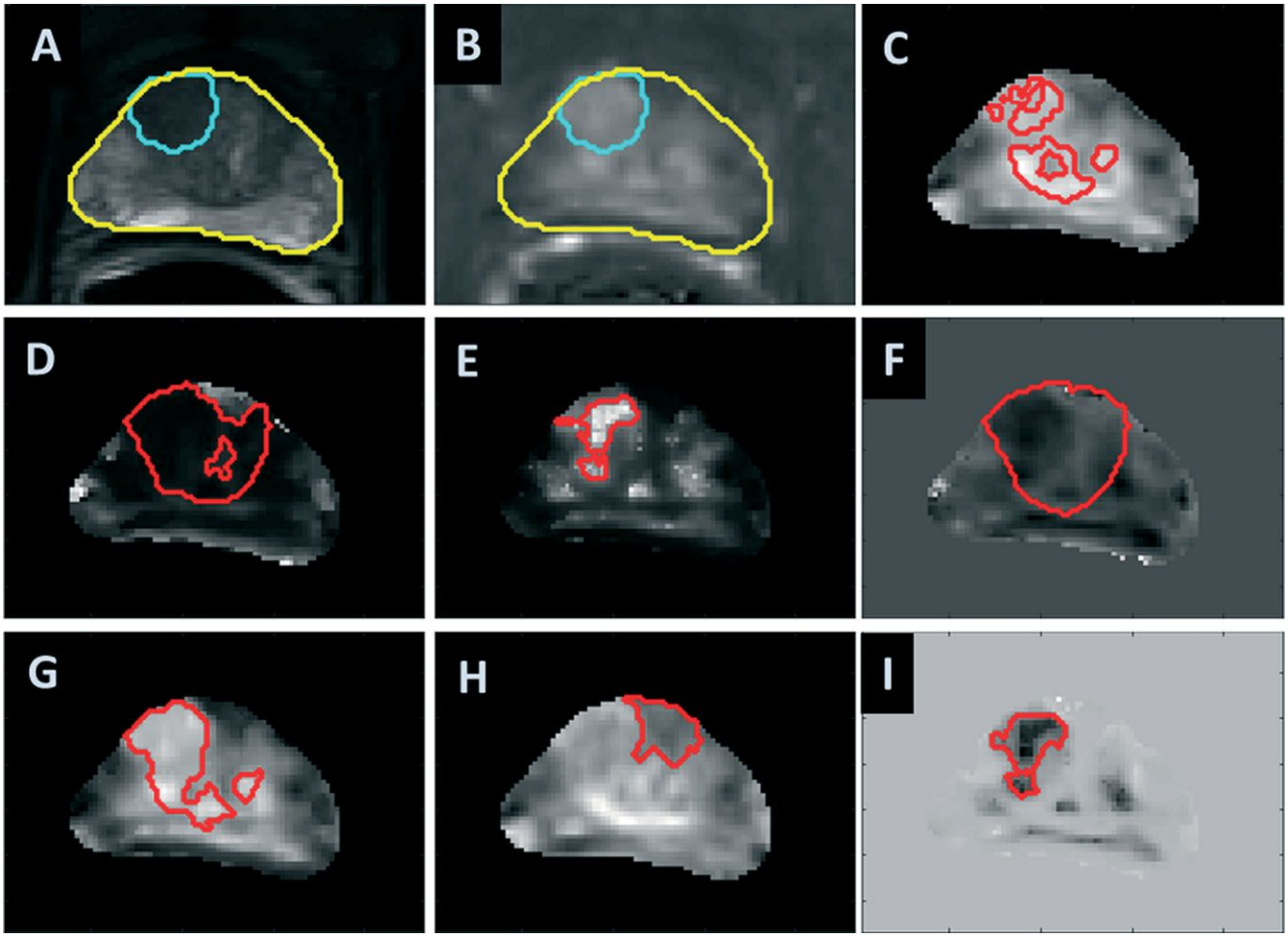


Figure 2. Example of prostate habitats based on DCE features. Radiologist's outline of an anterior lesion in the transition zone (TZ) (cyan) and prostate (yellow) contours overlapped with T2-weighted (T2w) imaging (A) and peak-enhancement DCE series (B). DCE feature maps for peak enhancement (C), time-to-peak (D), wash-in slope (E), wash-out slope (F), early AUC (G), late AUC (H), and slope product (I). DCE feature maps were built by parametrizing the DCE features for every voxel in the prostate.

brated so that both classes had matched sample size. Classifier performance was evaluated using *leave-1-out* cross-validation. Each classification experiment was repeated 50 times. Further, 95% confidence intervals for sensitivity, specificity, and AUC were estimated. Image processing and segmentations were performed on commercial imaging Picture Archive Communication System (PACS) workstation (MIM Corporation, Cleveland, OH, USA). Classifiers and feature computations were developed using custom code written in C++ and Matlab.

RESULTS

In this study we evaluated the predictive performance of *DCE (perfusion) habitats*, confined regions with similar perfusion behavior in the intra and peritumoral regions, using established characteristics of the DCE time activity curves. We determined a set of 7 parameters from a biexponential curve fitting of these curves (see online Supplemental Figure 1). These parameters generate feature maps (Figure 2) that were used to generate 1 habitat for each identified lesion. DICE score between habitats and manual lesion contours ranged between 0.17 and 0.42 (Table 2).

The discriminatory ability of each feature was evaluated performing univariate classification using SVM, repeated for each habitat (Table 3 and online Supplemental Table 1). The top performing habitat was the *slope product* habitat (*H-*m_{io}), with AUC for its DCE features in the range 0.46 – 0.78. The best predictive features were *tau* (AUC, 0.71 [0.69, 0.73]; sensitivity, 0.66 [0.64, 0.69]), followed by *wo* (AUC, 0.74 [0.73, 0.75]; sensitivity, 0.62 [0.60, 0.64]) and m_{io} (AUC, 0.78 [0.77, 0.79]; sensitivity, 0.68 [0.68, 0.68]).

Additionally, we separated the samples for consistent scanner types. For 1.5 T/ERC, the top performing habitats were the *late AUC* habitat (*H-AUCf*) and *H-*m_{io} (Table 4 and online Supplemental Table 2). For *H-AUCf*, the best predictive feature was *wi* (AUC, 0.81 [0.80, 0.82]; sensitivity, 0.74 [0.72, 0.75]), and for *H-*m_{io}, the best predictive feature was s_p (AUC, 0.78 [0.76, 0.80]; sensitivity, 0.79 [0.76, 0.81]). For 3 T, the top performing habitats were *H-AUCf* and the *peak enhancement* habitat (*H-*s_p)

Table 3. Univariate Evaluation of DCE-Based Habitats Versus DCE Features

	Feature						
	s_p	*tau*	*wi*	*wo*	*AUCi*	*AUCf*	m_{io}
Habitat							
H-s_p							
Sensitivity	0.54	0.76	0.60	0.68	0.43	0.56	0.64
Specificity	0.58	0.66	0.69	0.52	0.50	0.53	0.63
AUC	0.56	0.71	0.65	0.60	0.47	0.54	0.63
H-tau							
Sensitivity	0.44	0.55	0.71	0.59	0.49	0.71	0.71
Specificity	0.37	0.54	0.61	0.52	0.53	0.62	0.46
AUC	0.41	0.55	0.66	0.55	0.51	0.67	0.58
H-wi							
Sensitivity	0.48	0.40	0.54	0.51	0.58	0.58	0.44
Specificity	0.49	0.55	0.52	0.45	0.54	0.49	0.53
AUC	0.48	0.48	0.53	0.48	0.56	0.53	0.49
H-wo							
Sensitivity	0.55	0.71	0.60	0.65	0.59	0.56	0.58
Specificity	0.48	0.57	0.62	0.70	0.55	0.50	0.70
AUC	0.52	0.64	0.61	0.67	0.57	0.53	0.64
H-AUCi							
Sensitivity	0.57	0.49	0.47	0.59	0.55	0.55	0.55
Specificity	0.48	0.63	0.45	0.46	0.52	0.59	0.45
AUC	0.53	0.56	0.46	0.52	0.53	0.57	0.50
H-AUCf							
Sensitivity	0.55	0.68	0.66	0.63	0.68	0.57	0.71
Specificity	0.68	0.74	0.49	0.51	0.58	0.52	0.57
AUC	0.62	0.71	0.58	0.57	0.63	0.54	0.64
H-m_{io}							
Sensitivity	0.65	0.66	0.42	0.62	0.45	0.57	0.68
Specificity	0.57	0.75	0.69	0.86	0.46	0.41	0.88
AUC	0.61	0.71	0.56	0.74	0.46	0.49	0.78

Seven habitats were outlined by thresholding DCE feature maps (columns). For each habitat, the mean DCE feature values were computed (rows). Mean sensitivity, mean specificity, and mean AUC for classification between clinically insignificant and clinically significant cancer, based on MRI-guided biopsies. SVMs were used as classifiers with leave-1-out cross-validation. All patients in the study were included.

(Table 5 and online Supplemental Table 3). For the habitat based on the late area under the DCE time–activity curve (*H-AUCf*), the best predictive feature was *tau* (AUC, 0.83 [0.82, 0.85]; sensitivity, 0.69 [0.69, 0.70]), and for the *H-*s_p, the best predictive feature was *wo* (AUC, 0.81 [0.80, 0.83]; sensitivity, 0.85 [0.83, 0.86]).

The late AUC habitat (*H-AUCf*) was selected for pair-wise feature analysis because it had shown accurate features for both cohorts being robust for scanner strength/acquisition coil. Pair-wise analysis of this habitat showed that 2 pairs of features were predictive in both the 1.5/ERC data set and the 3 T data set. These pairs were (*tau*, *wi*) and (*wo*, *AUCi*) (Table 6 and online Supplemental Tables 4 and 5). Classification using the feature pair (*tau*, *wi*) had an AUC of 0.80 [0.79, 0.81] and a sensitivity 0.71 [0.70, 0.72] for 1.5 T and an AUC of 0.84 [0.83, 0.85] and a sensitivity 0.76 [0.75, 0.77] for 3 T. Classification using the feature pair (*wo*, *AUCi*) had an AUC of 0.82 [0.81, 0.83] and a sensitivity 0.80 [0.79, 0.81] for 1.5 T and an AUC of 0.81 [0.80, 0.82] and a sensitivity 0.73 [0.72, 0.75] for 3 T.

DISCUSSION

In our current work we present an approach to converge on a region (habitat) and quantify its DCE (perfusion) characteristics to discriminate clinically aggressive cancers. Prior work on perfusion characterization has shown DCE values extracted from ROIs correlates with pathological assessment (GS), using intra-subject nonlinear matrix factorization to identify a suspicious region (10). Owing to varied scanner types, using direct voxel intensity values, coupled with the nondeterministic nature of non-negative matrix factorization, limits the ability of the

Table 4. Univariate Evaluation of 1.5 T ERC DCE-Based Habitats Versus DCE Features

	Feature						
1.5 T ERC	**s_p**	**tau**	**wi**	**wo**	**AUCi**	**AUCf**	**m_{io}**
Habitat							
H-s_p							
Sensitivity	0.54	0.63	0.58	0.7	0.58	0.51	0.52
Specificity	0.45	0.7	0.67	0.53	0.59	0.5	0.42
AUC	0.49	0.66	0.63	0.62	0.59	0.5	0.47
H-tau							
Sensitivity	0.47	0.6	0.64	0.72	0.5	0.53	0.62
Specificity	0.45	0.53	0.68	0.53	0.55	0.52	0.49
AUC	0.46	0.57	0.66	0.62	0.53	0.52	0.56
H-wi							
Sensitivity	0.39	0.56	0.6	0.52	0.49	0.51	0.47
Specificity	0.53	0.59	0.43	0.63	0.54	0.51	0.44
AUC	0.46	0.57	0.52	0.57	0.52	0.51	0.46
H-wo							
Sensitivity	0.42	0.59	0.67	0.43	0.55	0.57	0.57
Specificity	0.65	0.51	0.6	0.49	0.58	0.42	0.52
AUC	0.54	0.55	0.64	0.46	0.57	0.49	0.54
H-AUCi							
Sensitivity	0.43	0.71	0.64	0.52	0.51	0.37	0.59
Specificity	0.55	0.69	0.52	0.64	0.43	0.45	0.55
AUC	0.49	0.7	0.58	0.58	0.47	0.41	0.57
H-AUCf							
Sensitivity	0.42	0.55	**0.74**	0.46	0.72	0.56	0.65
Specificity	0.42	0.74	**0.88**	0.48	0.76	0.64	0.66
AUC	0.42	0.64	**0.81**	0.47	0.74	0.6	0.66
H-m_{io}							
Sensitivity	**0.79**	0.65	0.61	0.66	0.72	0.72	0.63
Specificity	**0.78**	0.63	0.81	0.59	0.79	0.66	0.46
AUC	**0.78**	0.64	0.71	0.62	0.75	0.69	0.55

Seven habitats were outlined by thresholding DCE feature maps (columns). For each habitat, the mean DCE feature values were computed (rows). Mean sensitivity, mean specificity, and mean AUC for classification between clinically insignificant and clinically significant cancer, based on MRI-guided biopsies. SVMs were used as classifiers with leave-1-out cross-validation. All patients in the study were included. The two features with the largest AUC amongst all habitats have been indicated in boldface.

method to reproduce across varied cohorts. The habitat model presented here addresses the key issue of showing a means to localize the ROI before quantification. We use SVM classifiers to discern the habitat and quantified features on this habitat that improved the ability to discriminate aggressive cancers (22).

The use of parameters from pharmacokinetics modeling has shown to lack robustness. A recent study has shown usability of Ktrans map to localize the tumor region and these maps have been reported to be predictive of tumor aggressiveness (13). It has also been reported that repeatability of Ktrans maps across institutions has been low, and a recent report shows a coefficient of variation to be as high as 0.59 (23).

There is an open debate about the accuracy of ERC and pelvic phased-array coil for the detection of prostate cancer. At 1.5 T, ERC produces a higher-quality imaging of the prostate with common artifacts in the PZ. At 3 T, the pelvic phased-array coil produces high-quality images without the inconvenience and cost of an ERC. Because both of these technologies are currently used in the clinic, we strive to find DCE features that are robust to both acquisition coil and magnetic field strength of the scanner. It is imperative to develop prognostic features that work well with both types of coils. In this paper we review the robustness of DCE features in the prediction of clinically aggressive cancers, with respect to the acquisition settings.

To improve the accuracy and reproducibility of classifications, the patients were divided according to the MRI acquisition characteristics, and their habitats were analyzed separately, identifying DCE features that were common in both subsets. The *late AUC*

Table 5. Univariate Evaluation of 3 T Pelvic Coil DCE-Based Habitats Versus DCE Features

	Feature						
3 T PELVIC	s_p	*tau*	*wi*	*wo*	*AUCi*	*AUCf*	m_{io}
Habitat							
H-s_p							
Sensitivity	0.52	0.59	0.6	**0.85**	0.7	0.65	0.67
Specificity	0.62	0.8	0.89	**0.78**	0.54	0.71	0.79
AUC	0.57	0.7	0.74	**0.81**	0.62	0.68	0.73
H-tau							
Sensitivity	0.67	0.49	0.64	0.58	0.4	0.68	0.69
Specificity	0.69	0.66	0.6	0.56	0.61	0.79	0.58
AUC	0.68	0.57	0.62	0.57	0.51	0.73	0.63
H-wi							
Sensitivity	0.78	0.31	0.54	0.51	0.68	0.59	0.57
Specificity	0.68	0.64	0.57	0.46	0.57	0.69	0.57
AUC	0.73	0.48	0.55	0.48	0.63	0.64	0.57
H-wo							
Sensitivity	0.66	0.46	0.45	0.56	0.45	0.72	0.5
Specificity	0.56	0.36	0.47	0.67	0.48	0.52	0.61
AUC	0.61	0.41	0.46	0.61	0.46	0.62	0.56
H-AUCi							
Sensitivity	0.54	0.5	0.56	0.58	0.58	0.69	0.7
Specificity	0.63	0.6	0.55	0.67	0.62	0.8	0.6
AUC	0.59	0.55	0.55	0.63	0.6	0.75	0.65
H-AUCf							
Sensitivity	0.66	**0.69**	0.64	0.53	0.56	0.66	0.58
Specificity	0.62	**0.97**	0.68	0.87	0.65	0.64	0.89
AUC	0.64	**0.83**	0.66	0.7	0.6	0.65	0.73
H-m_{io}							
Sensitivity	0.68	0.59	0.51	0.47	0.55	0.48	0.52
Specificity	0.53	0.61	0.58	0.66	0.52	0.39	0.69
AUC	0.6	0.6	0.54	0.57	0.54	0.44	0.61

Seven habitats were outlined by thresholding DCE feature maps (columns). For each habitat, the mean DCE feature values were computed (rows). Mean sensitivity, mean specificity, and mean AUC for classification between clinically insignificant and clinically significant cancer, based on MRI-guided biopsies. SVMs were used as classifiers with leave-1-out cross-validation. All patients in the study were included. The two features with the largest AUC amongst all habitats have been indicated in boldface.

habitat (*H-AUCf*) showed good performance (with features having an AUC greater than 0.8) for both scanner settings. The *peak enhancement* habitat (*H-s_p*) in the 3 T data set had the largest sensitivity with the *wo* feature (AUC, 0.81 [0.80, 0.83]; sensitivity, 0.85 [0.83, 0.86]) but failed to be robust with the 1.5 T/ERC cohort (AUC, 0.62 [0.60, 0.64]; sensitivity, 0.70 [0.68, 0.72]).

Further pair-wise analysis of the *H-AUCf* habitat showed improvement for classification accuracy. For 3 T, 10 pairs of features showed AUC larger or equal to 0.8, while for 1.5 T/ERC, there were only 3 pairs. This may suggest that 3 T acquisition provides better predictive features on DCE images compared to 1.5 T with endorectal coil. The *H-AUCf* habitat had a DICE score of 0.22, suggesting that this habitat was mostly exploring the peritumoral region, adding information to the model from the surrounding environment. Two DCE feature pairs performed well: (*tau*, *wi*) and (*wo*, *AUCi*). The (*wo*, *AUCi*) pair had a sensitivity of 0.80 for 1.5 T/ERC and the (*tau*, *wi*) pair had a sensitivity of 0.76 for 3 T. Because we are aiming for features with high accuracy and high sensitivity, future experiments should evaluate if the tuple (*wo*, *AUCi*, *tau*, *wi*) would provide robust accuracy with high sensitivity. The main limitation of this study is the small sample size used for training; we expect using a conservative approach such as ours, would have a better chance of reproducibility.

CONCLUSION

We present a systematic quantitative methodology to identify DCE perfusion regions that provide quantitative assessment of DCE characteristics in these regions. We show that these metrics identify

Table 6. Evaluation of pairs of DCE features for habitat *H-AUCf*

	1.5 T ERC							3 T						
	s_p	*tau*	*wi*	*wo*	*AUCi*	*AUCf*	m_{io}	s_p	*tau*	*wi*	*wo*	*AUCi*	*AUCf*	m_{io}
Sensitivity														
s_p	0.42	0.74	0.64	0.75	0.70	0.60	0.64	0.66	0.76	0.61	0.70	0.74	0.65	0.74
tau		0.55	**0.71**	0.76	0.76	0.72	0.70		0.69	**0.76**	0.75	0.70	0.70	0.68
wi			0.74	0.63	0.73	0.68	0.60			0.64	0.63	0.61	0.70	0.73
wo				0.46	**0.80**	0.69	0.72				0.53	**0.73**	0.75	0.69
AUCi					0.72	0.73	0.69					0.56	0.77	0.64
AUCf						0.56	0.63						0.66	0.74
m_{io}							0.65							0.58
Specificity														
s_p	0.42	0.91	0.89	0.75	0.77	0.56	0.62	0.62	0.82	0.81	0.91	0.83	0.77	0.97
tau		0.74	**0.89**	0.72	0.81	0.86	0.70		0.98	**0.93**	0.98	0.84	0.83	0.94
wi			0.88	0.77	0.84	0.87	0.84			0.68	0.93	0.85	0.82	0.95
wo				0.48	**0.84**	0.66	0.67				0.87	**0.88**	0.91	0.94
AUCi					0.76	0.79	0.84					0.65	0.83	0.93
AUCf						0.64	0.63						0.64	0.91
m_{io}							0.66							0.89
AUC														
s_p	0.42	0.83	0.76	0.75	0.74	0.58	0.63	0.64	0.79	0.71	0.80	0.79	0.71	0.85
tau		0.64	**0.80**	0.74	0.79	0.79	0.70		0.83	**0.84**	0.87	0.77	0.76	0.81
wi			0.81	0.70	0.78	0.78	0.72			0.66	0.78	0.73	0.76	0.84
wo				0.47	**0.82**	0.67	0.69				0.70	**0.81**	0.83	0.82
AUCi					0.74	0.76	0.76					0.60	0.80	0.79
AUCf						0.60	0.63						0.65	0.83
m_{io}							0.66							0.73

Sensitivity, specificity, and AUC for classification between clinically insignificant and significant cancer is shown, based on MRI-guided biopsies. Support vector machines were used as classifiers. Leave-1-out cross-validation was used. The diagonal corresponds to the univariate case. The two features with the largest average AUC between 1.5 T and 3 T acquisitions have been indicated in boldface.

clinically significant cancers. In particular, we find that habitat regions identified by the late area under the DCE time–activity curve (*H-AUCf*) yield features to be related to clinically significant cancers. We also find that using a cohesive cohort with higher magnetic field strength (3 T) seems to improve the predictor performance.

Supplemental Materials

Supplemental Figure 1: http://dx.doi.org/10.18383/j.tom.2018.00037.sup.01

Supplemental Tables 1-5: http://dx.doi.org/10.18383/j.tom.2018.00037.sup.02

ACKNOWLEDGMENTS

This publication was supported by Grants R01CA189295 and R01CA190105 from the National Cancer Institute.

Disclosures: No disclosures to report.

Conflict of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA Cancer J Clin. 2016; 66:7–30.
2. Dianat SS, Carter HB, Macura KJ. Performance of multiparametric magnetic resonance imaging in the evaluation and management of clinically low-risk prostate cancer. Urol Oncol. 2014;32:39.e1–39.e10.
3. Lin K, Lipsitz R, Janakiraman S. Benefits and Harms of Prostate-Specific Antigen Screening for Prostate Cancer: An Evidence Update for the U.S. Preventive Services Task Force. Ann Intern Med. 2008;149:192–199.
4. Akizhanova M, Iskakova EE, Kim V, Wang X, Kogay R, Turebayeva A, Sun Q, Zheng T, Wu S, Miao L, Xie Y. PSA and Prostate Health Index based prostate cancer screening in a hereditary migration complicated population: implications in precision diagnosis. J Cancer. 2017;8:1223–1228.
5. Harvey CJ, Pilcher J, Richenberg J, Patel U, Frauscher F. Applications of transrectal ultrasound in prostate cancer. Br J Radiol. 2012;85 Spec Iss 1:S3–S17.

6. Tofts PS, Kermode AG. Measurement of the blood-brain barrier permeability and leakage space using dynamic MR imaging. 1. Fundamental concepts. Magn Reson Med. 1991;17:357–367.
7. Vajuvalli NN, Chikkemenahally DK, Nayak KN, Bhosale MG, Geethanath S. The Tofts model in frequency domain: fast and robust determination of pharmacokinetic maps for dynamic contrast enhancement MRI. Phys Med Biol. 2016;61:8462–8475.
8. Vos EK, Kobus T, Litjens GJ, Hambrock T, Hulsbergen-van de Kaa CA, Barentsz JO, Maas MC, Scheenen TW. Multiparametric magnetic resonance imaging for discriminating low-grade from high-grade prostate cancer. Invest Radiol. 2015;50:490–497.
9. Ginsburg SB, Algohary A, Pahwa S, Gulani V, Ponsky L, Aronen HJ, Boström PJ, Böhm M, Haynes AM, Brenner P, Delprado W, Thompson J, Pulbrock M, Taimen P, Villani R, Stricker P, Rastinehad AR, Jambor I, Madabhushi A. Radiomic features for prostate cancer detection on MRI differ between the transition and peripheral zones: Preliminary findings from a multi-institutional study. J Magn Reson Imaging. 2017;46:184–193.
10. Parra NA, Pollack A, Chinea FM, Abramowitz MC, Marples B, Munera F, Castillo R, Kryvenko ON, Punnen S, Stoyanova R. Automatic detection and quantitative DCE-MRI scoring of prostate cancer aggressiveness. Front Oncol. 2017;7:259.
11. American College of Radiology. Prostate Imaging Reporting and Data System (PIRADS) version 2. 2015. https://www.acr.org/~/media/ACR/Documents/PDF/QualitySafety/Resources/PIRADS/PIRADS%20V2.pdf.
12. Huisman HJ, Engelbrecht MR, Barentsz JO. Accurate estimation of pharmacokinetic contrast-enhanced dynamic MRI parameters of the prostate. J Magn Reson Imaging. 2001;13:607–614.
13. Vos EK, Litjens GJ, Kobus T, Hambrock T, Hulsbergen-van de Kaa CA, Barentsz JO, Huisman HJ, Scheenen TW. Assessment of prostate cancer aggressiveness using dynamic contrast-enhanced magnetic resonance imaging at 3 T. Eur Urol. 2013;64:448–455.
14. Litjens G, Debats O, Barentsz J, Karssemeijer N, Huisman H. Computer-aided detection of prostate cancer in MRI. IEEE Trans Med Imaging. 2014;33:1083–1092.
15. Chen YJ, Chu WC, Pu YS, Chueh SC, Shun CT, Tseng WY. Washout gradient in dynamic contrast-enhanced MRI is associated with tumor aggressiveness of prostate cancer. J Magn Reson Imaging. 2012;36:912–919.
16. Wu J, Cao G, Sun X, Lee J, Rubin DL, Napel S, Kurian AW, Daniel BL, Li R. Intratumoral spatial heterogeneity at perfusion MR imaging predicts recurrence-free survival in locally advanced breast cancer treated with neoadjuvant chemotherapy. Radiology. 2018;288:26–35.
17. Lee J, Narang S, Martinez J, Rao G, Rao A. Spatial habitat features derived from multiparametric magnetic resonance imaging data are associated with molecular subtype and 12-month survival status in glioblastoma multiforme. PLoS One. 2015;10:e0136557.
18. Stoyanova R, Chinea F, Kwon D, Reis IM, Tschudi Y, Parra NA, Breto AL, Padgett KR, Pra AD, Abramowitz MC, Kryvenko ON, Punnen S, Pollack A. An automated multiparametric MRI quantitative imaging prostate habitat risk scoring system for defining external beam radiotherapy boost volumes. Int J Radiat Oncol Biol Phys. 2018;102:821–829.
19. Farjam R, Tsien CI, Lawrence TS, Cao Y. DCE-MRI defined subvolumes of a brain metastatic lesion by principle component analysis and fuzzy-c-means clustering for response assessment of radiation therapy. Med Phys. 2014;41:011708.
20. Haase A. Snapshot FLASH MRI. Applications to T1, T2, and chemical-shift imaging. Magn Reson Med. 1990;13:77–89.
21. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–357.
22. Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods. Cambridge: Cambridge University Press. 2000.
23. Huang W, Li X, Chen Y, Li X, Chang MC, Oborski MJ, Malyarenko DI, Muzi M, Jajamovich GH, Fedorov A, Tudorica A, Gupta SN, Laymon CM, Marro KI, Dyvorne HA, Miller JV, Barbodiak DP, Chenevert TL, Yankeelov TE, Mountz JM, Kinahan PE, Kikinis R, Taouli B, Fennessy F, Kalpathy-Cramer J. Variations of dynamic contrast-enhanced magnetic resonance imaging in evaluation of breast cancer therapy response: a multicenter data analysis challenge. Transl Oncol. 2014;7:153–166.

Phantom Validation of DCE-MRI Magnitude and Phase-Based Vascular Input Function Measurements

Warren Foltz[1,2], Brandon Driscoll[1], Sangjune Laurence Lee[2], Krishna Nayak[3], Naren Nallapareddy[3], Ali Fatemi[2], Cynthia Ménard[4,6], Catherine Coolens[1,2,4,6], and Caroline Chung[5,7]

[1]Department of Medical Physics, Princess Margaret Cancer Center and University Health Network, Toronto, ON, Canada; [2]Department of Radiation Oncology, University of Toronto, Toronto, ON, Canada; [3]Ming Hsieh Department of Electrical Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles, CA; [4]Department of Radiation Oncology, Centre Hospitalier Universite de Montreal, Montreal, Canada; [5]TECHNA Institute, University Health Network, Toronto, ON, Canada; [6]Department of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, ON, Canada; and [7]Department of Radiation Oncology, MD Anderson Cancer Center, Houston, TX

Corresponding Author:
Catherine Coolens, PhD
Princess Margaret Cancer Centre, Radiation Medicine Program,
Rm 6:306 - 700 University Avenue, Toronto, ON M5G 1Z5, Canada;
E-mail: catherine.coolens@rmp.uhn.ca.

Key Words: dynamic contrast-enhanced MRI (DCE-MRI), permeability, arterial input function (AIF), quantification, MRI phase, phantom

Abbreviations: Arterial input functions (AIF), magnetic resonance imaging (MRI), computed tomography (CT), dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI), field of view (FOV), magnetic resonance (MR), echo time (TE), repetition time (TR), regions of interest (ROIs), signal-to-noise ratio (SNR)

ABSTRACT

Accurate, patient-specific measurement of arterial input functions (AIF) may improve model-based analysis of vascular permeability. This study investigated factors affecting AIF measurements from magnetic resonance imaging (MRI) magnitude (AIF_{MAGN}) and phase (AIF_{PHA}) signals, and compared them against computed tomography (CT) (AIF_{CT}), under controlled conditions relevant to clinical protocols using a multimodality flow phantom. The flow phantom was applied at flip angles of 20° and 30°, flow rates (3–7.5 mL/s), and peak bolus concentrations (0.5–10 mM), for in-plane and through-plane flow. Spatial 3D-FLASH signal and variable flip angle T1 profiles were measured to investigate in-flow and radiofrequency-related biases, and magnitude- and phase-derived Gd-DTPA concentrations were compared. MRI AIF performance was tested against AIF_{CT} via Pearson correlation analysis. AIF_{MAGN} was sensitive to imaging orientation, spatial location, flip angle, and flow rate, and it grossly underestimated AIF_{CT} peak concentrations. Conversion to Gd-DTPA concentration using T1 taken at the same orientation and flow rate as the dynamic contrast-enhanced acquisition improved AIF_{MAGN} accuracy; yet, AIF_{MAGN} metrics remained variable and significantly reduced from AIF_{CT} at concentrations above 2.5 mM. AIF_{PHA} performed equivalently within 1 mM to AIF_{CT} across all tested conditions. AIF_{PHA}, but not AIF_{MAGN}, reported equivalent measurements to AIF_{CT} across the range of tested conditions. AIF_{PHA} showed superior robustness.

INTRODUCTION

Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) is a useful tool to measure blood vessel permeability and volume fractions within heterogeneous lesions, such as tumors (1). There is growing interest in the role of early changes in tumor vascularity as predictive biomarkers of tumor response to therapy, particularly with increasing use of antiangiogenic agents, recognizing that changes in tumor physiology can often precede tumor volume changes (1-3). Biomarkers of early response to treatment introduce the potential to individualize cancer treatment based on individual responses, but the current challenge is determining the optimal approach for acquiring and interpreting these biomarker measures. To date there has been wide variability in the reported DCE-MRI findings and responses across different institutions and this may, at least in part, reflect the variability in image acquisition and analysis (4, 5).

Analysis of DCE-MRI data commonly assumes a 2-compartmental model to generate functional parameters, such as the permeability surface area product per unit volume (K^{trans}), size of the extracellular extravascular space (v_e), and efflux rate constant (k_{ep}) (6, 7). Accurate quantification of these permeability kinetic parameters is dependent on the application of an accurately measured arterial input function (AIF) from a major vessel in the vicinity of the tumor (8). Typically, the AIF has been evaluated by using the magnetization magnitude signal in an artery, but the conversion from magnitude signal to absolute gadolinium contrast agent concentration (eg, Gd-DTPA) is susceptible to a number of factors including blood inflow effects, radiofrequency transmit field (B_1) inhomogeneity, slice profile

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2019.00001

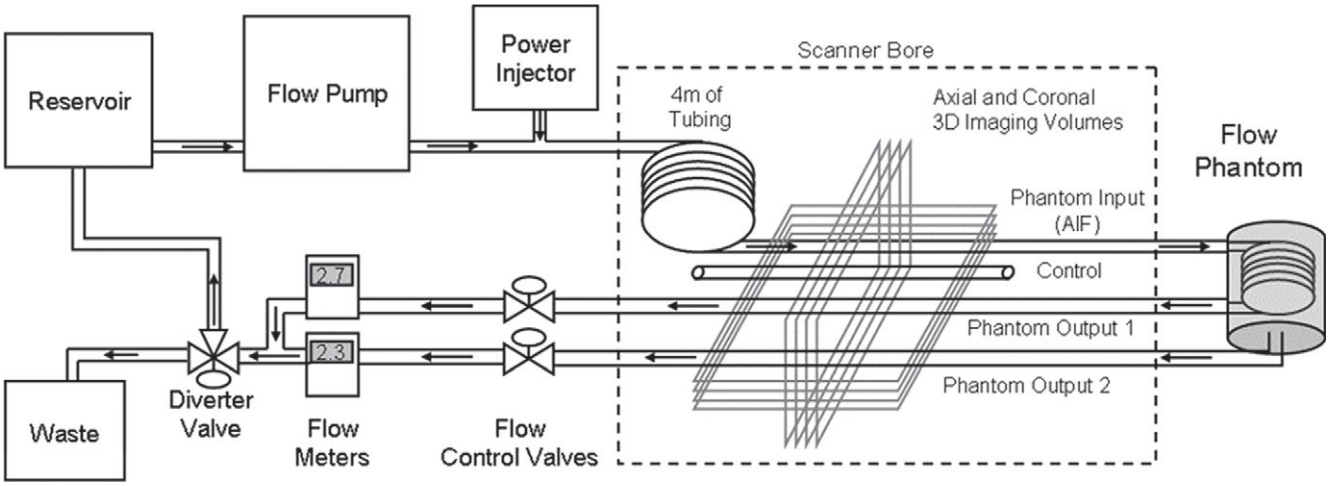


Figure 1. Simplified layout of the flow-phantom experiment. A high-concentration bolus is delivered from the pump and power injector through the phantom input tube into the flow phantom, where it divides into phantom output tubes 1 and 2, so that arterial input functions (AIFs) corresponding to each tube are captured within the same dynamic acquisition. The output flow ratios of the 2 phantom output tubes were set to 50:50 so that the velocity in each of the output tubes is half of that in the input tube. The diverter valve can channel the returning fluid to the reservoir for recirculation or to the waste for disposal (22).

effects, mis-registration susceptibility shifts, contrast agent dispersion, and hematocrit variation (9-12). Owing to these challenges, many studies have used the population-average AIF provided by vendor software for the generation of kinetic parameters from DCE-MRI data. Use of a population-average AIF may improve reproducibility in permeability kinetic parameters, but it may not result in accurate and meaningful quantification of kinetic parameters for individual patients (13, 14). Accurate measurement of an individual AIF may help improve both the accuracy and reproducibility of kinetic analysis (15, 16).

A growing body of work supports the use of the MRI signal phase for AIF measurement, for improved robustness relative to the magnitude signal (17-20). This study used an in-house-developed dynamic flow phantom (21) to investigate factors affecting the magnitude signal-derived AIF (AIF_{MAGN}), and to compare AIF_{MAGN} to the phase signal-derived AIF (AIF_{PHASE}) in a controlled environment with validation against the gold-standard computed tomography (CT)-derived AIF (AIF_{CT}). Both accuracy and robustness of the respective input functions were tested against varying imaging orientation, flip angles, flow rates, and peak AIF gadolinium contrast agent concentrations.

MATERIALS AND METHODS

Multimodal CT/MRI Flow Phantom

The basis of experimentation was an in-house-developed flow phantom (Figure 1), currently in use for accreditation of centers participating in multicenter clinical trials using DCE-CT in the province of Ontario (21). Physiological flow was simulated by a positive displacement pump (Compuflow 1000MR, Shelley Medical Imaging Technologies, London, ON), which pushed a blood-mimicking fluid consisting of a 15%–85% glycerol–water by volume mixture through the flow circuit. The 15%–85% glycerol–water mixture was pumped from an external reservoir through 1/4″ (6.35 mm) polyvinyl chloride tubing, and an in-line clinical power injector (Optistar Elite, Mallinckrodt, Cincinnati, OH) was used to simulate the contrast bolus representing the AIF (Phantom Input) by injecting various dilutions of Gadovist 1.0 (604 mg/ml, Bayer Corp., Leverkusen, DE). The flow phantom, based on a 2-compartmental exchange model, has 2 output tubes roughly representing the venous output function (phantom output 1) and the tissue signal function (phantom output 2). Fluid from the phantom outputs was fed back to the external reservoir for noncontrast experiments or to a waste container for contrast experiments.

Within flow rates up to 7.5 mL/s, the flow phantom provides high intrarun and intraday reproducibility with an error of $<2\%$ as validated through CT imaging.

The flow through the phantom output tubes is controlled by a set of flow control valves such that the output flow rates in each output tube is equal to half that of the input tube. The relationship between the phantom input peak concentration and that of the output tube peaks is variable, based on the choice of flow rate because exchange happens more quickly under higher rates of flow, but it is fully predictable based on the known geometry of the system.

Gold Standard CT_{AIF}

Gold standard CT_{AIF} measurements, shown in Figure 2, were acquired with a 320-slice scanner (Toshiba Medical Systems, Aquilion ONE) using a dynamic volume–time sequence operating at 120 kV, 300 mA gantry rotation of 0.5 seconds, and image frequency of 1 vol (160 mm longitudinal coverage) every 1.5

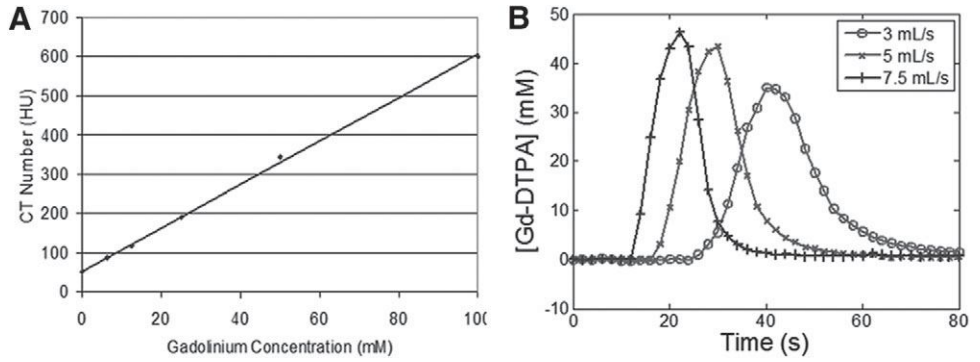


Figure 2. Gold standard AIF_{CT}: linearity of Hounsfield unit calibration with Gd-DTPA (A), Gold standard AIF_{CT} for 10 mM bolus injection at 3, 5, and 7.5 mL/s flow velocity (B).

seconds, with spatial resolution of 0.625 × 0.625 × 1 mm^3. The DCE-CT studies were performed at flow rates of 3, 5, and 7.5 mL/s, corresponding to average flow velocities of 9.5, 15.8, and 23.7 cm/s, at peak AIF Gd-DTPA concentration of 50 mM to improve CT signal-to-noise ratio (SNR). The peak CT-measured Gd-DTPA concentrations are linearly scaled to match those of corresponding MRI experiments, based on *a priori* validation of linearity between the CT Hounsfield Units and Gd-DTPA concentration (Figure 2).

MRI Methods

A consistent setup was used for both DCE-CT and DCE-MRI acquisitions. The in-flow and out-flow tubes were oriented parallel to B_0 to minimize susceptibility artifacts (22), and these tubes were placed above the spine array coil and below the 2-coil body array to allow for sensitive experimentation. A separate polyvinyl chloride tube filled with the 15%–85% glycerol–water mixture was placed within the imaging stack to provide a signal reference for MRI analysis. An additional 4 m of coil tubing was also wound within the MRI bore to allow for polarization of in-flowing spins.

All magnetic resonance (MR) imaging used a 3 T Verio System (IMRIS, Winnipeg, CA). Variable-flip-angle (VFA) T1 quantification and DCE experimentation used a 3-dimensional Fast Low Angle SHot (3D-FLASH) pulse sequences with shared geometric features (23). For axially oriented slice packages (eg, through-plane flow), 3D data sets were acquired over a 12.8- × 6.4- × 12-cm field of view (FOV) with 124 × 64 × 24 matrix, providing 1 × 1 × 5 mm voxels. For coronally oriented slice packages (eg, in-plane flow), the FOV was 19.2 × 9.6 × 7.2 cm, matrix size was 192 × 96 × 24, and voxels were 1 × 1 × 3 mm^3. All acquisitions used an echo time (TE) and repetition time (TR) of 1.86 milliseconds and 4.8 milliseconds, and a 500 Hz/pixel readout bandwidth. For dynamic scans, 3D-FLASH temporal resolution was 5 seconds with 36 repetitions including at least 3 repetitions at baseline flow before contrast agent injection to determine the average preinjection signal. The acquisition times were 37 seconds per flip angle for VFA-T1 (4 flip angles of 2°, 10°, 20°, 30°, 5 averages, iPAT factor 1), and 3 minutes 4 seconds for DCE-MRI (experiment-dependent flip angle, iPAT factor 2, 5-second temporal resolution, 38 repetitions).

Static Experiments

Gd-DTPA was diluted into 15%/85% glycerol/water and water-only at concentrations between 0 and 10 mM within 15-cc conical tubes. Within the 8-channel head coil of the 3 T Verio system, shimming was then performed using the 0 mM tube centrally placed, and surrounded by 6 control tubes containing water. 3D-FLASH acquisitions, including magnitude and phase reconstructions, were then performed. The central tube was then replaced with a tube of higher Gd-DTPA concentration and imaged without reshimming. This design reduced biases from coil sensitivity, and from shimming to an asymmetric distribution of samples with varying magnetic susceptiblity. It also provided a background phase correction, measured as the average phase drift across all 6 control tubes.

T1 relaxivity was measured using the body coil for RF transmit, and spine array coil elements and anteriorly placed small flexible coil for RF receive. T1 values were measured from all samples at once using an inversion recovery spin-echo technique (slice-selective inversion pulse; TE 12 milliseconds; TR 9350 milliseconds; inversion times 25, 50, 100, 200, 400, 800, 1200, 1600, 2000, 3500, 5000 milliseconds; FOV 240 × 192 mm; matrix 138 × 102; 5 mm slice thickness; iPAT factor 2; readout bandwidth 130 Hz/pixel; 10 minutes per inversion time). T1 relaxivity was extracted from the T1 and concentration data pairs via linear regression (OriginLab, Northampton, MA).

Dynamic Experiments

The complete set of dynamic experiments is summarized in Table 1. The input flow rate varied between 3, 5, and 7.5 mL/s (average flow velocities of 9.5, 15.8, and 23.7 cm/s), consistent with the physiologic range of internal carotid artery blood flow rates (24). For most runs, a peak concentration of 10 mM was delivered through the in-flow tube, which also provided assessment of peak concentrations of ~5 and 2.5 mM in the 2 outflow tubes. Lower peak AIF Gd-DTPA concentrations of ~0.5, 1, and 2 mM in the in-flow tube were also considered. This concentration range from 0.5 to 10 mM provided coverage of the full range of Gd-DTPA concentrations expected in a clinical DCE-MRI examination (25). Gd-DTPA concentration at peak enhancement was programmed by varying the dilution of Gd-DTPA within the power injector at constant injection volume of 16 mL and duration of 10 seconds.

Magnitude-Derived AIF: In-flow, RF, and Slice Profile Effects. Inhomogeneity of the RF transmit field and inflow affect the spatial profile and accuracy of the 3D-FLASH magnitude signal (23). Furthermore, inflow and RF transmit field inhomogeneity prolong the transition of the 3D-FLASH signal to steady-state

Table 1. List of AIF Experiments

Run #	Flow Rate (ml/s)	Peak Concentration (mM)	Imaging Plane	Flip Angle (Degrees)
1	7.5	10	Through-plane	20
2	7.5	10	Through-plane	20
3	5	10	Through-plane	20
4	5	10	Through-plane	20
5	5	10	Through-plane	20
6	5	10	Through-plane	20
7	3	10	Through-plane	20
8	3	10	Through-plane	20
9	5	5	Through-plane	20
10	5	5	Through-plane	20
11	5	2	Through-plane	20
12	5	2	Through-plane	20
13	5	1	Through-plane	20
14	5	1	Through-plane	20
15	5	0.5	Through-plane	20
16	5	0.5	Through-plane	20
17	7.5	10	Through-plane	30
18	5	10	Through-plane	30
19	3	10	Through-plane	30
20	5	5	Through-plane	30
21	5	2	Through-plane	30
22	5	1	Through-plane	30
23	5	0.5	Through-plane	20
24	7.5	10	In-plane	30
25	5	10	In-plane	20
26	7.5	10	In-plane	30

Imaging plane is stated as relative to the direction of flow; through-plane corresponds to an axial slice package; and in-plane corresponds to a coronal slice package. Flow rates of 3, 5, and 7.5 mL/s correspond to flow velocities of 9.5, 15.8, and 23.7 cm/s, respectively.

(9). Different RF inhomogeneity, slice profile, in-flow, and steady-state errors are to be expected for through-plane flow with an axial slice excitation and for in-plane flow with a sagittal or coronal slice excitation.

Spatial profiles of the 3D-FLASH signal were measured via acquisitions at flow rates from 0 to 23.7 cm/s and at flip angles from 2° to 30° for both axially and coronally oriented slice packages. Effects of flow rate and flip angle on the 3D-FLASH signal profile were confirmed via magnitude AIF measurements (through-plane and in-plane dynamic acquisitions at flip angles of 20° and 30° at flow rates of 5 mL/s and 7.5 mL/s and peak concentration of 10 mM. The magnitude AIF signals was scaled to Gd-DTPA concentration using T1 measured in a central slice using the variable flip angle technique.

Magnitude-Derived AIF: Correction Using VFA-T1. Endogenous T1 scales the conversion between the magnitude signal and Gd-DTPA concentration. However, in-flow accelerates the measured T1 relaxation based on the extent of spin displacement (26). Commonly used VFA-T1 measurements are very prone to bias from RF inhomogeneity, RF mistuning, and slice profile (23). Therefore, endogenous T1 values for blood taken from the literature or measured from static volumes may not be representative of true rates of repolarization at any location within the 3D-FLASH slice package in vivo. Geometrically equivalent 3D-FLASH VFA-T1 and DCE acquisitions at matched TR should be affected similarly by in-flow and RF errors. If so, MR signal to concentration conversion using position- and velocity-matched VFA-T1 instead of assumed T1 may improve the AIF_{MAGN} measurement.

The following acquisitions tested for improved AIF_{MAGN} using position- and velocity-matched VFA-T1. First, through-plane and in-plane 3D-FLASH image sets were measured at 2°, 10°, 20°, and 30° under static conditions. These image sets confirmed VFA-T1 at each location along the slice profile of the 3D-FLASH RF excitation pulse. Second, VFA-T1 maps were reconstructed from equivalent 3D-FLASH image sets acquired at flow rates of 3, 5, and 7.5 mL/s, to validate T1 acceleration with in-flow (26). Third, dynamic acquisitions at matching flow rates and flip angles of 20° and 30° were acquired during bolus Gd-DTPA injection with peak concentration of 10 mM and compared against DCE-CT to verify improved AIF accuracy when position- and velocity-matched VFA-T1 values were used rather than the VFA-T1 value at the center of the RF slice profile.

Phase-Derived AIF: Velocity and Concentration Effects. Compared to the magnitude-derived AIF, the phase-derived AIF should be insensitive to slice profile and in-flow effects. Phase-

and CT-derived AIFs for both axially and coronally oriented slice packages were compared at variable peak Gd-DTPA concentrations (0.5, 1, 2, 5, and 10 mM) and flow rates (9.5, 15.8, and 23.7 cm/s), across the entire slice profile.

Phase- and Magnitude-Derived AIFs: Comparison to CT. MRI AIF performance was tested against the CT gold standard via Bland–Altman difference and Pearson correlation analyses for all through-plane acquisitions in Table 1 (sample size of 69 given 23 runs, with readings from 1 phantom input and 2 phantom output tubes for each run). Further, 95% limits of agreement between the techniques for measurements in a single central slice were reported as the mean difference ± 1.96 standard deviance of the difference for the peak AIF concentration and the area under the curve (AUC) for the first 120 seconds after injection.

Phase- and Magnitude-Derived AIFs: Spatial Heterogeneity. Given the insensitivity to RF and in-flow effects, AIF_{PHA} should prove robust away from the central imaging slice. Spatial AIF_{MAGN} and AIF_{PHA} profiles were generated for 10 mM Gd-DTPA bolus injection at 23.7 cm/s flow velocity (5 mL/s). T1 measurements were matched to both velocity and slice position. Mean and standard deviations of T1-corrected AIF_{MAGN} and AIF_{PHA} were reported across the slice package for through-plane flow and across the field-of-view for in-plane flow.

Image Analysis

All MRI and CT signal and image data processing and analysis used Matlab® (MathWorks, Natick, MA). MR signal modeling used standard equations for magnitude and phase signal conversion to Gd-DTPA concentration, as follows:

$$S = S_0 \sin(\alpha)(1 - E_1)/(1 - E_1\cos(\alpha))e^{-TE/T2^*} \quad (1)$$

where $E_1 = \exp(-TR/T1)$, α is the flip angle, and S_0 and S are the relative signal enhancements before contrast injection and after contrast injection, respectively (27).

Magnitude signal enhancement was converted to concentration according to Schabel and Parker (28) with the following equation:

$$1/T_1(C) = 1/T_{10} + r_1 C \quad (2)$$

where T_{10} and T1 are the spin-lattice relaxation times before and after contrast injection, respectively; r_1 is the relaxivity of the contrast agent in the 15%–85% glycerol–water mixture; and C is the concentration of the Gd-DTPA contrast agent (27). For dynamic image analysis, the average of signals at the first 3 time points provided an estimate of the signal baseline.

The change in signal phase was converted to Gd-DTPA concentration with the following equation:

$$\Delta\varphi = TE\,\pi\gamma B_0 \chi_m \Delta C(\cos^2\theta - 1/3) \quad (3)$$

where γ is the proton gyromagnetic ratio (4.258×10^7 Hz/T), B_0 is the magnitude of the main magnetic field in Tesla, χ_m is the molar susceptibility of the Gd-DTPA concentration (3.4×10^7 mM^{-1} for Gd, in MKS units), and θ is the angle of the vessel relative to the main magnetic field ($\theta = 0$ being parallel with that field) (19). Concentration profiles of AIF_{PHA} were compensated for background phase drifts by subtraction of the phase signal within the 15%–85% glycerol–water control tube (18). The background phase drifts between baseline and final dynamic frames corresponded to Gd-DTPA concentration changes of 0.6 ± 0.2 mM, averaged across all 23 through-plane acquisitions and 2 control tubes.

Mean and standard deviations of signals were extracted from regions of interest (ROIs) for each of the 3 flow tubes. Analysis of axially oriented images used circular ROIs drawn on the AIF and control tubes in each of the 24 axial reconstructed slices. Coronal image analysis used 12 ROIs drawn equally spaced along the z-direction with both AIF and control tubes on a single coronal section that bisected each tube. Phase images were manually unwrapped if the ROI contained a phase 360° to 0° discontinuity by shifting modulo 360° until the discontinuity disappeared.

Statistical Analysis

Pearson correlations and linear regressions of peak Gd-DTPA concentrations and AUC measurements between CT and different MR data sets (magnitude, magnitude T1-corrected, and phase) were performed in MATLAB (The MathWorks) for both peak and AUC.

RESULTS

Static Experiments

Using the inversion recovery technique, an endogenous VFA-T1 of 1935 ± 40 milliseconds and T1 relaxivity of 7.5 ± 0.1 1/mM* milliseconds (R = 0.9998) were measured (Figure 3, A and B). The corresponding values for the Gd-DPTA–water solutions were 3007 ± 76 milliseconds and 5.0 ± 0.1 1/mM* milliseconds. Gd-DTPA concentrations derived from the magnitude signal were badly truncated to 2.5 mM using a 10° flip angle, but the Gd-DTPA concentration to 5 mM using a 20° flip angle, and to 10 mM using a 30° flip angle (Figure 3C). Linear and accurate measurement of Gd-DTPA concentration from the phase signal was observed after background phase correction (Figure 3D).

Dynamic Experiments

Magnitude-Derived AIF: In-flow, RF, and Slice Profile Effects. Spatial 3D-FLASH signal profiles are presented in Figure 4. In this figure, a completed transition to steady state was visualized as an equalization of signal magnitude with the static (0 cm/s) case. Flow data acquired with a 2° flip angle did not deviate much from the static experiment. At a 10° flip angle, considerable inflow bias was observed across the entire slice package at flow velocities above 3.2 cm/s for through-plane flow data (Figure 4C), whereas in-plane flow data were effectively acquired in steady-state at 0.6 relative to the FOV for flow velocities up to 30 cm/s (9.5 mL/s) (Figure 4D). With increasing flip angle above 10°, the extent of in-flow bias was reduced. Correspondingly, Figure 5 confirms improved, yet underestimated, magnitude-derived AIF estimation using higher flip angle acquisitions (20° and 30° for through-plane flow; 20° for in-plane flow at flow rates of 5 and 7.5 mL/s).

Magnitude-Derived AIF: Correction Using VFA-T1. Figure 6 confirms that in-flow and RF-related biases on the 3D-FLASH magnitude signal are encoded in VFA-T1 for through-plane measurements. Under no-flow conditions, axial VFA-T1 was uniform within 10% across 40% of the 12-cm slice package, and reduced sharply towards zero outside of the plateau region of the RF pulse profile. In comparison, coronal VFA-T1 was uniform

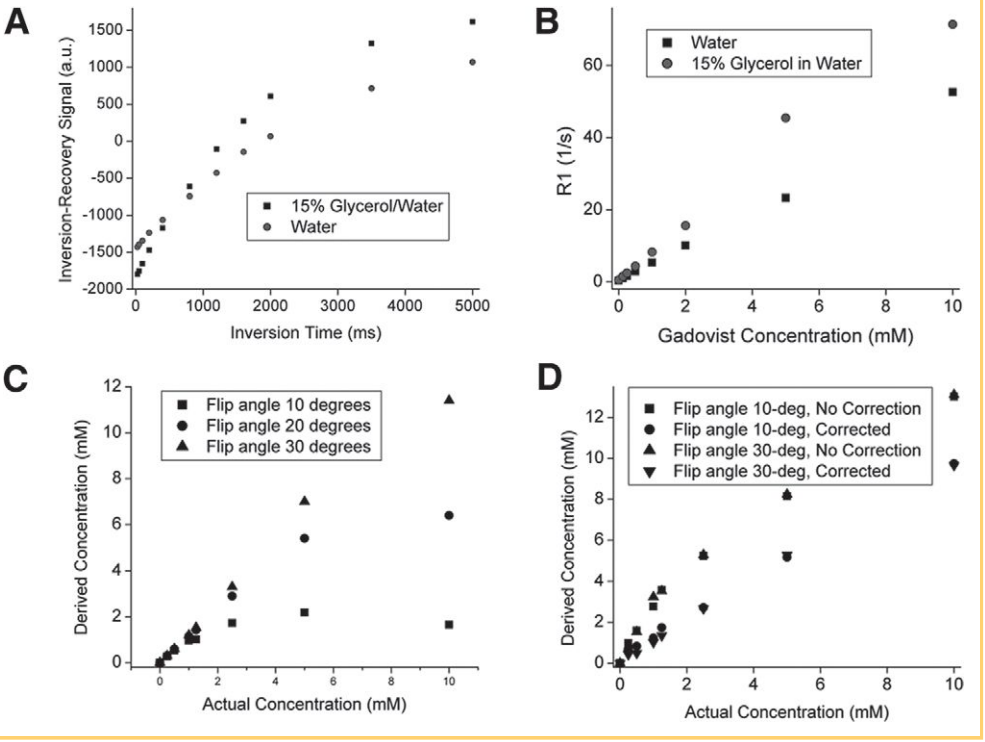


Figure 3. Static experiment results: Inversion-recovery signal as a function of inversion time for pure water and the 15% glycerol/water mixture (A). T1 and concentration data pairs used for T1 relaxivity calculations (B). Magnitude-derived concentration estimates against truth for flip angles of 10°, 20°, and 30° (C). Phase-derived concentration estimates against truth for flip angles of 10° and 30°, with and without background phase correction (D).

within 10% over 60% of the 19.2-cm FOV, and then reduced gradually, reflecting the RF inhomogeneity of the body transmit coil. Through-plane flow shifted the VFA-T1 profile in the direction of flow at increasing velocities, yet in-plane flow introduced only minor deviations to the VFA-T1 profile. Figure 7 confirms considerable improved, yet underestimated, AIF_{MAGN} accuracy compared to AIF_{CT} by using velocity-matched VFA-T1 measurements for signal-to-concentration conversion.

Phase-Derived AIF: Velocity and Concentration Effects. Figure 8A compares AIF_{PHA} and AIF_{CT} within a central slice across varying Gd-DTPA concentrations (0.5 to 10 mM) for through-plane flow at a fixed input flow rate of 5 mL/s. After

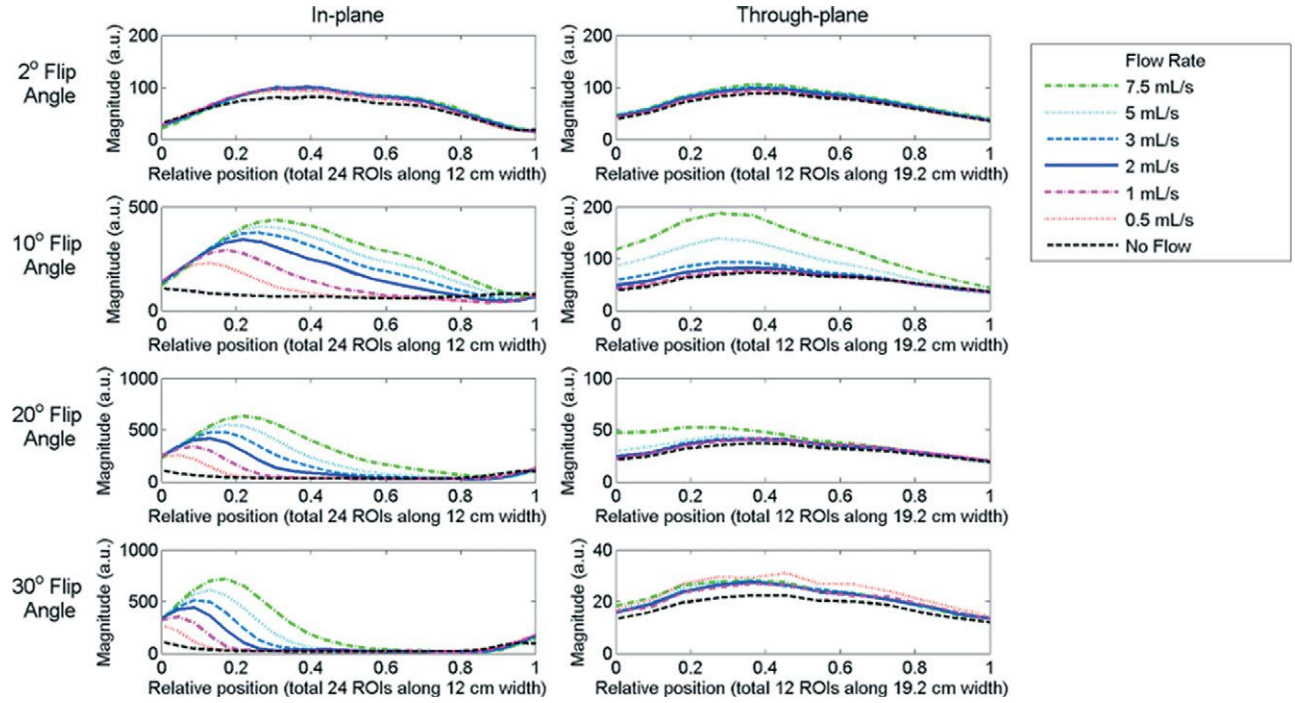


Figure 4. 3D FLASH signal profiles at varying flip angles and flow rates, corresponding to the range of velocities in the phantom input and 2 output tubes without Gd-DTPA for through-plane and in-plane orientations.

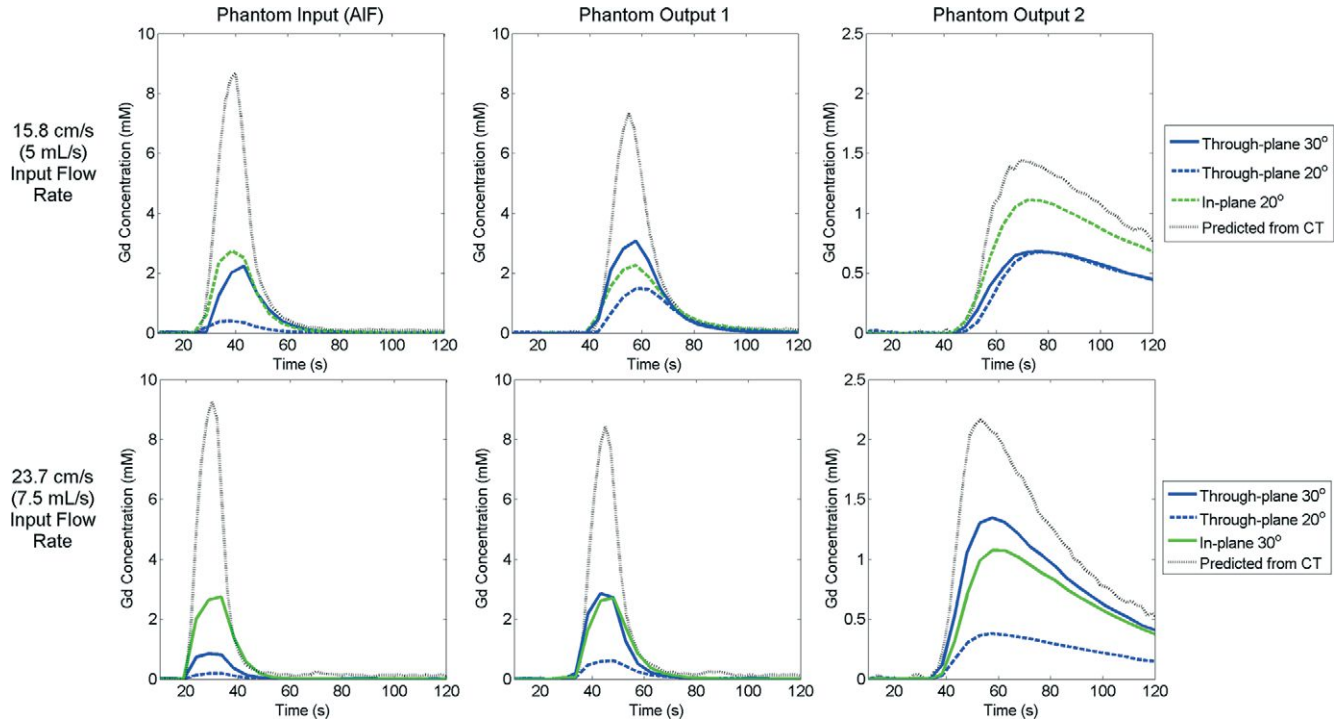


Figure 5. Magnitude and computed tomography (CT)-derived AIFs acquired at 10 mM peak Gd-DTPA concentration acquired for in-plane and through-plane flow orientations at multiple flip angles and 2 flow velocities.

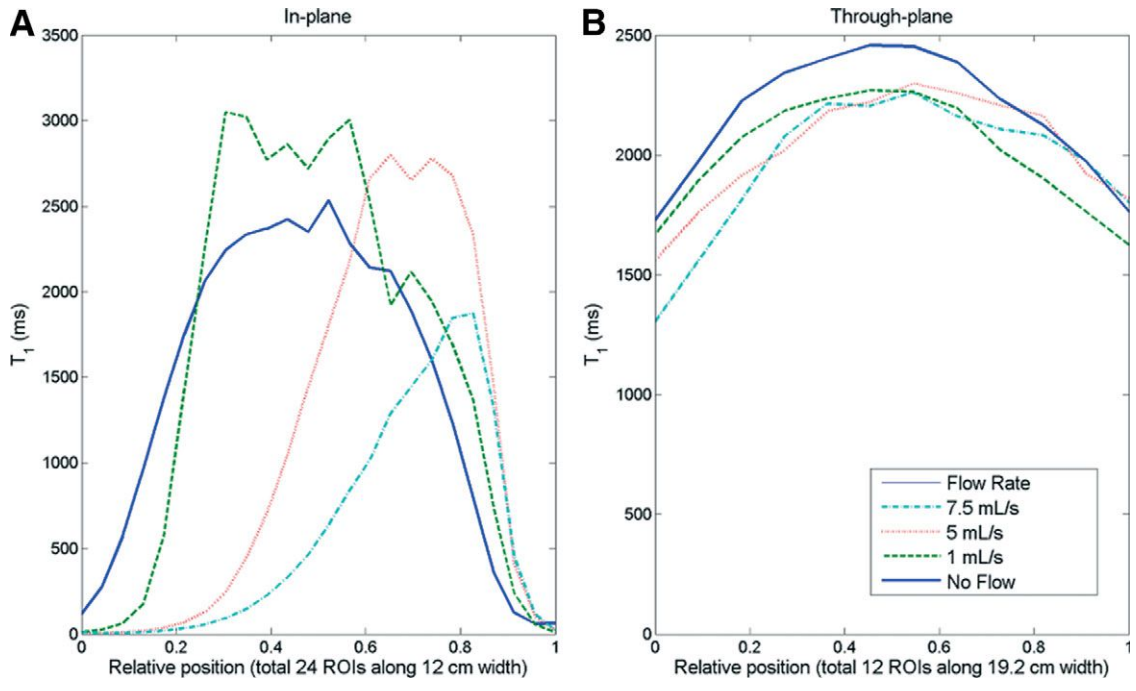


Figure 6. VFA-T1 measured at flow rates ranging from 0 to 7.5 mL/s. VFA-T1 accelerated by through-plane flow across the 12-cm slice profile in the axial orientation (A). VFA-T1 accelerated by in-plane flow across the 19.2 cm field of view (FOV) at zero flow in the coronal orientation (B).

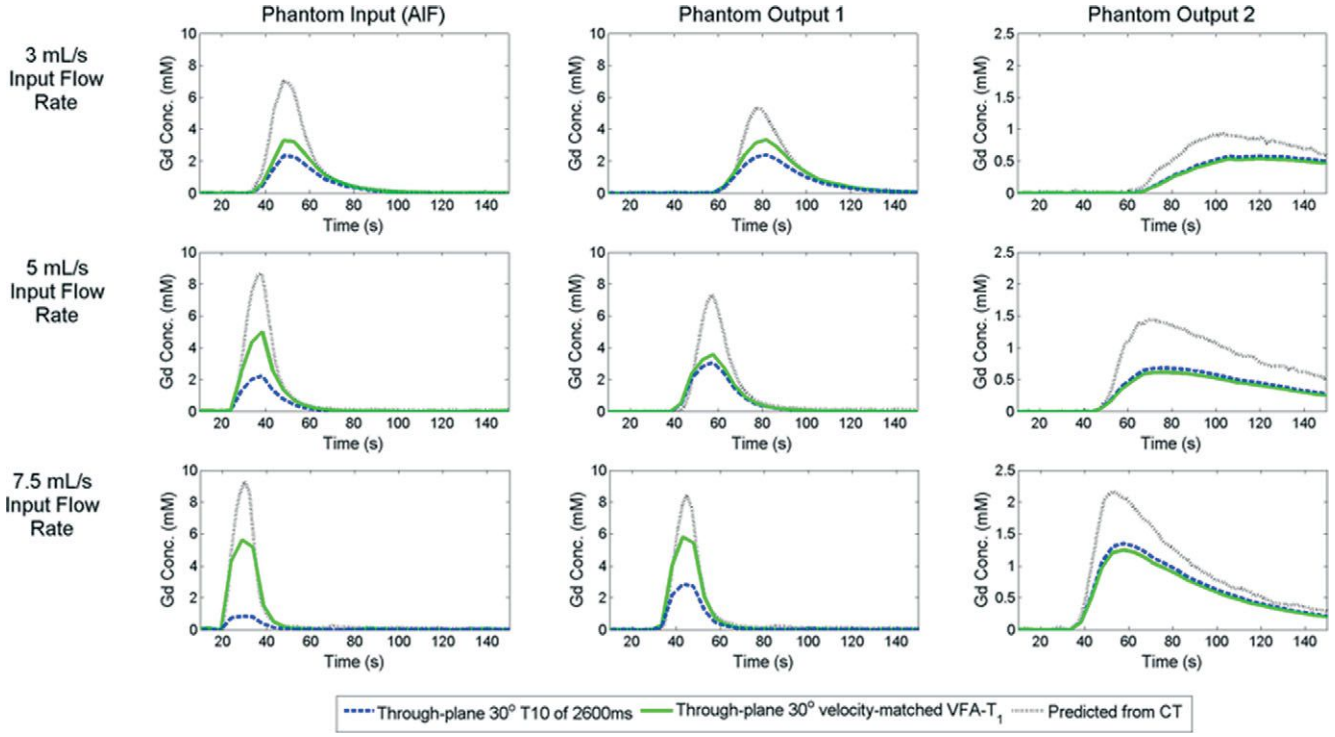


Figure 7. Accuracy in calculation of AIF_{MAGN} is improved using velocity-matched VFA-T1 at higher flow rates and Gd-DTPA concentrations. AIF_{MAGN} is compared to AIF_{CT} in a central slice at set peak concentration of 10 mM for through-plane flow at flow rates of (upper) 3, (middle) 5, and (lower) 7.5 mL/s for all 3 flow tubes (left—phantom input, middle—phantom output 1, right—phantom output 2).

background phase correction, the phase measurement tracked AIF_{CT} at Gd-DTPA concentrations above 1.6 mM, corresponding to phantom input and phantom output 1 tubes for input peak bolus >2 mM; yet, deviations were apparent in all phantom output 2 measurements and in all 1 mM peak bolus experiments. Figure 8B compares the same AIFs across varying input flow rates (3–7.5 mL/s) at a fixed input concentration of 10 mM. Under these conditions, AIF_{PHA} again tracked AIF_{CT} accurately.

Phase- and Magnitude-Derived AIFs: Comparison to CT. Pearson correlation analysis presented in Figure 9 reported the following trends: (A) bias in AIF_{MAGN} increased with Gd-DTPA concentration and reduced with flip angle; (B) T1 correction improved the AIF_{MAGN} measurement, but the 95% limits of agreement were prohibitively broad; and (C) phase-corrected AIF_{PHA} tracked AIF_{CT}. Difference analysis, summarized in Table 2, reported equivalence of AIF_{PHA} with AIF_{CT} within 1 mM for both peak concentration and within 20 mM*s AUC across all tested conditions, and that AIF_{MAGN} measurements approached equivalence with AIF_{CT} only at concentrations below 2 mM.

Phase- and Magnitude-Derived AIFs: Spatial Heterogeneity. For the through-plane flow, AIF_{PHA} reported mean and standard deviation values of 9.6 ± 0.5 mM for peak concentration and 28 ± 7 mM*s for AUC, across the middle 60% package of slices. For in-plane flow, AIF_{PHA} reported mean and standard deviation values of 9.2 ± 1 mM for peak concentration and 27 ± 14 mM*s for AUC, across the central 60% of the FOV. In comparison, T1-corrected AIF_{MAGN} using flip angles of 20° and 30° reported 4.0 ± 0.3 mM (through-plane) and 7 ± 2 mM (in-plane) for peak concentration, and 13 ± 1 mM*s (through-plane) and 20 ± 4 mM*s (in-plane) for AUC.

DISCUSSION

In this study a multimodality flow phantom was used to compare the AIFs derived from 3D-FLASH magnitude and phase signals, against the gold standard DCE-CT under similar conditions (including use of Gd-DTPA for CT investigation). Magnitude signal-derived AIF is sensitive to imaging orientation, flip angle, and in-flow effects, as demonstrated by prior authors. We show that implementation of position and velocity-matched T1 measurements can improve the magnitude signal-derived AIF measurement, yet equivalence to CT was noted only at peak Gd-DTPA concentrations to 2.5 mM. In comparison, phase-derived AIF showed equivalence to CT within 1 mM across the range of tested conditions, plus robustness to imaging orientation, flip angle, and in-flow effects. However, the phase AIF overshot the CT AIF for low concentrations.

Magnitude Signal-Derived AIF Measurements

Conversion of the magnitude signal to concentration using the standard FLASH signal equation leaves the AIF_{MAGN} measurement prone to a number of biases. Saturation of the nonlinear FLASH signal owing to T1 and T_2* properties of gadolinium is a known problem (29). In addition, the magnetic susceptibility

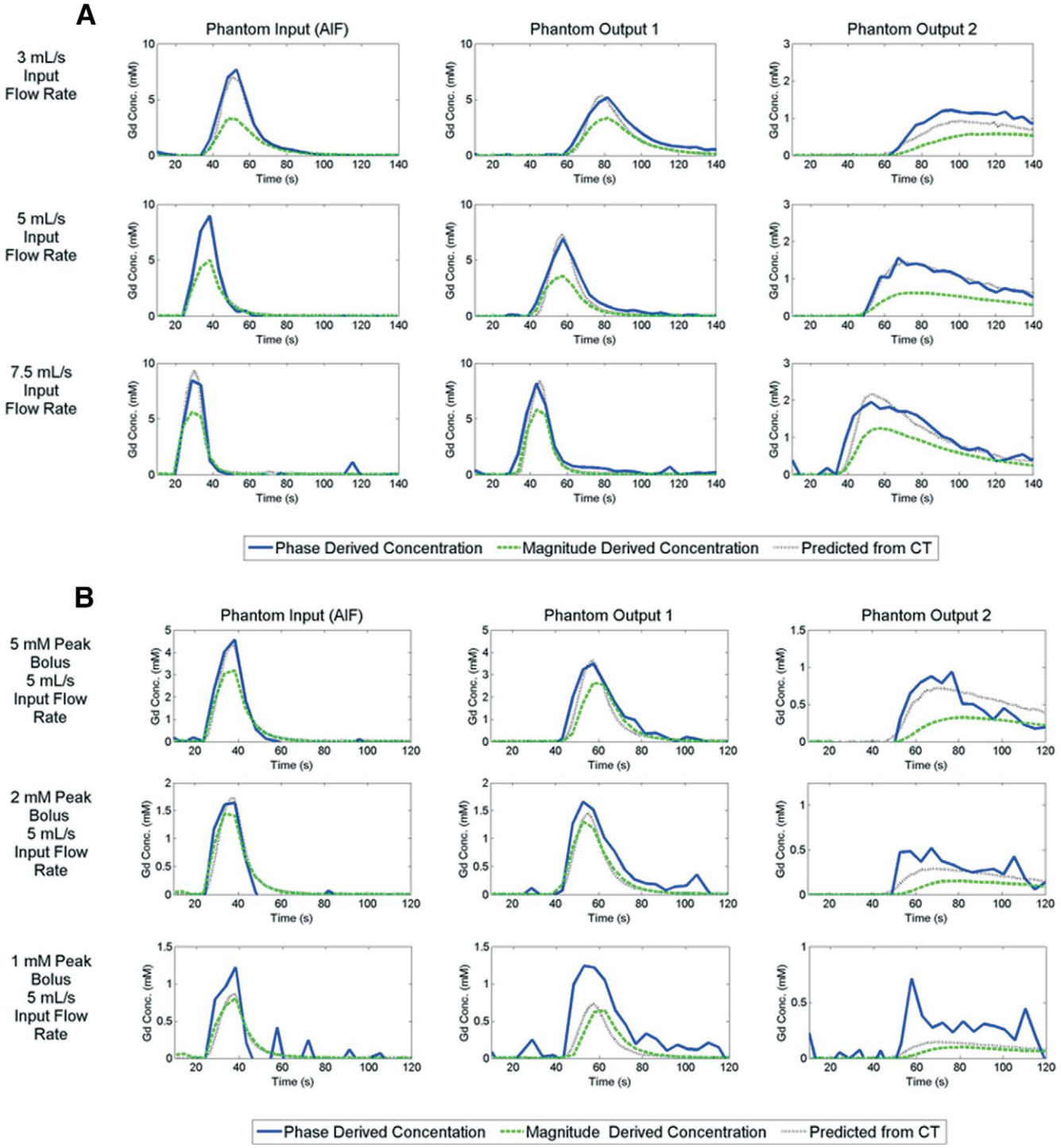


Figure 8. AIF_{PHA} compared to AIF_{MAGN} at a flip angle of 30° and AIF_{CT} for: (A) peak concentrations of 10 mM at increasing flow rates and for (B) 5 mL/s through-plane flow rate with increasing peak Gd-DTPA concentrations.

offset at peak bolus concentration may introduce a mis-registration artifact (18).

This research investigated RF and inflow biases to the 3D-FLASH magnitude signal, and it showed dramatic underestimation of AIF_{MAGN} metrics (28). RF and inflow biases were particularly severe for through-plane flow compared to in-plane flow, consistent with Garpenbring et al. (10). Inflow effects were partially compensated by incorporation of flow and position-matched VFA-T1 measurements into equation (1). The T1 correction is intuitive because the rate of inflow from outside of the imaging volume is captured as an acceleration of R_1 (26). However, even with this correction, our AIF_{MAGN} measurements were significantly different from AIF_{CT} except at peak Gd-DTPA concentrations <2.5 mM. Korporaal et al. also presented with a

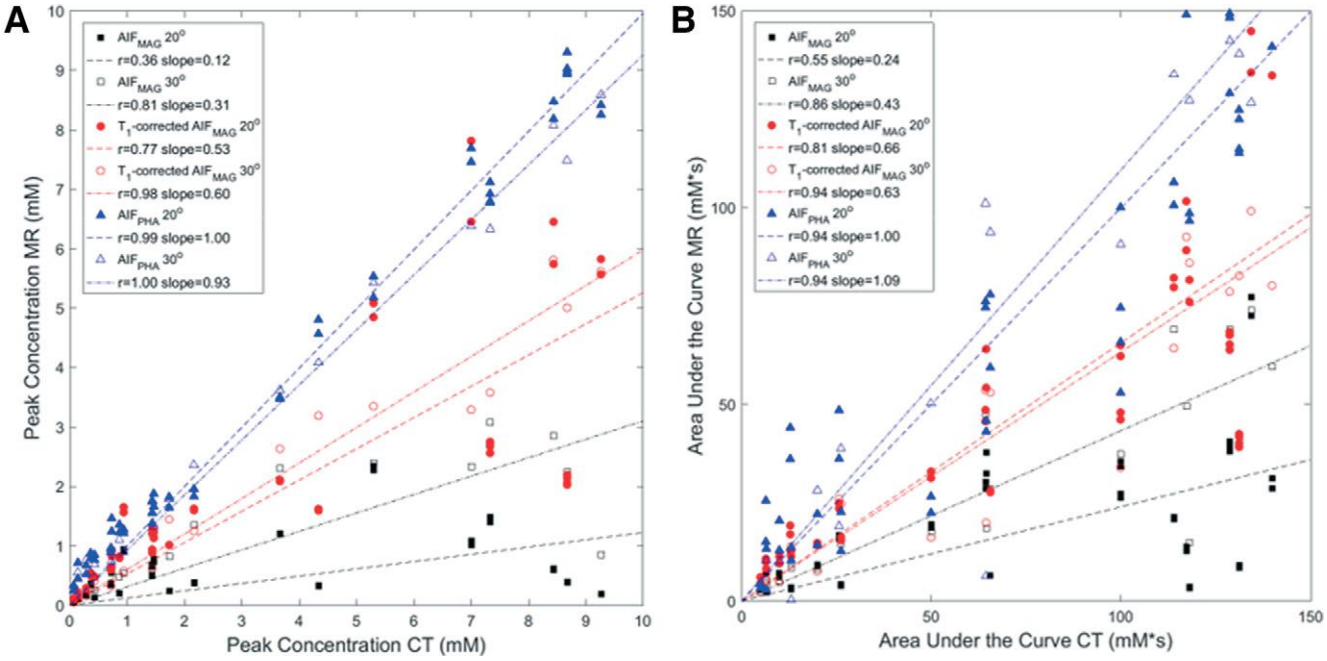


Figure 9. Pearson correlation results for AIF_{MAGN} and AIF_{PHA} compared to AIF_{CT}, for (A) peak concentration, and (B) AUC measurements within a central slice. The different shapes represent the 3 data types evaluated, pooled across velocities and concentrations for through-plane flow at flip angles of 20 (filled symbols, 16 runs, 48 measurements in input and output tubes) and 30° (open symbols, 7 runs, 21 measurements in input and output tubes).

considerably underestimated AIF_{MAGN} compared to AIF_{PHA} and AIF_{CT} targeting the femoral artery of patients with prostate cancer (17).

A key issue for the current multimodal phantom design is the need for a higher T1 relaxivity glycerol/water mixture to sustain CompuFlow pump performance. A peristaltic pump would convect water instead, at the cost of reproducibility in performance because of residual pulsatility, deviation from expected flow under high downstream pressure, and tube stretch over time. At the measured T1 relaxivity of 7.5 1/mM*s, the MR magnitude signal to Gd-DTPA concentration conversion saturates at lower concentrations than would be expected for Gd-DTPA within saline or plasma (T1 relaxivity, ~5.0 1/mM*s). This saturation is exacerbated at lower flip angles (eg, saturation of 10° flip angle data at ~2.5 mM in Figure 3C).

Dynamic measurement analysis at high Gd-DTPA concentrations can also be compromised by T_2* relaxation, because T_2* relaxation times of the 15% glycerol/water mixture appear to be well within an order of magnitude of the TE at concentrations above 5 mM. Our image processing assumed negligible T_2* relaxation, in part because experimental TE values are generally short relative to T2*, but it was also infeasible to measure T_2* at each Gd-DTPA concentration during the dynamic experiment. Schabel et al. published a nonlinear concentration-independent solution to the dynamic analysis problem, but logic and accurate knowledge of T_2* are necessary for selection of the correct concentration following saturation (28). Sufficient signal-to-noise must also exist for differentiation between concentrations, and the dynamic range of the signal across concentrations reduces with flip angle.

Another likely issue affecting the 3D-FLASH magnitude signal during dynamic experiments is its transient nature. At high flow rates, the magnetization may be exposed to an insufficient number of RF pulses to achieve steady-state condition, which is further compounded by spatially varying RF amplitudes. At low flow rates for through-plane flow, the magnetization may also be exposed to spatially varying RF amplitude along the shoulder region of the RF pulse slice profile. Consequently, some groups advocate AIF_{MAGN} measurement in ROI locations, where the steady-state condition is better satisfied (11, 21). Use of higher flip angles also improves AIF_{MAGN} robustness by accelerating the transition to steady state (25).

One may also expect better comparative performance at higher flip angles because of improved 3D-FLASH signal linearity with Gd-DTPA concentration, consistent with our results in Figure 2 and the published comparative measurements from Cron et al. (20). However, the improved signal linearity comes at a price of SNR and specific absorption ratio, factors which can prohibit implementation of high spatial resolution and high temporal resolution brain protocols with considerable coverage (eg, 1.5 × 1.5 × 3 mm spatial resolution, 6-second temporal resolution, 12 cm of through-plane coverage).

Phase-Signal-Derived AIF Measurements

The phase of the MR signal provides a mechanism for AIF quantification based on the magnetic susceptibility of Gd-DTPA. Our findings show that $AIFP_{HA}$ peak concentration measurements are equivalent within 1 mM to gold standard AIF_{CT} to a concentration of 10 mM, which covers the clinically relevant concentration range for AIF measurement (25). The AIF_{PHA} was

Table 2. 95% Limits of Agreement (LoA) for Peak Concentration and AUC Measurements Defined from Bland–Altman Difference Analysis Between MRI- and CT-derived AIF

Data Type	Data Range (mM)	Peak Concentration (mM)	AUC (mM*s)
AIF_{PHA}, FA 20°	0–10	0.1 ± 0.7	0.6 ± 37.0
AIF_{PHA}, FA 30°	0–10	−0.1 ± −0.9	3.6 ± 43.6
AIF_{PHA}, FA 20°	0–5	0.3 ± 0.4	3.4 ± 27.3
AIF_{PHA}, FA 30°	0–5	0.1 ± 0.4	2.6 ± 32.6
AIF_{PHA}, FA 20°	0–2	0.3 ± 0.4	5.9 ± 27.3
AIF_{PHA}, FA 30°	0–2	0.2 ± 0.4	−3.8 ± 20.3
Uncorrected AIF_{MAGN}, FA 20°	0–10	−2.8 ± 6.2	−52.5 ± 85.9
Uncorrected AIF_{MAGN}, FA 30°	0–10	−1.9 ± 4.9	−35.4 ± 64.7
Uncorrected AIF_{MAGN}, FA 20°	0–5	−0.9 ± 2.4	−15.5 ± 34.7
Uncorrected AIF_{MAGN}, FA 30°	0–5	−0.6 ± 1.6	−10.7 ± 24.1
Uncorrected AIF_{MAGN}, FA 20°	0–2	−0.4 ± 2.4	−6.8 ± 34.7
Uncorrected AIF_{MAGN}, FA 30°	0–2	−0.2 ± 0.5	−4.7 ± 8.2
T_1-corrected AIF_{MAGN}, FA 20°	0–10	−1.4 ± 4.3	−23.4 ± 58.4
T_1-corrected AIF_{MAGN}, FA 30°	0–10	−1.2 ± 2.8	−23.3 ± 44.6
T_1-corrected AIF_{MAGN}, FA 20°	0–5	−0.4 ± 1.7	−6.7 ± 23.9
T_1-corrected AIF_{MAGN}, FA 30°	0–5	−0.3 ± 0.8	−6.6 ± 19.6
T_1-corrected AIF_{MAGN}, FA 20°	0–2	−0.1 ± 1.7	−0.8 ± 23.9
T_1-corrected AIF_{MAGN}, FA 30°	0–2	−0.1 ± 0.2	0.6 ± 37.0

The table entries report the 95% LoA for each parameter as the average difference ± 1.96 standard deviation of the difference across variable concentrations and flow rates. The statistical analyses are repeated across 3 ranges of input tube concentrations.

also spatially robust in both in-plane and through-plane flow orientations, and to flow velocity and flip angle. This improved and more robust performance is consistent with the results from other groups. Korporaal et al. compared AIF_{MAGN}, AIF_{PHA}, and AIF_{CT}, targeting the femoral artery in patients with prostate cancer (17). Cron et al. favorably compared AIF_{PHA} to AIF_{MAGN} in the femoral artery, but without measurement of vascular T1 or AIF_{CT} validation (20). The same group has also applied phase imaging to calculate the venous input function in the superior sagittal sinus (30).

A limiting factor for the phase AIF measurement is precision at low Gd-DTPA concentrations, because the phase shift is small and imprecisely measured. Gd-DPTA increases the SNR in T1-weighted magnitude images, and phase noise varies inversely with the SNR in magnitude images (31). For this reason, the analysis of data presented in experiment 6 considers only the middle 60% of the imaging slices.

A number of factors can further impact the use of AIF_{PHA} in the clinical setting. First, 3D-FLASH phase reconstruction may not be accessible on clinical MRI systems in the absence of a research license or key. Second, vessel selection for AIF_{PHA} measurements may be limited to large vessels (eg, femoral artery, sagittal sinus, internal carotid artery) owing to the need for sufficient vessel diameter to reduce partial volume effects, and the increasing complexity of modeling of magnetic susceptibility effects when vessel orientation and shape diverges from that of a cylinder aligned in parallel with B_0 (22). Finally, phase wrap during bolus passage requires automated postprocessing, or manual correction by modulo 360° shifts in the phase image (32, 33).

A need for background phase correction is a complication of the AIF_{PHA} measurement. Our flow phantom experiments used a single control tube satisfactorily; yet, the phase signal calibration was stabilized against off-resonance effects when the Gd-DTPA-doped sample tube was surrounded with a hexagonal array of control tubes. Residual static field inhomogeneity will also exist across the brain in vivo, and within the sagittal sinus itself (34), which may complicate selection of an ROI for background phase correction in vivo. Our own experiences (unpublished) also suggest that the sagittal sinus AIF_{PHA} measurement is improved at lower flip angles for reasons that require further investigation and yet may suggest a combination of off-resonance and RF heating effects. However, any RF heating effects should be apparent in the brain parenchyma but not in the sinus owing to convection. Alternatively, field camera technology can capture the temporal and spatial history of resonance frequency changes during phase-sensitive acquisitions (35, 36). Also, the fat resonance provides a temperature-insensitive phase reference that should enable tracking of instrumentation-related resonance frequency changes (37).

Clinical Relevance

The value of individualized patient AIF acquisition is not yet fully understood. Port et al. showed that in 23% of patients, the individual's AIF differs from the population average by >50%

(38), and it is possible that the use of population-average AIFs may limit our ability to meaningfully interpret DCE-MRI findings, and this variability may contribute to the high inconsistency in permeability measures reported in prior DCE-MRI studies. Ashton et al. showed a 70% reduction of visit-to-visit coefficient of variation in permeability parameters using individual compared with population AIFs (39). However, several publications have shown equivalence in pharmacokinetic output parameters when DCE-MRI data are analyzed using population average or individualized measurements (13, 40). These findings may reflect on-going challenges to measure individual AIF accurately. This study shows that AIF_{PHA} could provide a feasible supplemental method for individual AIF acquisition with greater accuracy and robustness such that this approach may improve the consistency in the results of future DCE-MRI studies.

The technical requirements for a dynamic MRI phantom for quality assurance testing are currently not well understood. However, recognizing that site and system factors that compromise shim performance and temperature regulation of hardware components may compromise phase-based AIF evaluation within clinical trials, the roles of a flow phantom may include to measure and consider differences across sites and scanners when interpreting data, as well as to monitor system performance. It is important to note that factors affecting phase may not be captured by standardized QA protocols that focus on magnitude signal metrics. The phantom could also be modified to account for partial voluming, and vessels of smaller calibers.

In summary, we use a controlled multimodal flow phantom that is validated against AIF_{CT} to show that AIF_{PHA} tracks peak Gd-DTPA concentration within 1 mM, and AUC within 44 mM*s, over a range of tested conditions. The robustness of the AIF_{PHA} measurements was also apparent across the imaged volume. In comparison, AIF_{MAGN} measurements were highly sensitive to imaging plane orientation, flip angle selection, and flow velocity, and equivalent performance to AIF_{CT} was shown at only Gd-DTPA concentrations <2 mM. Improving the accuracy of the AIF should reduce variability in pharmacokinetic output parameters, and thereby, it should increase the potential for meaningful interpretation of the changes in vascular permeability using DCE-MRI.

ACKNOWLEDGMENTS

This work was partially supported by Discovery Grant 386277 from the National Science and Research Council of Canada (NSERC), the Brain Tumor Foundation of Canada Pilot Grant, and the Princess Margaret Cancer Foundation—Department of Radiation Oncology Academic Enrichment Fund.

Equal Contribution: Coolens C. and Chung C. contributed equally to this paper.

Disclosures: No disclosures to report.

Conflict of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. O'Connor JP, Jackson A, Parker GJ, Roberts C, Jayson GC. Dynamic contrast-enhanced MRI in clinical trials of antivascular therapies. Nat Rev Clin Oncol. 2012;9:167–177.
2. Leach M, Morgan B, Tofts P, Buckley DL, Huang W, Horsfield MA, Chenevert TL, Collins DJ, Jackson A, Lomas D, Whitcher B, Clarke L, Plummer R, Judson I, Jones R, Alonzi R, Brunner T, Koh DM, Murphy P, Waterton JC, Parker G, Graves MJ, Scheenen TW, Redpath TW, Orton M, Karczmar G, Huisman H, Barentsz J, Padhani A; Experimental Cancer Medicine Centres Imaging Network Steering Committee. Imaging vascular function for early stage clinical trials using dynamic contrast-enhanced magnetic resonance imaging. Eur Radiol. 2012;22:1451–1464.
3. O'Connor JPB, Jackson A, Parker GJM, Jayson GC. DCE-MRI biomarkers in the clinical evaluation of antiangiogenic and vascular disrupting agents. Brit J Cancer. 2007;96:189–195.
4. Heye T, Davenport MS, Horvath JJ, Feuerlein S, Breault SR, Bashir MR, Merkle EM, Boll DT. Reproducibility of dynamic contrast-enhanced MR imaging. Part I. Perfusion characteristics in the female pelvis by using multiple computer-aided diagnosis perfusion analysis solutions. Radiology. 2013;266:801–811.
5. Heye T, Merkle EM, Reiner CS, Davenport MS, Horvath JJ, Feuerlein S, Breault SR, Gall P, Bashir MR, Dale BM, Kiraly AP, Boll DT. Reproducibility of dynamic contrast-enhanced MR imaging. Part II. Comparison of intra- and interobserver variability with manual region of interest placement versus semiautomatic lesion segmentation and histogram analysis. Radiology. 2013;266:812–821.
6. Tofts PS. Modeling tracer kinetics in dynamic Gd-DTPA MR imaging. J Magn Reson Imaging. 1997;7:91–101.
7. Yang X, Knopp MV. Quantifying tumor vascular heterogeneity with dynamic contrast-enhanced magnetic resonance imaging: a review. J Biomed Biotechnol. 2011;2011:1–12.
8. Yang X, Liang J, Heverhagen JT, Jia G, Schmalbrock P, Sammet S, Koch R, Knopp MV. Improving the pharmacokinetic parameter measurement in dynamic contrast-enhanced MRI by use of the arterial input function: Theory and clinical application. Magn Reson Med. 2008;59:1448–1456.
9. Peeters F, Annet L, Hermoye L, van Beers BE. Inflow correction of hepatic perfusion measurements using T1-weighted, fast gradient-echo, contrast-enhanced MRI. Magn Reson Med. 2004;51:710–717.
10. Garpebring A, Wirestam R, Östlund N, Karlsson M. Effects of inflow and radiofrequency spoiling on the arterial input function in dynamic contrast-enhanced MRI: A combined phantom and simulation study. Magn Reson Med. 2011;65:1670–1679.
11. Roberts C, Little R, Watson Y, Zhao S, Buckley DL, Parker GJ. The effect of blood inflow and B1-field inhomogeneity on measurement of the arterial input function in axial 3D spoiled gradient echo dynamic contrast-enhanced MRI. Magn Reson Med. 2011;65:108–119.
12. van Osch MJP, Vonken EPA, Viergever MA, van der Grond J, Bakker CJ. Measuring the arterial input function with gradient echo sequences. Magn Reson Med. 2003;49:1067–1076.
13. Meng R, Chang SD, Jones EC, Goldenberg SL, Kozlowski P. Comparison between population average and experimentally measured arterial input function in predicting biopsy results in prostate cancer. Acad Radiol. 2010;17:520–525.
14. Li X, Welch EB, Arlinghaus LR, Chakravarthy AB, Xu L, Farley J, Loveless ME, Mayer IA, Kelley MC, Meszoely IM, Means-Powell JA, Abramson VG, Grau AM, Gore JC, Yankeelov TE. A novel AIF tracking method and comparison of DCE-MRI parameters using individual and population-based AIFs in human breast cancer. Phys Med Biol. 2011;56:5753–5769.
15. Yankeelov TE, Cron GO, Addison CL, Wallace JC, Wilkins RC, Pappas BA, Santyr GE, Gore JC. Comparison of a reference region model with direct measurement of an AIF in the analysis of DCE-MRI data. Magn Reson Med. 2007;57:353–361.
16. Lavini C, Verhoeff JJC. Reproducibility of the gadolinium concentration measurements and of the fitting parameters of the vascular input function in the superior sagittal sinus in a patient population. Magn Reson Imaging. 2010;28:1420–1430.
17. Korporaal JG, van den Berg CA, van Osch MJ, Groenendaal G, van Vulpen M, van der Heide UA. Phase-based arterial input function measurements in the femoral arteries for quantification of dynamic contrast-enhanced (DCE) MRI and comparison with DCE-CT. Magn Reson Med. 2011;66:1267–1274.
18. Akbudak E, Norberg RE, Conturo TE. Contrast-agent phase effects: An experimental system for analysis of susceptibility, concentration, and bolus input function kinetics. Magn Reson Med. 1997;38:990–1002.
19. Garpenbring A, Wirestam R, Yu J, Asklund T, Karlsson M. Phase-based arterial input functions in humans applied to dynamic contrast-enhanced MRI: potential usefulness and limitations. Magn Reson Mater Phys. 2011;24:233–245.

20. Cron GO, Foottit C, Yankeelov TE, Avruch LI, Schweitzer MR, Cameron I. Arterial input functions determined from MR signal magnitude and phase for quantitative dynamic contrast-enhanced MRI in the human pelvis. Magn Reson Med. 2011; 66:498–504.
21. Driscoll B, Keller H, Jaffray DA, Coolens C. Development of a dynamic quality assurance testing protocol for multisite clinical trial DCE-CT accreditation. Med Phys. 2013; 40:081906.
22. Schenck JF. The role of magnetic susceptibility in magnetic resonance imaging: MRI magnetic compatibility of the first and second kinds. Med Phys. 1996;23:815–850.
23. Cheng H-LM, Wright GA. Rapid high-resolution T1 mapping by variable flip angles: accurate and precise measurements in the presence of radiofrequency field inhomogeneity. Magn Reson Med. 2006;55:566–574.
24. Takeuchi S, Karino T. Flow patterns and distributions of fluid velocity and wall shear stress in the human internal carotid and middle cerebral arteries. World Neurosurg. 2010;73:174–185.
25. De Naeyer D, Verhulst J, Ceelen W, Segers P, De Deene Y, Verdonck P. Flip angle optimization for dynamic contrast-enhanced MRI-studies with spoiled gradient echo pulse sequences. Phys Med Biol. 2011;56:5373–5395.
26. Bauer WR, Hiller KH, Roder F, Rommel E, Ertl G, Haase A. Magnetization exchange in capillaries by microcirculation affects diffusion-controlled spin-relaxation: a model which describes the effect of perfusion on relaxation enhancement by intravascular contrast agents. Magn Reson Med. 1996;35:43–55.
27. Brookes JA, Redpath TW, Gilbert FJ, Murray AD, Staff RT. Accuracy of T1 measurement in dynamic contrast-enhanced breast MRI using two- and three-dimensional variable flip angle fast low-angle shot. J Magn Reson Imaging. 1999; 9:163–171.
28. Schabel MC, Parker DL. Uncertainty and bias in contrast concentration measurements using spoiled gradient echo pulse sequences. Phys Med Biol. 2008;53: 2345–2373.
29. de Bazelaire C, Rofsky N, Duhamel G, Zhang J, Michaelson MD, George D, Alsop DC. Combined T2* and T1 measurements for improved perfusion and permeability studies in high field using dynamic contrast enhancement. Eur Radiol. 2006;16:2083–2091.
30. Foottit C, Cron GO, Hogan MJ, Nguyen TB, Cameron I. Determination of the venous output function from MR signal phase: feasibility for quantitative DCE-MRI in human brain. Magn Reson Med. 2010;63:772–781.
31. Conturo TE, Smith GD. Signal-to-noise in phase angle reconstruction: dynamic range extension using phase reference offsets. Magn Reson Med. 1990;15:420–437.
32. Liu J, Drangova M. Intervention-based multidimensional phase unwrapping using recursive orthogonal referring. Magn Reson Med. 2012;68:1303–1316.
33. Chavez S, Xiang Q-S, An L. Understanding phase maps in MRI: a new cutline phase unwrapping method. IEEE T Med Imaging. 2002;21:966–977.
34. Jain V, Langham MC, Wehrli FW. MRI estimation of global brain oxygen consumption rate; J Cereb Blood Flow Metab. 2010;30:1598–1607.
35. De Zanche N, Barmet C, Nordmeyer-Massner JA, Pruessmann KP. NMR probes for measuring magnetic fields and field dynamics in MR systems. Magn Reson Med. 2008;60:176–186.
36. Boer VO, van de Bank BL, van Vliet G, Luijten PR, Klomp DW. Direct B0 field monitoring and real-time B0 field updating in the human breast at 7 Tesla. Magn Reson Med. 2012;67:586–591.
37. Hofstetter LW, Yeo DT, Dixon WT, Kempf JG, Davis CE, Foo TK. Fat-referenced MR thermometry in the breast and prostate using IDEAL. J Magn Reson Imaging. 2012;36:722–732.
38. Port RE, Knopp MV, Brix G. Dynamic contrast-enhanced MRI using Gd-DTPA: interindividual variability of the arterial input function and consequences for the assessment of kinetics in tumors. Magn Reson Med. 2001;45:1030–1038.
39. Ashton E, Raunig D, Ng C, Kelcz F, McShane T, Evelhoch J. Scan-rescan variability in perfusion assessment of tumors in MRI using both model and data-derived arterial input functions. J Magn Reson Imaging. 2008;28:791–796.
40. Hormuth DA 2nd, Skinner JT, Does MD, Yankeelov TE. A comparison of individual and population-derived vascular input functions for quantitative DCE-MRI in rats. Magn Reson Imaging 2014;32:397–401.

TOMOGRAPHY® RESEARCH ARTICLE

Early Prediction of Breast Cancer Therapy Response using Multiresolution Fractal Analysis of DCE-MRI Parametric Maps

Archana Machireddy, Guillaume Thibault, Alina Tudorica, Aneela Afzal, May Mishal, Kathleen Kemmer, Arpana Naik, Megan Troxell, Eric Goranson, Karen Oh, Nicole Roy, Neda Jafarian, Megan Holtorf, Wei Huang, and Xubo Song

Oregon Health and Science University, Portland, OR

Corresponding Author:
Xubo Song, PhD
Department of Medical Informatics and Clinical Epidemiology & Center for Spoken Language Understanding, Oregon Health and Science University, Portland, OR, USA;
E-mail: songx@ohsu.edu

Key Words: breast cancer, DCE-MRI, neoadjuvant chemotherapy, early prediction, multiresolution fractal analysis

Abbreviations: Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI), neoadjuvant chemotherapy (NACT), area under the curve (AUC), pathological complete response (pCR), gray-level co-occurrence matrix (GLCM), run length matrix (RLM), receiver operating curve (ROC), fractal dimension (FD), residual cancer burden (RCB), region of interest (ROI)

ABSTRACT

We aimed to determine whether multiresolution fractal analysis of voxel-based dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) parametric maps can provide early prediction of breast cancer response to neoadjuvant chemotherapy (NACT). In total, 55 patients underwent 4 DCE-MRI examinations before, during, and after NACT. The shutter-speed model was used to analyze the DCE-MRI data and generate parametric maps within the tumor region of interest. The proposed multiresolution fractal method and the more conventional methods of single-resolution fractal, gray-level co-occurrence matrix, and run-length matrix were used to extract features from the parametric maps. Only the data obtained before and after the first NACT cycle were used to evaluate early prediction of response. With a training (N = 40) and testing (N = 15) data set, support vector machine was used to assess the predictive abilities of the features in classification of pathologic complete response versus non-pathologic complete response. Generally the multiresolution fractal features from individual maps and the concatenated features from all parametric maps showed better predictive performances than conventional features, with receiver operating curve area under the curve (AUC) values of 0.91 (all parameters) and 0.80 (K^{trans}), in the training and testing sets, respectively. The differences in AUC were statistically significant ($P < .05$) for several parametric maps. Thus, multiresolution analysis that decomposes the texture at various spatial-frequency scales may more accurately capture changes in tumor vascular heterogeneity as measured by DCE-MRI, and therefore provide better early prediction of NACT response.

INTRODUCTION

Breast cancer is the second leading cause of cancer death among all cancers occurring in American women (1). The survival rate and prognosis of a patient with breast cancer is dependent on the stage of cancer at diagnosis. Locally advanced breast cancers (generally with tumor size >2 cm) are often treated with neoadjuvant chemotherapy (NACT) before surgery to reduce the tumor size for breast-conserving surgery (2, 3). A pathological complete response (pCR) to NACT is considered a surrogate marker for overall and long-term disease-free survival (4). However, the pCR rate is only 6%–45% depending on breast cancer subtypes and treatment regimen (5, 6, 7, 8). It is therefore important to identify the nonresponders at an early stage so that their treatment regimen can be modified, sparing them the long- and short-term toxicities from ineffective chemotherapies. Currently, in standard of care, the response to NACT is evaluated based on the histological examination of a surgical specimen taken after the completion of NACT. Noninvasive or minimally invasive methods that can predict therapy response at the early stages of NACT can potentially play an important role in the emerging era of precision medicine to help guide regimen de-escalation/alteration in NACT treatment of breast cancer (9).

A significant change in the microenvironment of the tumor, such as perfusion/permeability and metabolism, usually precedes a reduction in tumor size as response to chemotherapy (10–13). As a noninvasive imaging method for assessment of microvascular perfusion/permeability, dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) is increasingly used in research and early-phase clinical trial settings to predict and evaluate cancer response to treatment (9, 10). Several stud-

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2018.00046

ies (9, 14-22) have shown that changes in quantitative parameters estimated from pharmacokinetic modeling of DCR-MRI data can be useful markers for early prediction of breast cancer response to NACT.

When compared to normal tissue vasculature, tumor vasculature exhibits greater spatiotemporal heterogeneity. The heterogeneity of the tumor vasculature reflects the tumor stage and the disease progression (23). The aforementioned DCE-MRI studies (9, 14-22) generally reported changes in mean parameter values of the entire breast tumor, masking the potential changes in spatial heterogeneity of the microvasculature in response to NACT. Image texture features that can capture the heterogeneity of tumor vasculature from DCE-MRI images or voxel-based parametric maps could be highly useful in assessing tumor response to therapy. Several texture analysis methods such as gray-level co-occurrence matrix (GLCM) and gray-level run length matrix (RLM) have been frequently used in DCE-MRI analysis (24, 25). They were initially used on DCE-MRI images directly. Teruel et al. (24) analyzed T1-weighted DCE-MRI images using GLCM features to predict breast cancer response to NACT. They extracted 16 textural features at each time point of a DCE-MRI acquisition, and the most significant feature yielded a receiver operating curve (ROC) area under the curve (AUC) of 0.77 for prediction of pCR versus stable disease. Similarly, Golden et al. (25) used GLCM features from pre- and post-NACT 2-dimensional (2D) DCE image slices to evaluate NACT response. The pre-NACT features were able to predict pCR with an AUC of 0.68. Although the post-NACT features showed more favorable performances in predicting pCR, these were obtained after the completion of NACT and were not useful for early prediction of pCR. Several studies have performed the same texture analysis on voxel-based maps of pharmacokinetic parameters estimated from pharmacokinetic modeling of DCE-MRI data. Banerjee et al. (26) extracted a combination of intensity, texture, shape, and edge-based features from 2D maps of pharmacokinetic parameters before and after NACT to assess treatment response. Their best model obtained an AUC of 0.83, using a concatenation of Riesz and first-order statistical features. However, the use of only pre- and post-NACT data limits the utility of this model for early prediction of NACT response. In our previous study, we (27) have extracted multiple statistical texture features from 3-dimensional (3D) pharmacokinetic parametric maps before and after 1 cycle of NACT, and found that 3D GLCM features were most effective for early prediction of NACT response through correlation with index values of residual cancer burden (RCB) using a regression model.

In all the analysis methods described above, texture has been studied on a statistical level, by analyzing the spatial distribution of the gray-level values. Textures can also be characterized by fractals, which describe irregular structures that show self-similarity at various scales. Fractal-based texture analysis correlates texture heterogeneity to fractal dimension (FD), which is a mathematical descriptor of a structure's geometrical complexity, based on the concept of spatial pattern self-similarity. Rose et al. (28) showed that fractal analysis could be used to quantify spatial heterogeneity in DCE-MRI parametric maps and differentiate between low- and high-grade tumors. Several other studies (29, 30) have used fractal analysis of breast DCE-MRI images to classify benign versus malignant tumors.

Another important aspect while considering textures is the scale. It has been shown that human visual system processes information in a multiscale approach (different cells in the visual cortex respond to different frequencies and orientations) (31). Owing to the highly heterogeneous nature of the tumor vasculature, analyzing images at a single resolution may not be able to capture the entire complexity of the tumor vasculature. A multiresolution approach can decompose an image into different levels of resolution, giving an opportunity to extract informative features at each level. Lower resolution levels best represent large structures or high contrast, while higher resolutions describe small size or low-contrast objects (32). Multiresolution analysis gives the advantage of analyzing both small- and large-object characteristics in a single image at several resolutions and therefore may be better suited to describe the highly heterogeneous tumor vasculature structure. Multiresolution methods, such as the wavelet analysis, transform images into a representation containing both frequency and spatial information (33). The mean and entropy values extracted from the subimages resulting from wavelet decomposition of DCE-MRI images have been used to classify malignant and benign breast tumors (34, 35). Braman et al. (36) used Gabor wavelet, co-occurrence measures and energy measures to generate 1980 features from DCE images to predict breast cancer response to NACT. A feature selection step was carried out to select top 10 features for final classification. Al-Kadi et al. combined wavelet analysis with fractal analysis and used multiresolution fractal descriptors on ultrasonography images to characterize the tissue and showed that tumor heterogeneity described by this feature improved prediction of response to therapy and disease characterization (37). To the best of our knowledge, fractal analysis at multiple resolutions has not been conducted on breast MRI images for prediction of response to NACT. In this preliminary study, we evaluated the potential of multiresolution fractal analysis of volumetric DCE-MRI pharmacokinetic parametric maps for early prediction of breast cancer response to NACT, and compared it with the conventional methods of GLCM, RLM and single-resolution fractal analysis.

MATERIALS AND METHODS

Patient Cohort and Study Schema

In total, 55 patients diagnosed with locally advanced breast cancer received standard-of-care NACT. They were consented to participate in a longitudinal research DCE-MRI study approved by the local IRB. The NACT regimen typically consists of 4 cycles of doxorubicin–cyclophosphamide administration every 2 weeks followed by 4 cycles of taxane every 2 weeks, or 6 cycles of the combination of all 3 drugs every 3 weeks (9, 27). The targeted agent trastuzumab was added to the regimen for tumors with positive HER2 (human epidermal growth factor receptor 2) receptor status. A full NACT course therefore would normally last 4–5 months.

In total, 4 DCE-MRI examinations were performed before, during, and after the NACT course: pre-NACT (visit-1), after the first NACT cycle (visit-2), at NACT midpoint (visit-3; usually

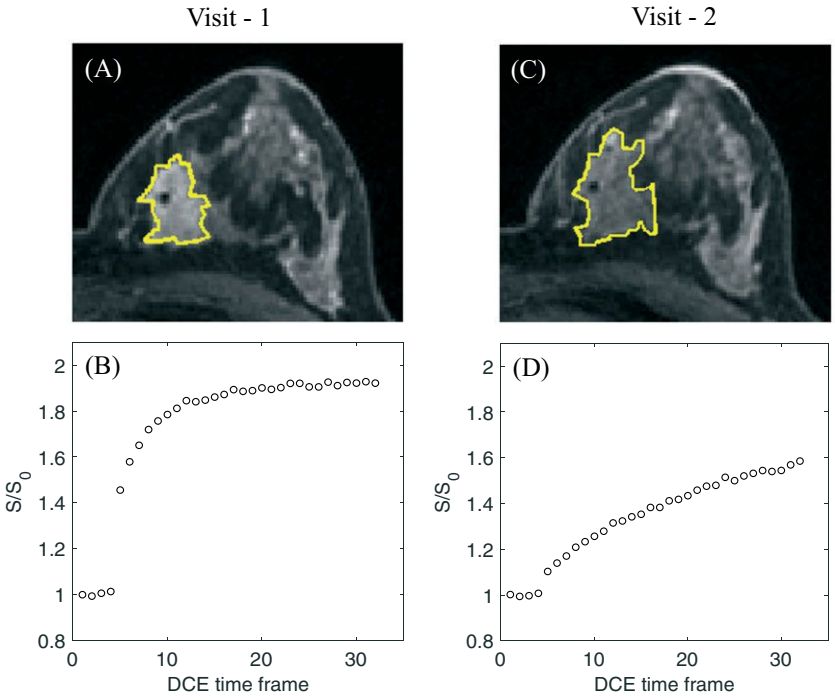


Figure 1. The visit-1 and visit-2 postcontrast dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) image slice (A and C, respectively) through the center of the primary breast tumor of a pathologic complete response (pCR) patient [35 years, grade 2 invasive ductal carcinoma, 2.9 cm in the longest diameter at visit-1, ER (estrogen receptor) −, PR (progesterone receptor) +, HER2 + receptor status]. The tumor ROI boundaries are shown in yellow. The time courses of mean signal intensity ratio, S/S_0, in the tumor ROI are shown in B and D for visit-1 and visit-2, respectively. S_0: signal intensity at baseline before contrast injection.

after 3 or 4 cycles of NACT, or before the change of NACT agents), and after the completion of NACT but before surgery (visit-4). Except for the visit-1 examination, all examinations were performed at least a week after administering the latest cycle of NACT agents to allow time for the drugs to take effect. Pathological analysis of the post-NACT surgical specimens was performed to determine the status of pathologic response to NACT. The values of cross-sectional size of the tumor in 2D, tumor cell density, number of lymph nodes involved, and the greatest dimension in the largest involved node were measured and used in the equation given by Symmans et al. (38) to compute the RCB. A pCR is defined as the absence of residual invasive tumor, indicated by RCB = 0. Non-pCR includes all cases with RCB > 0.

In this preliminary study, only data from the visit-1 and visit-2 DCE-MRI studies were used for image feature analysis and correlations with response endpoint of pCR versus non-pCR to assess the capability for early prediction of breast cancer response to NACT.

DCE-MRI Data Acquisition and Analysis

DCE-MRI data acquisition was performed using a Siemens 3 T system (Siemens, Erlangen, Germany) with the body coil as the transmitter and a 4-channel bilateral phased-array breast coil as the receiver. During each MRI session, following pilot scans and precontrast axial T1- and T2-weighted MRI acquisitions, axial bilateral DCE-MRI images with full breast coverage were acquired using a 3D gradient echo-based Time-resolved angiography With Stochastic Trajectories (TWIST) sequence (9). DCE-MRI acquisition parameters included the following: flip angle = 10°, echo time/repetition time = 2.9/6.2 milliseconds, parallel imaging acceleration factor of 2, field of view = 30 to 34 cm, in-plane matrix size = 320 × 320, and slice thickness = 1.4 mm. About 32–34 image volume sets of 104–128 slices each were acquired over a period of about 10 minutes with a temporal resolution of 14–20 seconds. The contrast agent gadolinium (HP-DO3A) was injected intravenously (0.1 mmol/kg at 2 mL/s) using a programmable power injector after acquisition of 2 baseline image volumes, followed by a 20-mL saline flush at the same injection rate.

Three experienced breast radiologists manually delineated the tumor region of interest (ROI) on postcontrast (90–120 seconds after the injection of the contrast agent) DCE-MRI image slices that contained the contrast-enhanced tumor. To minimize interobserver variability in tumor ROI drawing for the same patient, 1 radiologist drew ROIs for the entire longitudinal study of a single patient. With only 2 patients having multifocal disease, ROIs were drawn for the primary breast tumors only. Figure 1 shows an example of postcontrast DCE images from a pCR patient, drawn tumor ROIs, and ROI mean signal intensity ratio time-courses at visit-1 and visit-2. For pharmacokinetic analysis, precontrast tissue T1 value, T10, was determined using a proton density method (9) by acquiring proton density images just before DCE-MRI that were spatially coregistered with the DCE images. The DCE time-course data from the voxels within the tumor ROI was fitted with a 2-compartment–3-parameter shutter-speed model (9, 39), using a population-averaged arterial input function from the axillary artery (9). This pharmacokinetic analysis yielded the following 4 parameters: K^{trans} (volume transfer rate constant), v_e (volume fraction of extravascular and extracellular space), k_{ep} ($=K^{trans}/v_e$, efflux rate constant), and τ_i (mean intracellular water lifetime). Figure 2 shows examples of voxel-based parametric maps of these 4 parameters

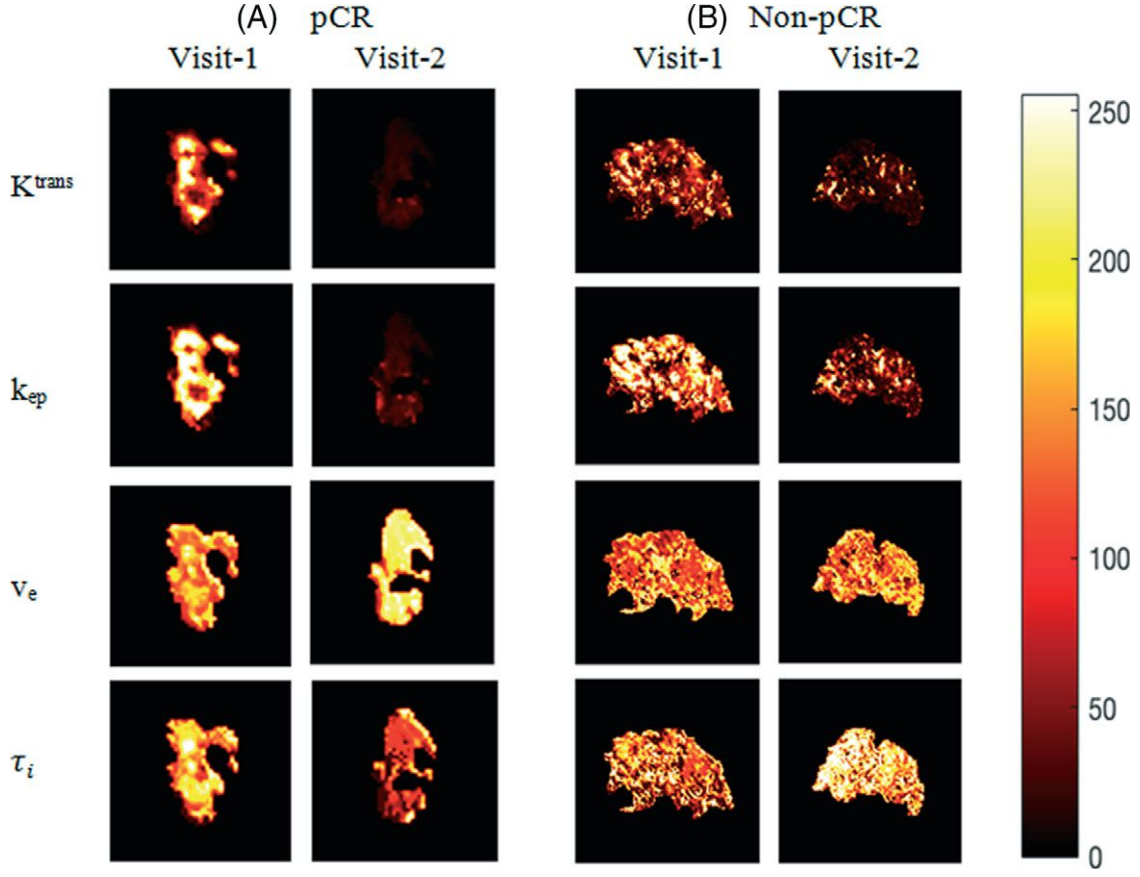


Figure 2. The visit-1 and visit-2 parametric maps of K^{trans}, k_{ep}, v_e, and τ_i of the tumor ROI on an image slice through the center of the tumor: a 27-year-old pCR patient with a grade 3 invasive ductal carcinima (5.0 cm in the longest diameter at visit-1) and ER −, PR −, HER2 + receptor status (A); a 45-year-old non-pCR patient with a grade 2 invasive mammary carcinoma (11.9 cm in the longest diameter at visit-1) and ER +, PR +, HER2 − receptor status (B).

for a pCR (Figure 2A) and a non-pCR (Figure 2B) tumor at visit-1 and visit-2. The parametric maps from the visit-1 and visit-2 studies of all patients were subjected to multiresolution fractal analysis described in detail below, as well as the traditional texture analysis methods of GLCM, RLM and single-resolution fractal analysis.

Multiresolution Fractal Analysis

Each of the parametric maps was decomposed into a multiresolution representation using wavelet analysis, and subsequently, FDs were calculated at each resolution level.

Wavelet Analysis for Multiresolution Decomposition. Wavelet analysis is used to decompose the parametric maps into a set of frequency sub-bands based on small basis functions of varying frequency and limited time duration called wavelets, enabling the characterization of texture at appropriate frequency levels. The wavelet is scaled and translated to cover the time-frequency domain. The discrete wavelet transform for a function $f(x, y, z)$ of size (M, N, K) can be represented as:

$$W_\varphi(j_0, m, n, k) = \frac{1}{\sqrt{MNK}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} \sum_{z=0}^{K-1} f(x, y, z) \varphi_{j_0,m,n,k}(x, y, z) \quad (1)$$

$$W^i_\psi(j_0, m, n, k) = \frac{1}{\sqrt{MNK}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} \sum_{z=0}^{K-1} f(x, y, z) \psi^i_{j, m, n, k}(x, y, z) \quad i = \{H, V, D\} \quad (2)$$

where $W_\varphi(j_0, m, n, k)$ is the approximation of $f(x, y, z)$ at scale j_0, $W^i_\psi(j, m, n, k)$ coefficients define the horizontal (H), vertical (V, and diagonal (D) details for scales $j \geq j_0$, φ is the scaling function, and ψ is the wavelet function (40). The wavelet transform depends mainly on the scaling (φ) and wavelet (ψ) functions, but it is not necessary to define their explicit form. Instead, a low-pass and high-pass filter that characterize the interaction of these functions are used. The process of decomposing the parametric maps can be viewed as passing them through a series of low-pass and high-pass filters and down-sampling successively. The 3D volume is first filtered along the columns resulting in a low-pass-filtered subvolume and a high-pass-filtered subvolume. These resulting subvolumes are further filtered along rows and slices resulting in 8 decomposed subvolumes. We have used Daubechies wavelets, as these filters have been designed to account for signal discontinuities and self-similarity, which make them the most suitable wavelet for describing signals exhibiting fractal patterns (41). Unlike Haar wavelet, they use overlapping windows that help capture changes in high frequency, and they also demonstrate

better recognition of fine characteristic structures (40). One level of decomposition results in 8 subvolumes. The FD for each of these subvolumes is calculated.

Multiresolution Fractal Analysis. The FD is calculated based on the power spectrum analysis of the 3D Fourier transformation of the subvolumes (42). The 3D discrete Fourier transform is defined as:

$$F(x,\ y,\ z) = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\sum_{k=0}^{K-1} I(m,\ n,\ k)e^{-j2\pi\left(x\frac{m}{M}+y\frac{n}{N}+z\frac{k}{K}\right)} \quad (3)$$

where I(m, n, k) is the 3D volume of size (M, N, K) and x, y, and z are the spatial frequencies. The power spectral density (P) is estimated as:

$$P(x,\ y,\ z) = |F(x,\ y,\ z)|^2 \quad (4)$$

The frequency space is evenly divided into 12 zenith and 24 azimuth directions, and 30 points are uniformly sampled along the radial component in each of these directions. The power spectral density is plotted against sampled radial frequency in a log-log plot. The slope β of a least-squares regression line of the log-log plot is related to the FD as:

$$FD = \frac{11-\beta}{2} \quad (5)$$

In standard wavelet analysis, the energy of the subvolume is used to guide further decomposition, but this value is highly dependent on the intensity values of the subvolume. In this work, instead of using energy, the subvolume with the highest FD was selected for further decomposition.

Each parametric map was decomposed down to 4 levels, using FD to guide the sub-band tree structure. Finally, we concatenated the highest and the lowest FD at each level of decomposition to form a feature vector. Therefore, for each parametric map an 8-dimensional feature vector is generated from multiresolution FD analysis.

Conventional Texture Feature Analysis

We compared the performance of multiresolution fractal analysis features with that of GLCM, RLM, and single-resolution fractal analysis. GLCM is a second-order statistical method, which estimates the joint probability $P(i, j \mid d\ \theta)$, where 2 voxels with intensity i and j are separated by distance d and direction θ. A GLCM matrix was constructed by averaging the matrices obtained over 13 directional offsets at distance d = 1 (27). Twelve Harlick features (43) were derived from this GLCM matrix. RLM $P(i,\ r \mid \theta)$ is defined as the number of pixels with gray-level i and run-length r, for a given direction θ. RLM was computed by adding all possible run lengths in the 13 directions of the 3D space and 13 statistical features were derived from this matrix (44). Fractal analysis describes the roughness or smoothness of the texture through the FD measure. Here, single-resolution fractal analysis refers to the estimation of FD of the tumor ROI from 3D parametric maps directly (39).

Evaluation of Predictive Performance for NACT Response

For each of the features obtained from the GLCM, RLM, multiresolution, and single-resolution fractal analysis, the percentage change in the feature values was calculated between the visit-1 and visit-2 DCE-MRI studies. These percentage changes were given as input to support vector machine (45), a robust classifier, to generate a predictive model for classification of pCR versus non-pCR. The performances of the models were evaluated using the ROC AUC, sensitivity and specificity analysis. Sensitivity here refers to the proportion of pCRs correctly identified as pCRs, while specificity refers to the proportion of non-pCRs correctly identified as non-pCRs.

The support vector machine classification performance was evaluated by calculating the average over 10 random partitions of the data for training and testing. For each partition, pCRs and non-pCRs were randomly divided into training and testing data sets as described below. The mean and standard deviation of AUC, sensitivity, and specificity values obtained over the 10 partitions of training and testing data sets are reported. The predictive performance was assessed for the features extracted from each of the 4 parametric maps as well as those constructed by concatenating the texture features from all 4 parametric maps of K^{trans}, k_{ep}, v_e, and τ_i, designated as "All."

The ROC AUC values of the multiresolution fractal features were compared with those of the conventional features by calculating the critical ratio according to the Hanley and McNeil formula (46). The statistical significance was set at $P < .05$.

Table 1. Clinicopathological Characteristics of pCR and non-pCR Groups

	pCR (n = 14)	non-pCR (n = 41)
Age at Diagnosis (years)	27–63	27–79
Tumor Type	14–IDC	34–IDC
		3–ILC
		4–IMC
Tumor Grade		
1	1	4
2	7	16
3	6	21
Tumor Size in Longest Diameter (cm)	1.0–6.9	1.2–12.8
ER		
Positive	2	24
Negative	12	17
PR		
Positive	3	26
Negative	11	15
HER-2		
Positive	12	25
Negative	2	16

Abbreviations: IDC, invasive ductal carcinoma; ILC, invasive lobular carcinoma; IMC, invasive mammary carcinoma.

RESULTS

Among the 55 patients in the study cohort, 14 achieved pCR to NACT, while the other 41 patients were non-pCRs based on pathological analysis of the surgical specimens. Table 1 shows the clinicopathological characteristics of the pCR and non-pCR groups. Nine pCRs/31 non-pCRs and 5 pCRs/10 non-pCRs were

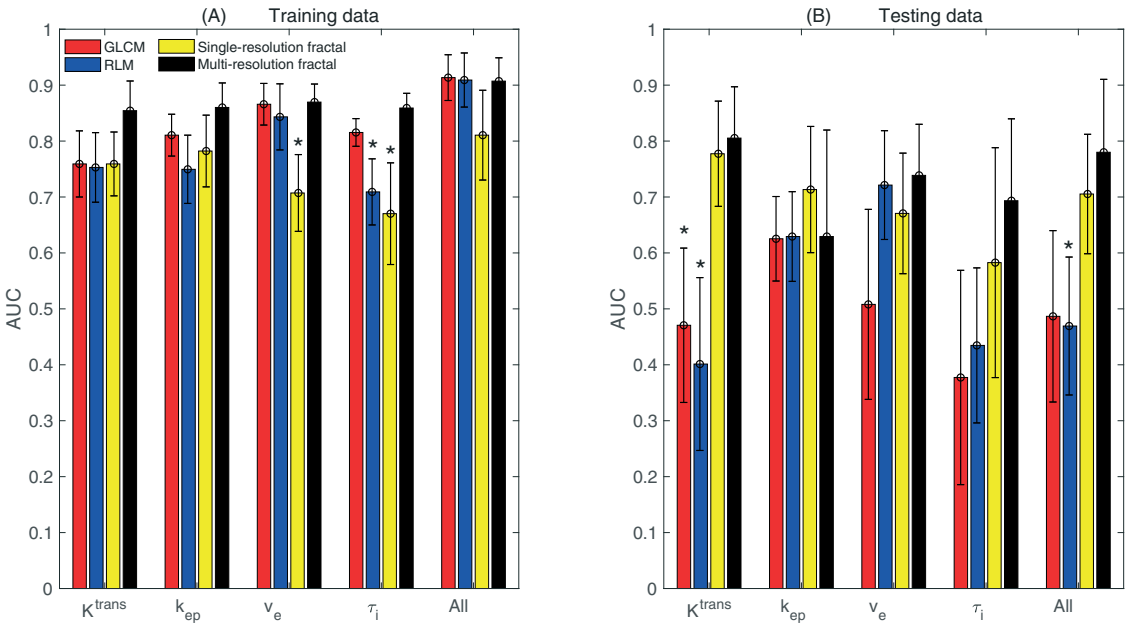


Figure 3. The ROC AUC values for classification of pCR from non-pCR patients in the training (A) and testing (B) data sets using the GLCM (red), RLM (blue), single-resolution fractal (yellow), and multiresolution fractal (black) features extracted from the K^{trans}, k_{ep}, v_e, and τ_i parametric maps. The final column "All" represents the concatenated features from all 4 parametric maps. The error bars represent the standard deviation obtained over the 10 different partitions of train and test data. *: significant ($P < .05$) difference in AUC compared to that of multiresolution fractal features.

selected randomly to form the training and testing sets, respectively. Figure 3 shows the ROC AUC values for classification of pCR versus non-pCR using the GLCM, RLM, single-resolution fractal, and multiresolution fractal features from parametric maps of different DCE-MRI parameters considered individually and the concatenated feature from all 4 parametric maps. GLCM and RLM features seemed to overfit on the training data, as they had high training AUCs but low testing AUCs. For example, the AUC values were 0.76 and 0.75 from the training K^{trans} maps, and 0.47 and 0.40 from the testing K^{trans} maps for the GLCM and RLM methods, respectively. Overall, the multiresolution fractal features from each DCE-MRI parametric map and the concatenated features performed the best in prediction of pCR versus non-pCR in both the training and testing data sets with AUC = 0.85, 0.86, 0.87, 0.86, 0.91 (for K^{trans}, k_{ep}, v_e, τ_i, and All, respectively), and 0.80, 0.63, 0.74, 0.70, 0.78 for the training and testing data sets, respectively. The only exception was from the k_{ep} maps in the testing data sets, where the single-resolution fractal analysis provided the highest AUC of 0.71 among the 4 feature analysis methods. Within the testing or training sets, the predictive performances of multiresolution fractal features were significantly better than the GLCM features from the K^{trans} map (AUC = 0.47, $P = .022$) in the testing set, RLM features from the τ_i map (AUC = 0.71, $P = .012$) in the training set, RLM features from the K^{trans} (AUC = 0.40, $P = .013$), and "All" (AUC = 0.47, $P = .049$) maps in the testing set, and single-resolution fractal features from the v_e (AUC = 0.71, $P = .012$) and τ_i (AUC = 0.67, $P = .037$) maps in the training set.

We evaluated the specificities of the classification models at 2 levels of sensitivities for the testing data sets: 60% (3 out of 5 pCRs were classified correctly) and at 80% (4 out of 5 pCRs were classified correctly), as shown in Table 2. At both sensitivity levels, with a few exceptions, fractal features presented higher specificities than the GLCM and RLM features, with the multi-

Table 2. Specificity Values [Mean (Standard Deviation)] in the Testing Data Set for GLCM, RLM, Single-Resolution Fractal, and Multiresolution Fractal Methods With Sensitivity Set at 60% and 80%

	GLCM	RLM	Single-Resolution Fractal	Multi-Resolution Fractal
	Sensitivity = 60			
K^{trans}	49.3 (22.3)	34.7 (27.5)	84.0 (16.1)	89.3 (11.4)
k_{ep}	73.3 (9.4)	67.3 (17.9)	84.0 (7.2)	70.7 (23.3)
v_e	64.0 (31.6)	82.0 (14.4)	75.3 (7.1)	80.7 (12.4)
τ_i	40.7 (28.4)	44.0 (21.6)	57.3 (27.6)	68.7 (25.0)
All	49.3 (21.6)	42.0 (18.1)	82.0 (15.7)	82.7 (17.0)
	Sensitivity = 80			
K^{trans}	19.3 (16.2)	25.3 (14.7)	63.3 (25.4)	68.7 (13.7)
k_{ep}	49.3 (19.9)	45.3 (19.1)	55.3 (29.3)	49.3 (27.8)
v_e	22.0 (30.0)	67.3 (19.7)	56.0 (20.7)	62.0 (17.2)
τ_i	18.0 (23.1)	30.0 (24.6)	47.3 (24.2)	62.0 (37.7)
All	32.0 (21.5)	28.7 (16.3)	62.0 (26.9)	62.0 (17.8)

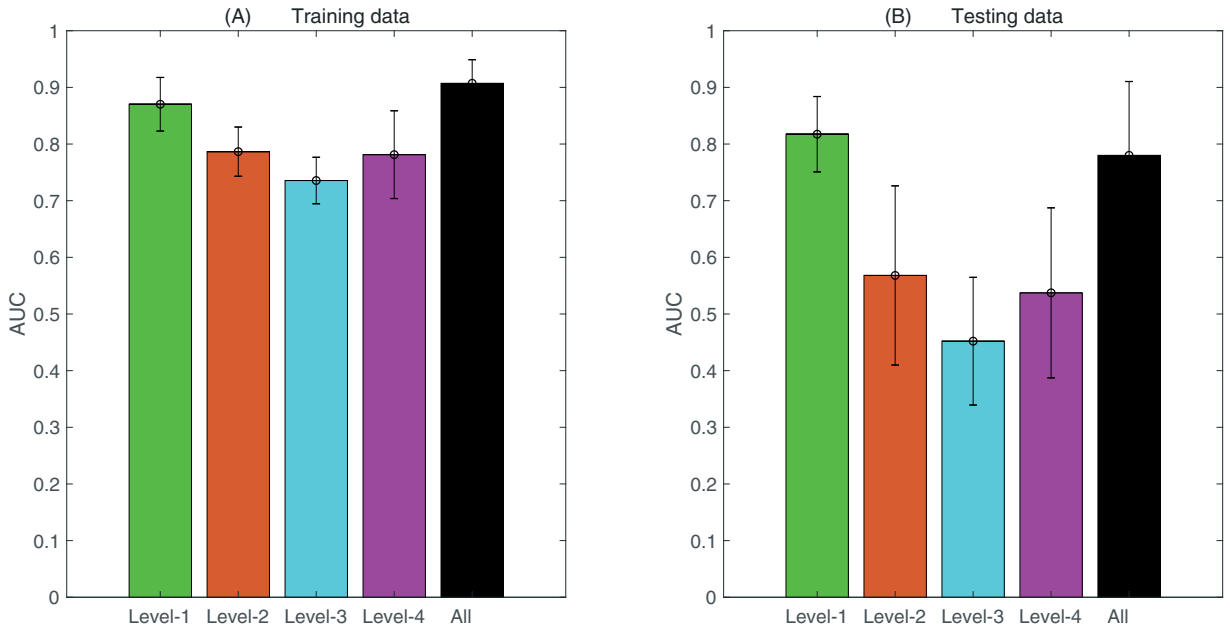


Figure 4. The receiver operating curve (ROC) area under curve (AUC) values for classification of pCR from non-pCR patients in the training (A) and testing (B) data sets using concateneted feature vectors obtained by combining multiresolution fractal features extracted from the K^{trans}, k_{ep}, v_e and τ_i maps. The green, orange, teal, and magenta columns represent the first, second, third, and fourth levels of decompositions, respectively, while the black column corresponds to the combination of features from all 4 levels. The error represents the standard deviation obtained over the 10 different partitions of train and test data.

resolution method generally outperforming the single-resolution method (except when they were applied to the k_{ep} map).

Figure 4 shows the ROC AUC values (for the training and testing data sets) for each level of decomposition of the concatenated feature vectors obtained by combining multiresolution fractal features extracted from the K^{trans}, k_{ep}, v_e, and τ_i maps. The combination of features from all 4 levels and all 4 parametric maps is represented by the black bar "All." It can be observed that the first level of decomposition alone achieved a predictive performance (AUC = 0.87 and 0.82 for the training and testing sets, respectively) comparable to that of the combined features from all levels (AUC = 0.91 and 0.78 for the training and testing sets, respectively), while features from levels 2, 3, and 4 individually had rather poor performances in the test set with AUC = 0.57, 0.45, and 0.54, respectively.

DISCUSSION

This preliminary study shows that multiresolution fractal analysis has the potential to better capture the heterogeneity in the breast tumor vasculature as measured by DCE-MRI, and the extracted features from voxel-based DCE-MRI parametric maps are good early predictors of breast cancer response to NACT. In general, the concatenated features extracted from parametric maps of all the DCE-MRI parameters provide the best predictive performance. Multiresolution analysis filters out irrelevant features and noise at different resolutions, rendering more emphasis on distinct features, and fractal analysis at each level appears to be able to capture these distinct features. The GLCM and RLM features reflect the overall correlation between adjacent voxels in terms of second-order and higher-order statistical features, respectively (27). For the small data set used in this study, the generally higher AUC values from the multiresolution fractal analysis when compared to GLCM and RLM methods suggest that decomposing the texture may give further insights into the heterogeneity of the tumor microvasculature shown on DCE-MRI parametric maps and help capture the subtle variations in the texture which cannot be assessed by the single-resolution approach. However, this observation needs to be validated with a larger patient cohort. Consistent with the studies reporting mean parameter changes (9, 14-22), the results from this study provide further proof that changes in vascular perfusion/permeability represented by DCE-MRI imaging biomarkers are important features in identifying responders and nonresponders at the early stage of NACT.

On inspecting the AUC values from Figure 3, it can be observed that single-resolution fractal features performed consistently well in prediction of response for both the training and testing sets, although not as well as the multiresolution approach. The higher AUCs for fractal-based features suggest that they provide a richer representation of the heterogeneity in the tumor when compared to GLCM and RLM methods. The low dimensionality (d = 1) of single-resolution fractal feature is less likely to cause overfitting for the small sample size of our data set and therefore could lead to a good discriminative model. This could be one of the reasons that contributed to its effectiveness. On the other hand, in spite of increased dimensionality (d = 32),

multiresolution fractal features exhibited better predictive performance, suggesting that analyzing heterogeneity at multiple resolutions provides a more comprehensive measure of the texture and thus increases the discriminative power of the feature. At each level of decomposition, the approximation coefficient [W_φ from equation (1)] represents the low-frequency component, which characterizes the coarse structure of the data, and the detail coefficients [W_ψ^i from equation (2)] represent the high-frequency components, which capture the discontinuities and singularities in the data. Therefore, combination of features from different scales and frequencies gives a richer representation of the overall underlying texture. The advantages of multiresolution fractals can be expected to be even more significant when the data set is large enough to offset their high dimensionality.

The tumor heterogeneity appears to be captured well at the first level of decomposition as shown by Figure 4. Each decomposition level analyzes the signal at a particular band of frequencies. Higher decomposition levels have better frequency resolution. The first level of decomposition encompasses the entire frequency band of the input data in its subvolumes. Thereafter, we select the subvolume with the highest FD and perform multiresolution fractal analysis on that sub-band alone. By doing this we are effectively looking at finer frequency resolutions of the selected sub-band frequencies alone. Considering features from these finer frequency resolutions in isolation do not appear to have as much discriminative power as the first level of decomposition, but combining the finer frequency resolutions features with the features from first level appears to enrich the representation and provide incremental improvement.

As shown in Table 2 for the test data set, at fixed sensitivity, the higher AUC values from the multiresolution fractal features generally resulted in higher specificity values compared to those from other features. It is important to have high sensitivities so that most pCR patients will be correctly identified and continue with the original or de-escalated NACT regimen. At 80% sensitivity, the >60% specificity (except for the k_{ep} features) of the multiresolution fractal features implies that were this method used in clinical care, more than half of the non-pCRs would be correctly classified after the first NACT cycle, potentially enabling alteration of treatment plans for these nonresponders at the early stage of NACT to receive more personalized care.

This study has several limitations. The first being the small size of the data set used. The preliminary results obtained need to be evaluated on a larger patient cohort. Also due to the small size of the data set, dimensionality increase in feature vectors impedes the performance of the classifier. Larger data set can enable the choice of a richer feature vector from different levels in the multiresolution fractal decomposition, which might consistently outperform the other features. Finally, the DCE-MRI parametric maps used for feature analysis were obtained with the shutter-speed model, which is not commonly used in pharmacokinetic analysis of DCE-MRI data. In future studies, parametric maps obtained with the widely used standard Tofts model (47, 48), which generates only the K^{trans}, v_e, and k_{ep} parameters and thus results in reduced dimensionality of the feature vector, will be used for feature extractions and the results will be compared with those presented here.

CONCLUSION

In this preliminary study, we have demonstrated that multiresolution fractal analysis of voxel-based DCE-MRI parametric maps could be a promising tool for early prediction of breast cancer response to NACT. The multiresolution fractal features generally have better predictive performances than those extracted with the more conventional methods of GLCM, RLM, and single-resolution fractal analysis. Furthermore, compared to features extracted from individual DCE-MRI parametric maps, the use of concatenated features from all DCE-MRI parameters generally further improves prediction of NACT response.

ACKNOWLEDGMENTS

This study was supported by National Institutes of Health grant U01-CA154602 and Circle of Giving grant from Oregon Health & Science University Center for Women's Health.

Disclosures: No disclosures to report.

Conflict of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. Smith RA, Andrews KS, Brooks D, Fedewa SA, Manassaram-Baptiste D, Saslow D, Brawley OW, Wender RC. Cancer screening in the United States, 2017: a review of current American Cancer Society guidelines and current issues in cancer screening. Cancer J Clini. 2017;67:100–121.
2. Chatterjee A, Erban JK. Neoadjuvant therapy for treatment of breast cancer: the way forward, or simply a convenient option for patients? Gland Surg. 2017;6:119–124.
3. Kaufmann M, Von Minckwitz G, Mamounas EP, Cameron D, Carey LA, Cristofanilli M, Denkert C, Eiermann W, Gnant M, Harris JR, Karn T. Recommendations from an international consensus conference on the current status and future of neoadjuvant systemic therapy in primary breast cancer. Ann Surg Oncol. 2012; 19:1508–1516.
4. Rastogi P, Anderson SJ, Bear HD, Geyer CE, Kahlenberg MS, Robidoux A, Margolese RG, Hoehn JL, Vogel VG, Dakhil SR, Tamkus D, King KM, Pajon ER, Wright MJ, Robert J, Paik S, Mamounas EP, Wolmark N. Preoperative chemotherapy: updates of National Surgical Adjuvant Breast and Bowel Project Protocols B-18 and B-27. J Clin Oncol. 2008;26:778–785.
5. Bonnefoi H, Litiere S, Piccart M, MacGrogan G, Fumoleau P, Brain E, Petit T, Rouanet P, Jassem J, Moldovan C, Bodmer A, Zaman K, Cufer T, Campone M, Luporsi E, Malmström P, Werutsky G, Bogaerts J, Bergh J, Cameron DA; EORTC 10994/BIG 1-00 Study investigators. Pathological complete response after neoadjuvant chemotherapy is an independent predictive factor irrespective of simplified breast cancer intrinsic subtypes: a landmark and two-step approach analyses from the EORTC 10994/BIG 1-00 phase III trial. Ann Oncol. 2014;25:1128–1136.
6. von Minckwitz G, Untch M, Blohmer JU, Costa SD, Eidtmann H, Fasching PA, Gerber B, Eiermann W, Hilfrich J, Huober J, Jackisch C, Kaufmann M, Konecny GE, Denkert C, Nekljudova V, Mehta K, Loibl S. Definition and impact of pathologic complete response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes. J Clin Oncol. 2012;30:1796–1804.
7. Houssami N, Macaskill P, von Minckwitz G, Marinovich ML, Mamounas E. Meta-analysis of the association of breast cancer subtype and pathologic complete response to neoadjuvant chemotherapy. Eur J Cancer. 2012;48:3342–3354.

8. Zambetti M, Mansutti M, Gomez P, Lluch A, Dittrich C, Zamagni C, Ciruelos E, Pavesi L, Semiglazov V, De Benedictis E, Gaion F, Bari M, Morandi P, Valagussa P, Luca G. Pathological complete response rates following different neoadjuvant chemotherapy regimens for operable breast cancer according to ER status, in two parallel, randomized phase II trials with an adaptive study design (ECTO II). Breast Cancer Res Treat. 2012;132:843–851.
9. Tudorica A, Oh KY, Chui SY, Roy N, Troxell ML, Naik A, Kemmer KA, Chen Y, Holtorf ML, Afzal A, Springer CS, Jr. Early prediction and evaluation of breast cancer response to neoadjuvant chemotherapy using quantitative DCE-MRI. Transl Oncol. 2016;9:8–17.
10. Leach MO, Morgan B, Tofts PS, Buckley DL, Huang W, Horsfield MA, Chenevert TL, Collins DJ, Jackson A, Lomas D, Whitcher B. Imaging vascular function for early stage clinical trials using dynamic contrast-enhanced magnetic resonance imaging. Eur Radiol. 2012;22:1451–1464.
11. Yankeelov TE, Mankoff DA, Schwartz LH, Lieberman FS, Buatti JM, Mountz JM, Erickson BJ, Fennessy FMM, Huang W, Kalpathy-Cramer J, Wahl RL, Linden HM, Kinahan PE, Zhao B, Hylton NM, Gillies RJ, Clarke L, Nordstrom R, Rubin DL. Quantitative imaging in cancer clinical trials. Clin Cancer Res. 2016;22:284–290.
12. Pickles MD, Gibbs P, Lowry M, Turnbull LW. Diffusion changes precede size reduction in neoadjuvant treatment of breast cancer. Magn Reson Imaging. 2006; 24:843–847.
13. Li SP, Padhani AR, Makris A. Dynamic contrast-enhanced magnetic resonance imaging and blood oxygenation level-dependent magnetic resonance imaging for the assessment of changes in tumor biology with treatment. J Natl Cancer Inst Monogr. 2011;2011:103–107.
14. Li X, Kang H, Arlinghaus LR, Abramson RG, Chakravarthy AB, Abramson VG, Farley J, Sanders M, Yankeelov TE. Analyzing spatial heterogeneity in DCE- and DW-MRI parametric maps to optimize prediction of pathologic response to neoadjuvant chemotherapy in breast cancer. Transl Oncol. 2014;7:14–22.
15. Huang W, Li X, Chen Y, Li X, Chang MC, Oborski MJ, Malyarenko DI, Muzi M, Jajamovich GH, Fedorov A, Tudorica A, Gupta SN, Laymon CM, Marro KI, Dyvorne HA, Miller JV, Barbodiak DP, Chenevert TL, Yankeelov TE, Mountz JM, Kinahan PE, Kikinis R, Taouli B, Fennessy F, Kalpathy-Cramer J. Variations of dynamic contrast-enhanced magnetic resonance imaging in evaluation of breast cancer therapy response: a multicenter data analysis challenge. Transl Oncol. 2014;7:153–166.
16. Li X, Arlinghaus LR, Ayers GD, Chakravarthy AB, Abramson RG, Abramson VG, Atuegwu N, Farley J, Mayer IA, Kelley MC, Meszoely IM, Means-Powell J, Grau AM, Sanders M, Bhave SR, Yankeelov TE. DCE-MRI analysis methods for predicting the response of breast cancer to neoadjuvant chemotherapy: pilot study findings. Magn Reson Med. 2014;71:1592–1602.
17. Tateishi U, Miyake M, Nagaoka T, Terauchi T, Kubota K, Kinoshita T, Daisaki H, Macapinlac HA. Neoadjuvant chemotherapy in breast cancer: prediction of pathologic response with PET/CT and dynamic contrast-enhanced MR imaging—prospective assessment. Radiology. 2012;263:53–63.
18. Li SP, Makris A, Beresford MJ, Taylor NJ, Ah-See MLW, Stirling JJ, d'Arcy JA, Collins DJ, Kozarski R, Padhani AR. Use of dynamic contrast-enhanced MR imaging to predict survival in patients with primary breast cancer undergoing neoadjuvant chemotherapy. Radiology. 2011;260:68–78.
19. Jensen LR, Garzon B, Heldahl MG, Bathen TF, Lundgren S, Gribbestad IS. Diffusion-weighted and dynamic contrast-enhanced MRI in evaluation of early treatment effects during neoadjuvant chemotherapy in breast cancer patients. J Magn Reson Imaging. 2011;34:1099–1109.
20. Ah-See MLW, Makris A, Taylor NJ, Harrison M, Richman PI, Burcombe RJ, Stirling JJ, d'Arcy JA, Collins DJ, Pittam MR, Ravichandran D, Padhani AR. Early changes in functional dynamic magnetic resonance imaging predict for pathologic response to neoadjuvant chemotherapy in primary breast cancer. Clin Cancer Res. 2008;14:6580–6589.
21. Yankeelov TE, Lepage M, Chakravarthy A, Broome EE, Niermann KJ, Kelley MC, Meszoely I, Mayer IA, Herman CR, McManus K, Price RR, Gore JC. Integration of quantitative DCE-MRI and ADC mapping to monitor treatment response in human breast cancer: initial results. Magn Reson Imaging. 2007;25:1–13.
22. Padhani AR, Hayes C, Assersohn L, Powles T, Makris A, Suckling J, Leach MO, Husband JE. Prediction of clinicopathologic response of breast cancer to primary chemotherapy at contrast-enhanced MR imaging: initial clinical results. Radiology. 2006;239:361–374.
23. Egeblad M, Nakasone ES, Werb Z. Tumors as organs: complex tissues that interface with the entire organism. Dev Cell. 2010;18:884–901.
24. Teruel JR, Heldahl MG, Goa PE, Pickles M, Lundgren S, Bathen TF, Gibbs P. Dynamic contrast-enhanced MRI texture analysis for pretreatment prediction of clinical and pathological response to neoadjuvant chemotherapy in patients with locally advanced breast cancer. NMR Biomed. 2014;27:887–896.
25. Golden DI, Lipson JA, Telli ML, Ford JM, Rubin DL. Dynamic contrast-enhanced MRI-based biomarkers of therapeutic response in triple-negative breast cancer. J Am Med Inform Assoc. 2013;20:1059–1066.
26. Banerjee I, Malladi S, Lee D, Depeursinge A, Telli M, Lipson J, Golden D, Rubin DL. Assessing treatment response in triple-negative breast cancer from quantitative image analysis in perfusion magnetic resonance imaging. J Med Imaging. 2017; 5:011008.
27. Thibault G, Tudorica A, Afzal A, Chui SY, Naik A, Troxell ML, Kemmer KA, Oh KY, Roy N, Jafarian N, Holtorf ML. DCE-MRI texture features for early prediction of breast cancer therapy response. Tomography. 2017;3:23.
28. Rose CJ, Mills SJ, O'Connor JP, Buonaccorsi GA, Roberts C, Watson Y, Cheung S, Zhao S, Whitcher B, Jackson A, Parker GJ. Quantifying spatial heterogeneity in dynamic contrast-enhanced MRI parameter maps. Magn Reson Med. 2009; 62:488–499.
29. Soares F, Janela F, Pereira M, Seabra J, Freire MM. 3D lacunarity in multifractal analysis of breast tumor lesions in dynamic contrast-enhanced magnetic resonance imaging. IEEE Trans Image Process. 2013;22:4422–4435.
30. Nirouei M, Pouladian M, Abdolmaleki P, Akhlaghpour S. Feature extraction and classification of breast tumors using chaos and fractal analysis on dynamic magnetic resonance imaging. Iranian Red Crescent Med J. 2017; 19:1–9, e41336.
31. Daugman JG. An information-theoretic view of analog representation in striate cortex. In Computational Neuroscience. Cambridge, MA: MIT Press; 1993:403–423.
32. Gonzalez RC, Woods RE. Digital Image Processing. 2nd ed. Upper Saddle River New Jersey: Prentice Hall; 2002.
33. Scheunders P, Livens S, Van de Wouwer G, Vautrot P, Van Dyck D. Wavelet-based texture analysis. Int J Comput Sci Inform Manag. 1998;1:22–34.
34. Tzalavra A, Dalakleidi K, Zacharaki EI, Tsiaparas N, Constantinidis F, Paragios N, Nikita KS. Comparison of multi-resolution analysis patterns for texture classification of breast tumors based on DCE-MRI. In *International Workshop on Machine Learning in Medical Imaging*. Cham: Springer: 2016:296–304.
35. Tzalavra AG, Zacharaki EI, Tsiaparas NN, Constantinidis F, Nikita KS. A multi-resolution analysis framework for breast tumor classification based on DCE-MRI. In *2014 IEEE International Conference on Imaging Systems and Techniques (IST)*. Santorini, Greece: IEEE; 2014:246–250.
36. Braman NM, Etesami M, Prasanna P, Dubchuk C, Gilmore H, Tiwari P, Plecha D, Madabhushi A. Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast DCE-MRI. Breast Cancer Res. 2017;19:57.
37. Al-Kadi OS, Chung DY, Carlisle RC, Coussios CC, Noble JA. Quantification of ultrasonic texture intra-heterogeneity via volumetric stochastic modeling for tissue characterization. Med Image Anal. 2015;21:59–71.
38. Symmans WF, Peintinger F, Hatzis C, Rajan R, Kuerer H, Valero V, Assad L, Poniecka A, Hennessy B, Green M, Buzdar AU. Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. J Clin Oncol. 2007;25:4414–4422.
39. Yankeelov TE, Rooney WD, Li X, Springer CS, Jr. Variation of the relaxographic "shutter-speed" for transcytolemmal water exchange affects the CR bolus-tracking curve shape. Magn Reson Med. 2003;50:1151–1169.
40. Mallat S. A Wavelet Tour of Signal Processing. Amsterdam: Elsevier; 1999 Sep 14.
41. Daubechies I. The wavelet transform, time-frequency localization and signal analysis. IEEE Trans Inform Theory. 1990;36:961–1005.
42. Kontos D, Bakic PR, Carton AK, Troxel AB, Conant EF, Maidment AD. Parenchymal texture analysis in digital breast tomosynthesis for breast cancer risk estimation: a preliminary study. Acad Radiol. 2009;16:283–298.
43. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Trans Syst Man Cybern. 1973;3:610–621.
44. Gallowy MM. Texture analysis using gray level run length. Comput Graph Image Process. 1975;4:172–179.
45. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–297.
46. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology. 1983;148: 839–843.
47. Tofts PS, Kermode AG. Measurement of the blood-brain barrier permeability and leakage space using dynamic MR imaging. Magn Reson Med. 1991;17:357–367.
48. Tofts PS, Brix G, Buckley DL, Evelhoch JL, Henderson E, Knopp MV, Larsson HB, Lee TY, Mayr NA, Parker GJ, Port RE, Taylor J, Weisskoff RM. Estimating kinetic parameters from dynamic contrast-enhanced T1-weighted MRI of a diffusable tracer: standardized quantities and symbols. J Magn Reson Imaging. 1999;10: 223–232.

The Impact of Arterial Input Function Determination Variations on Prostate Dynamic Contrast-Enhanced Magnetic Resonance Imaging Pharmacokinetic Modeling: A Multicenter Data Analysis Challenge, Part II

Wei Huang[1], Yiyi Chen[1], Andriy Fedorov[2], Xia Li[3], Guido H. Jajamovich[4], Dariya I. Malyarenko[5], Madhava P. Aryal[5], Peter S. LaViolette[6], Matthew J. Oborski[7], Finbarr O'Sullivan[8], Richard G. Abramson[9], Kourosh Jafari-Khouzani[10], Aneela Afzal[1], Alina Tudorica[1], Brendan Moloney[1], Sandeep N. Gupta[3], Cecilia Besa[4], Jayashree Kalpathy-Cramer[10], James M. Mountz[7], Charles M. Laymon[7], Mark Muzi[11], Paul E. Kinahan[11], Kathleen Schmainda[6], Yue Cao[5], Thomas L. Chenevert[5], Bachir Taouli[4], Thomas E. Yankeelov[12], Fiona Fennessy[2], and Xin Li[1]

[1]Oregon Health and Science University, Portland, OR; [2]Brigham and Women's Hospital and Harvard Medical School, Boston, MA; [3]General Electric Global Research, Niskayuna, NY; [4]Icahn School of Medicine at Mt Sinai, New York, NY; [5]University of Michigan, Ann Arbor, MI; [6]Medical College of Wisconsin, Milwaukee, WI; [7]University of Pittsburgh, Pittsburgh, PA; [8]University College, Cork, Ireland; [9]Vanderbilt University, Nashville, TN; [10]Massachusetts General Hospital and Harvard Medical School, Boston, MA; [11]University of Washington, Seattle, WA; and [12]The University of Texas, Austin, TX

Corresponding Authors:
Wei Huang, PhD
Oregon Health and Science University, 3181 SW Sam Jackson Park Rd, L452, Portland, OR 97239,
E-mail: huangwe@ohsu.edu; and Xin Li, E-mail: lxin@ohsu.edu

Key Words: DCE-MRI, arterial input function, variation, shutter-speed model, prostate

Abbreviations: Arterial input function (AIF), concordance correlation coefficient (CCC), confidence interval (CI), intra-class correlation coefficient (ICC), volume transfer rate constant (K^{trans}), efflux rate constant (k_{ep}), pre-contrast tissue longitudinal relaxation rate constant (R_{10}), shutter-speed model (SSM), mean intracellular water lifetime (τ_i), extravascular, extracellular volume fraction (v_e), within-subject coefficient of variation (wCV), dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI), contrast agent (CA), region of interest (ROI)

ABSTRACT

This multicenter study evaluated the effect of variations in arterial input function (AIF) determination on pharmacokinetic (PK) analysis of dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) data using the shutter-speed model (SSM). Data acquired from eleven prostate cancer patients were shared among nine centers. Each center used a site-specific method to measure the individual AIF from each data set and submitted the results to the managing center. These AIFs, their reference tissue-adjusted variants, and a literature population-averaged AIF, were used by the managing center to perform SSM PK analysis to estimate K^{trans} (volume transfer rate constant), v_e (extravascular, extracellular volume fraction), k_{ep} (efflux rate constant), and τ_i (mean intracellular water lifetime). All other variables, including the definition of the tumor region of interest and precontrast T_1 values, were kept the same to evaluate parameter variations caused by variations in only the AIF. Considerable PK parameter variations were observed with within-subject coefficient of variation (wCV) values of 0.58, 0.27, 0.42, and 0.24 for K^{trans}, v_e, k_{ep}, and τ_i, respectively, using the unadjusted AIFs. Use of the reference tissue-adjusted AIFs reduced variations in K^{trans} and v_e (wCV = 0.50 and 0.10, respectively), but had smaller effects on k_{ep} and τ_i (wCV = 0.39 and 0.22, respectively). k_{ep} is less sensitive to AIF variation than K^{trans}, suggesting it may be a more robust imaging biomarker of prostate microvasculature. With low sensitivity to AIF uncertainty, the SSM-unique τ_i parameter may have advantages over the conventional PK parameters in a longitudinal study.

INTRODUCTION

As a noninvasive method to measure tissue microvascular perfusion and permeability, dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) is increasingly used in oncologic imaging for cancer diagnosis and therapeutic monitoring (1, 2). DCE-MRI generally involves the serial acquisition of heavily T_1-weighted images before, during, and after the injection of a paramagnetic contrast agent (CA). Quantitative pharmacokinetic (PK) modeling of DCE-MRI time-course data allows estimation of imaging biomarkers, such as K^{trans} (volume transfer

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2018.00027

rate constant) and v_e (extravascular, extracellular volume fraction), that are direct measures of tissue biology and in principle independent of data acquisition details and MRI scanner platform (3). However, the accuracy and precision of the derived PK parameters can be largely affected by the selection of the PK model for data fitting (3-5), errors in quantification of the native tissue T_1 value (3, 4, 6), and variance in determination of arterial input function (AIF; the time-course of CA plasma concentration) (3, 4, 7-9). These challenges lead to substantial variations in the reported PK parameter values for the same disease and are fundamental obstacles in translating quantitative DCE-MRI into multicenter clinical trials and general clinical practice. Therefore, it is important for the DCE-MRI community to investigate the impact of variations/errors in different steps of PK data analysis on the estimated parameter values, establish ways to reduce parameter variance, and identify those parameters that are less sensitive to certain variations in data analysis and, therefore, the more robust imaging biomarkers for multicenter studies.

Quantification of the AIF is generally required in most PK models to fit the DCE-MRI time-course data from the tissue of interest. There are many approaches to determine the AIF, including blinded estimation (10), reference tissue (11-13), empirically derived population-averaged AIF (14), and direct measurement of AIF from a feeding artery if the artery is clearly visible within the image field of view (9). In a previous multicenter data analysis challenge (9) within the Quantitative Imaging Network (QIN) of the National Cancer Institute, we have shown, with shared DCE-MRI data sets from patients with prostate cancer, different extents of PK parameter variations owing to differences in individually measured AIFs using site-specific methods. We have shown that parameter variations could be reduced by using a reference-tissue method (15, 16) to adjust the amplitude of the measured AIF. The commonly used standard Tofts model (17, 18) with two independent fitting parameters (K^{trans} and v_e) was used for PK analysis in that study. A recent single-center prostate DCE-MRI study (19) also shows parameter variations when individually measured AIFs and literature population-average AIFs were used for PK analysis with the Tofts model, resulting in substantial variations in diagnostic accuracy of prostate cancer.

In this study, part II of the QIN multicenter data analysis challenge, the shutter-speed model (SSM) (20, 21) was used to perform PK analysis of the shared data sets with AIFs measured by multiple QIN centers. The main difference between the SSM and the Tofts model is that the former takes into account inter-tissue-compartment water-exchange kinetics. An additional parameter, the mean intracellular water lifetime (τ_i), is used in the SSM to account for the transcytolemmal (cross cell membrane) water-exchange kinetics. Recent studies show that the SSM-derived K^{trans} parameter is a more accurate diagnostic marker for both breast (22, 23) and prostate cancer (24), and pretreatment τ_i is predictive of breast cancer response to neoadjuvant chemotherapy (25) and overall survival in patients with head and neck cancer (26). Furthermore, recent results suggest that τ_i is potentially a new imaging biomarker of cellular metabolic activity (27-31), specifically the activity of the Na^+-K^+-ATPase pump, which is essential for all mammalian cells and is primarily responsible for maintaining the K^+ and Na^+ gradient in vivo. In addition, a simulation study (16) has shown low sensitivity of τ_i to AIF amplitude scaling compared with other conventional PK parameters such as K^{trans}. Thus, it is important to experimentally investigate the effect of uncertainty in AIF determination on parameters estimated with the SSM, which was the goal of this study.

MATERIALS AND METHODS

Data Sharing and Multicenter AIF Measurement

Axial prostate DCE-MRI data were collected by one QIN center (32) for pretreatment staging of patients with prostate cancer. Data sets from 11 patients were shared with other QIN centers through TCIA (The Cancer Imaging Archive). These data sets were acquired at 3 T using a 3-dimensional SPoiled Gradient Recalled (SPGR) sequence with repetition time = 3.6 milliseconds, echo time = 1.3 milliseconds, flip angle = 15°, a temporal resolution ranging from 4.4 to 5.3 seconds, and about 60 frames for a 4.5- to 6-minute acquisition time. Nine QIN centers, denoted as QIN1 to QIN9, downloaded the DCE-MRI data and performed AIF measurement from a single image slice for each individual data set using site-specific methods. The smaller circular region of interest (ROI) placed in the left femoral artery (Figure 1A insert) shows the most common location where the AIFs were measured. The derived AIFs in the form of signal intensity time-course data were then submitted to one of the 9 centers, the data managing center, for centralized PK analysis of the 11 DCE-MRI data sets. Additional details on DCE-MRI acquisition parameters and the methods used by each center for AIF measurement from the imaging data are described in Huang et al.'s study (9).

DCE-MRI Data Analysis

The AIF signal intensity time-course was converted by the managing center to blood R_1 ($\equiv 1/T_1$) time-course, $R_{1,b}(t)$, using the steady-state MRI signal intensity equation for a gradient pulse sequence (33) with the known acquisition parameters of flip angle, echo time, and repetition time, and a fixed precontrast blood R_1 of 0.61 s^{-1} (34), and then to plasma CA concentration time-course, $C_p(t)$, using the following equation:

$$R_{1,b}(t) = r_1 h C_P(t) + 0.61\ s^{-1} \qquad (1)$$

where r_1 is the CA relaxivity at 3 T, set at 3.8 $mM^{-1}s^{-1}$; h is the hematocrit, set at 0.45.

For comparison with the individually measured AIFs, a frequently cited and used population-averaged AIF published by Geoff Parker (GP) et al. (14) was also included in this study. The analytical expression of the GP AIF was implemented at the managing center and resampled to match the temporal features of the prostate DCE-MRI data sets.

For each data set, the prostate tumor ROI was defined on a single image slice through the central portion of the tumor by one investigator from the center where the data were generated. The signal intensity time-course for each voxel within the tumor ROI was converted by the managing center to R_1 time-course, $R_1(t)$, in the same way as for $R_{1,b}(t)$, but with a fixed precontrast R_1 for the tumor tissue, R_{10}, assumed to be 0.63 s^{-1} (7). Follow-

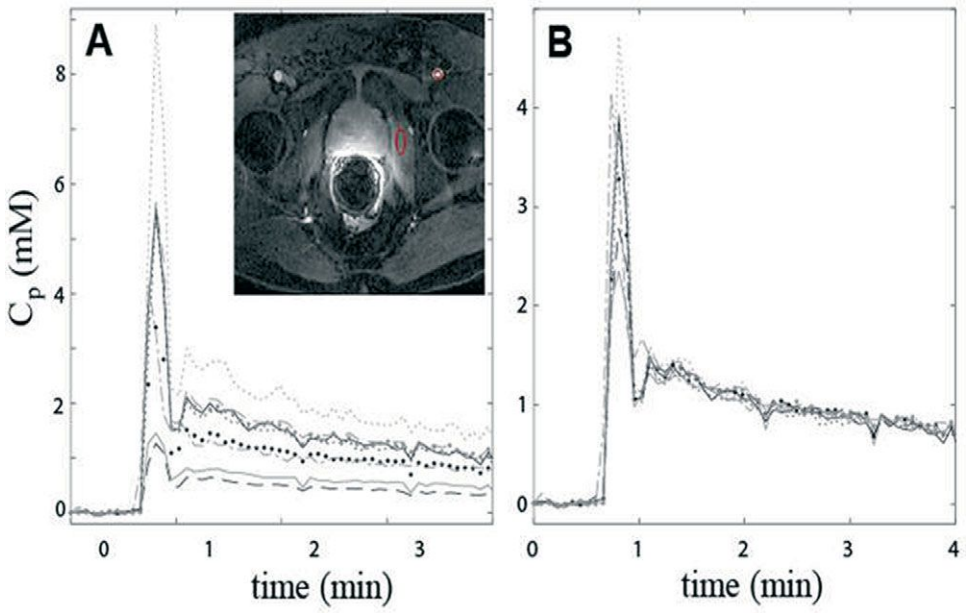


Figure 1. Individual arterial input functions (AIFs) measured from one subject's prostate dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) data set by 9 Quantitative Imaging Network (QIN) centers. The insert in (A) is a zoomed axial postcontrast DCE-MRI image slice showing the smaller red, circular region of interest (ROI) in the left femoral artery where the blood signals were measured for the AIF time-courses, and the larger red, ellipsoidal reference ROI in the normal-appearing obturator muscle adjacent to the prostate. Substantial variations in both the shape and magnitude can be observed in the AIF curves determined by the 9 QIN centers (A), which are clearly reduced following magnitude adjustment using the reference tissue method (B).

ing calculation of $C_p(t)$ [equation (1)] for each of the AIFs measured by the 9 QIN centers and the literature GP AIF, and $R_1(t)$ for each tumor voxel, the managing center performed PK analysis of the shared 11 prostate DCE-MRI data sets on a voxel-by-voxel basis using an in-house Python-based SSM software package. All AIF arrival times were manually aligned with the uptake phase of the average tissue response curves from the tumor ROIs. The 2-compartment–3-parameter version of the SSM (20, 21) was used for $R_1(t)$ data fitting in this study:

$$R_1(t) = (1/2)[\{2R_{1i} + r_1K^{trans}/v_e\int_0^t C_p(t')\exp((-K^{trans}/v_e) \times (t-t'))dt' + (R_{10} - R_{1i} + 1/\tau_i)/v_e\} - \{[2/\tau_i + (R_{1i} - R_{10} - 1/\tau_i)/v_e - r_1K^{trans}/-v_e\int_0^t C_p(t')\exp \times ((-K^{trans}/v_e)(t-t'))dt')]^2 + 4(1-v_e)/\tau_i^2v_e\}^{1/2}] \quad (2)$$

where R_{1i} is the intrinsic intracellular longitudinal relaxation rate constant and is assumed to be equal to the tissue R_{10}. The PK model fitting returned K^{trans}, v_e, and τ_i parameter values for each voxel within the tumor ROI, and the CA efflux rate constant, k_{ep}, was calculated as $k_{ep} = K^{trans}/v_e$. The mean parameter values of the single-slice tumor ROI were obtained by averaging the voxel parameter values within the ROI.

Owing to large differences in the site-specific methods for the AIF measurement (9), such as the placement of the ROI in the artery and the ROI size, substantial variations in AIF amplitude were observed in the AIFs measured from the same data set. A reference tissue method (15, 24) was used to adjust the amplitude of the measured AIFs, as well as the literature GP AIF, in an attempt to reduce the variations (9). In this approach, an ellipsoidal ROI (Figure 1A insert) was drawn in the adjacent, normal-appearing obturator muscle area on the same image slice as the one for the AIF measurement and used as the reference tissue ROI. The AIF amplitude was varied until the Tofts model fitting of the DCE-MRI data from the muscle reference tissue ROI returned a v_e value of 0.1 (35). In total, 20 AIFs, including unadjusted and reference tissue-adjusted AIFs measured by the 9 QIN centers and of the literature GP AIF, were used for PK modeling of each prostate DCE-MRI data set using the SSM, resulting in 20 sets of mean tumor K^{trans}, v_e, k_{ep}, and τ_i values that were then separated into two groups of results based on the unadjusted and reference tissue-adjusted AIF approaches.

Because a physically meaningful v_e is in the range of 0.0 to 1.0, these two values were used as the lower and upper boundaries, respectively, for SSM fitting of all voxel data. All returned voxel v_e values were within the two boundaries (none at boundary values) when the reference tissue-adjusted AIFs were used, while, on average, <3% voxels (range: 0%–6.6% for all the AIF and data set combinations) had returned v_e values reaching the upper boundary of 1.0 when the unadjusted AIFs were used. In the latter case, the parameter values from these limited number of voxels with v_e value of 1.0 were not excluded from the calculation of tumor mean parameter values.

Statistical Analysis

The mean parameter values for the tumor ROI obtained from all fittings were used for statistical analysis. Descriptive statistical analysis was conducted to summarize the PK parameter values returned using different AIFs, with the distribution graphically assessed by boxplots. Intraclass correlation coefficients (ICC), within-subject coefficient of variation (wCV), and concordance correlation coefficients (CCC) were calculated, and these were reported with the corresponding 95% confidence intervals (CIs). Although all three coefficients were computed to assess the reproducibility of the PK parameter values obtained with different AIFs, each had specific focus. The ICC measures the proportion of total variation contributed by between-subject differences, with a high ICC value indicating good agreement (36). The wCV is the ratio of within-subject standard deviation to the mean of a parameter, with smaller wCV value suggesting better reproducibility. Closely related to ICC, CCC represents the level of pairwise linear agreement to a 45° line of which the intercept is forced to be zero. A larger CCC indicates better agreement between results from a pair of measurements and thus better reproducibility. Bland–Altman plots were used to graphically demonstrate pairwise agreements in results from different AIF measurements. SAS 9.4 (Cary, NY) was used for all statistical analysis. SAS macro "%ICC9" and "%mccc" were used for the estimations of ICC, wCV, and CCC.

RESULTS

Variations in AIF Determination

For each data set, substantial variations in both the amplitude and shape of the $C_p(t)$ time-course can be observed as a result of direct AIF measurement from the DCE-MRI data by the 9 QIN centers using site-specific methods. A clear example of $C_p(t)$ variation is shown in Figure 1A. Following amplitude adjustment of $C_p(t)$ using the reference tissue (Figure 1A insert), the agreement among the individually measured AIF curves was clearly improved (Figure 1B). Table 1 lists the standard deviation (SD) of the $C_p(t)$ peak amplitude for unadjusted and reference tissue-adjusted AIFs from measurements by the 9 centers for each patient. Two-tailed paired *t* test shows that the AIF peak value SD of the reference tissue-adjusted AIFs is significantly ($P = .018$) smaller than that of the unadjusted AIFs.

Table 1. Standard Deviation of AIF Peak from Multicenter Measurements

Patient	SD of AIF Peak Value (mM)	
	Unadj. AIF	Adj. AIF[a]
1	0.88	0.54
2	2.36	0.72
3	4.74	1.98
4	0.75	0.65
5	0.55	0.32
6	0.68	0.32
7	0.55	0.76
8	1.63	0.64
9	0.41	0.42
10	1.28	0.56
11	4.45	2.27

[a] Standard deviation (SD) of AIF peak value is significantly smaller for reference tissue-adjusted (Adj.) AIFs in comparison with unadjusted (Unadj.) AIFs: 2-tailed paired *t* test, $P = .018$.

PK Parameter Variations Due to AIF Differences

Figure 2 shows the boxplots of K^{trans}, v_e, k_{ep}, and τ_i parameters estimated from SSM modeling of the 11 DCE-MRI data sets with adjusted and unadjusted AIFs (including those from the GP AIF). For most measurements, the mean is greater than the median, which is commonly seen when distributions are skewed toward larger parameter values. The dispersion of the estimated parameter values from the 11 patients varies substantially across the QIN centers (or AIFs), with K^{trans} showing clearly the largest variation, while v_e and τ_i exhibiting the least variations. As another marker of microvascular properties, k_{ep} shows less variation than K^{trans}. Comparing the boxplots between unadjusted and adjusted AIFs, it can be visually observed that the agreement in parameter dispersion among different centers (or AIFs) is improved for K^{trans} and v_e when the reference tissue-adjusted AIFs were used in data fitting, but this is not clearly the case for k_{ep} and τ_i. Similar observations can be obtained from Table 2, which shows the mean SSM parameter values and 95% CIs for each patient under the unadjusted and reference tissue-adjusted AIF approaches. The mean values were calculated by averaging the tumor parameter values derived with the individual AIFs determined by the 9 QIN centers.

Figure 3 shows a column graph of wCV for K^{trans}, v_e, k_{ep}, and τ_i obtained with the unadjusted (gray) and adjusted (white) AIFs. The error bars represent the 95% CIs. The larger the wCV value, the higher the variation in a measurement performed on the same subject by different methods. The wCV values for K^{trans}, v_e, k_{ep}, and τ_i are 0.58, 0.27, 0.42, and 0.24 for unadjusted AIFs, and 0.50, 0.10, 0.39, and 0.22 for adjusted AIFs, respectively. The wCV of K^{trans} is the largest among all 4 parameters with either unadjusted or adjusted AIFs, while those of v_e and τ_i are the smallest. From unadjusted to adjusted AIFs, the decrease in parameter variation is more prominent for K^{trans} and v_e (wCV value decreases from 0.58 to 0.50 and from 0.27 to 0.10, respectively), compared with k_{ep} and τ_i (0.42 to 0.39 and 0.24 to 0.22, respectively). Figure 4 shows a similar graph of ICC values for K^{trans}, v_e, k_{ep}, and τ_i obtained with the two AIF approaches. The ICC values for K^{trans}, v_e, k_{ep}, and τ_i are 0.44, 0.51, 0.72, and 0.92 for unadjusted AIFs, respectively, and 0.59, 0.91, 0.79, and 0.93 for adjusted AIFs, respectively. Consistent with the results shown in Figure 3 , K^{trans} has the smallest ICC value with either AIF approach, while τ_i has the largest ICC value. From unadjusted to adjusted AIFs, the increase in ICC is the most obvious for K^{trans} and v_e (ICC value increases from 0.44 to 0.59 and from 0.51 to 0.91, respectively) compared with k_{ep} and τ_i (0.72 to 0.79 and 0.92 to 0.93, respectively).

As an example of differences in AIF-caused variations in estimated PK parameters when unadjusted and reference tissue-adjusted AIFs were used for SSM analysis, Figure 5 shows voxel-based parametric maps of K^{trans} and τ_i of a prostate tumor generated from the SSM analysis. The tumor ROI was in the

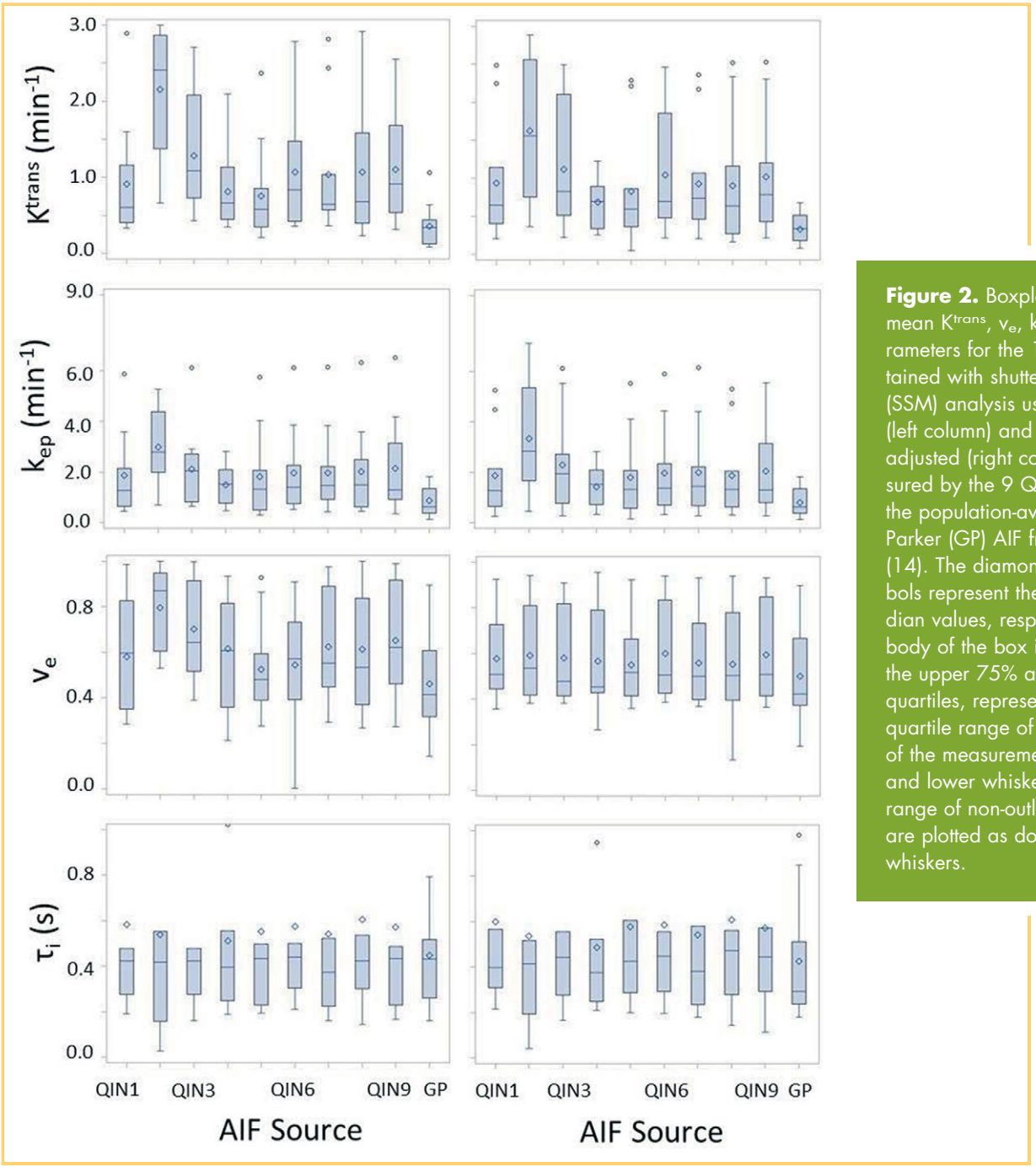


Figure 2. Boxplots of the tumor mean K^{trans}, v_e, k_{ep}, and τ_i parameters for the 11 subjects obtained with shutter-speed model (SSM) analysis using unadjusted (left column) and reference tissue-adjusted (right column) AIFs measured by the 9 QIN centers and the population-averaged Geoff Parker (GP) AIF from the literature (14). The diamond and bar symbols represent the mean and median values, respectively. The body of the box is bounded by the upper 75% and lower 25% quartiles, representing the interquartile range of the middle 50% of the measurements. The upper and lower whiskers define the range of non-outliers. The outliers are plotted as dots beyond the whiskers.

peripheral zone, as indicated by the arrow in the postcontrast DCE-MRI image. K^{trans} and τ_i maps obtained with unadjusted AIFs from the 9 QIN centers are shown on the left panels and those with reference tissue-adjusted AIFs are shown on the right. These maps are displayed under the same K^{trans} and τ_i color scales, respectively. With either AIF approach, substantially higher variations among the 9 K^{trans} maps can be visually observed compared with the 9 τ_i maps. While the variations among the K^{trans} maps can be seen reduced when the reference tissue-adjusted AIFs were used, there is no noticeable improvement in agreement among the τ_i maps going from unadjusted to adjusted AIFs. It is interesting to note, however, that despite considerable variations in K^{trans} maps owing to AIF differences, the spatial pattern of voxel K^{trans} distribution largely remains the same in all the maps. This was also observed in the τ_i maps, and in the maps of v_e and k_{ep} (data not shown for the latter two parameters).

Concordance Analysis

Concordance correlation analysis was conducted to assess parameter agreement between any two AIFs under the same condition (adjusted or unadjusted). Tables 3 and 4 tabulate the CCC values for K^{trans} and τ_i, respectively, with those for the unadjusted AIFs listed in the top right half and those for the adjusted AIFs in the lower left half. The CCC ranges for K^{trans} and τ_i are 0.005–0.937 and 0.558–0.993, respectively, for unadjusted AIFs, and 0.102–0.991 and 0.640–0.997, respectively, for ad-

Table 2. Mean and 95% Confidence Interval of the SSM PK Parameters Obtained with Unadjusted and Reference-Tissue-Adjusted AIFs

	Unadj. AIF				Adj. AIF			
Patient	K^{trans} (min^{-1})	v_e	k_{ep} (min^{-1})	τ_i (s)	K^{trans} (min^{-1})	v_e	k_{ep} (min^{-1})	τ_i (s)
1	0.52 (0.26, 0.77)	0.65 (0.55, 0.76)	0.80 (0.60, 1.00)	0.38 (0.32, 0.44)	0.35 (0.26, 0.43)	0.48 (0.46, 0.51)	0.75 (0.60, 0.90)	0.31 (0.28, 0.34)
2	0.99 (0.45, 1.52)	0.41 (0.25, 0.57)	2.26 (1.92, 2.61)	0.20 (0.14, 0.27)	0.94 (0.79, 1.08)	0.41 (0.39, 0.43)	2.27 (1.92, 2.60)	0.19 (0.16, 0.22)
3	1.89 (1.32, 2.46)	0.35 (0.26, 0.43)	5.58 (4.55, 6.61)	0.34 (0.24, 0.44)	2.29 (1.88, 2.65)	0.44 (0.41, 0.47)	5.44 (4.59, 6.49)	0.36 (0.25, 0.46)
4	2.67 (2.45, 2.89)	0.74 (0.65, 0.83)	3.73 (3.26, 4.20)	0.40 (0.29, 0.50)	2.15 (1.88, 2.42)	0.49 (0.46, 0.53)	4.47 (3.87, 5.07)	0.34 (0.24, 0.44)
5	0.60 (0.43, 0.77)	0.44 (0.39, 0.49)	1.45 (1.20, 1.71)	0.60 (0.42, 0.83)	0.48 (0.38, 0.58)	0.36 (0.34, 0.40)	1.44 (1.21, 1.68)	0.58 (0.41, 0.82)
6	1.21 (0.77, 1.65)	0.92 (0.89, 0.96)	1.26 (0.84, 1.68)	0.42 (0.37, 0.48)	1.06 (0.83, 1.28)	0.93 (0.92, 0.94)	1.11 (0.87, 1.34)	0.46 (0.44, 0.48)
7	0.44 (0.33, 0.54)	0.84 (0.60, 0.99)	0.48 (0.38, 0.58)	0.41 (0.36, 0.46)	0.22 (0.16, 0.28)	0.81 (0.76, 0.85)	0.29 (0.23, 0.36)	0.56 (0.54, 0.58)
8	0.68 (0.32, 1.04)	0.78 (0.62, 0.94)	0.82 (0.47, 1.18)	0.42 (0.37, 0.47)	0.63 (0.35, 0.91)	0.79 (0.74, 0.84)	0.79 (0.46, 1.12)	0.41 (0.39, 0.44)
9	1.10 (0.75, 1.45)	0.63 (0.58, 0.67)	2.01 (1.50, 2.52)	1.25 (1.18, 1.32)	0.80 (0.59, 1.02)	0.51 (0.49, 0.53)	1.97 (1.42, 2.51)	1.21 (1.13, 1.29)
10	1.25 (0.59, 1.91)	0.62 (0.49, 0.76)	1.87 (1.29, 2.44)	0.35 (0.24, 0.46)	1.28 (0.75, 1.82)	0.71 (0.66, 0.76)	1.81 (1.14, 2.40)	0.39 (0.32, 0.46)
11	1.13 (0.55, 1.71)	0.52 (0.42, 0.61)	2.13 (1.45, 2.81)	0.23 (0.20, 0.26)	0.90 (0.31, 1.40)	0.37 (0.30, 0.45)	2.36 (1.56, 3.25)	0.22 (0.19, 0.24)

The values in the parenthesis represent the lower and upper bounds of the 95% confidence interval.

justed AIFs. Reflective of the results shown in Figures 3 and 4, there is generally a considerable increase (comparing values that are symmetric to the diagonal line in Table 3) in the CCC value for pair-wise comparisons of the K^{trans} parameter going from unadjusted to adjusted AIFs, while little CCC changes are observed (comparing values that are symmetric to the diagonal line in Table 4) for the τ_i parameter. The CCC ranges for v_e and k_{ep} (tables not shown) are 0.334–0.986 and 0.145–0.957, respectively, for unadjusted AIFs, and 0.554–0.993 and 0.129–0.965, respectively, for adjusted AIFs. From unadjusted to adjusted AIFs, the changes in CCC for v_e and k_{ep} are similar to those for K^{trans} and τ_i, respectively. In addition, it is important to note that with either AIF approach, the CCC values for pair-wise comparisons that included the GP AIF are among the smallest values in the aforementioned CCC ranges.

Bland–Altman plots are shown in Figure 6 to show examples of pair-wise agreements in K^{trans} (Figure 6A) and τ_i (Figure 6B). The plots are displayed only for the AIF pairs with the largest (top rows in Figure 6, A and B) and smallest (bottom rows in Figure 6, A and B) CCC values for the unadjusted (left columns) and reference tissue-adjusted (right columns) AIFs. Although the differences between the measurements are mostly within the 95% CIs for all the plots, it is clear, with the vertical axis scales kept the same for the K^{trans} and τ_i plots, respectively, that the width of the CI band differs substantially between AIF pairs with greater CCC values and those with smaller CCC values: narrower for the former and wider for the latter. For K^{trans} and τ_i with the largest CCC values (ie, the best pair-wise agreements in the estimated K^{trans} and τ_i values), the means of parameter difference represented by the dotted lines are 0.22

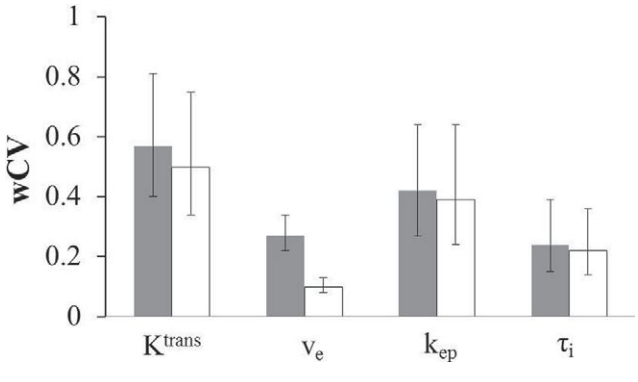


Figure 3. Column graphs of within-subject coefficient of variation (wCV) for the SSM K^{trans}, v_e, k_{ep}, and τ_i parameters obtained with the unadjusted (gray) and adjusted (white) AIFs. The respective 95% confidence intervals (CI) are shown as error bars.

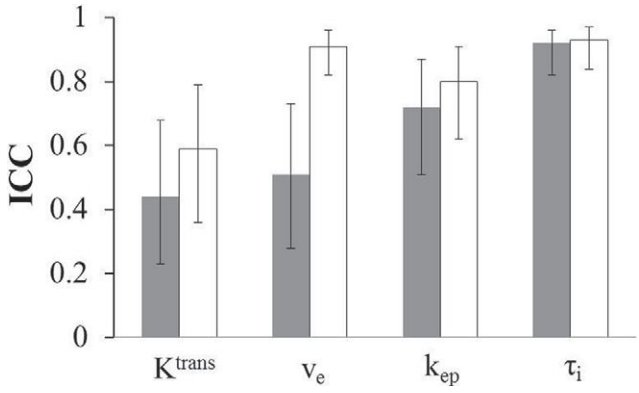


Figure 4. Column graphs of intraclass correlation coefficient (ICC) for the SSM K^{trans}, v_e, k_{ep}, and τ_i parameters obtained with the unadjusted (gray) and adjusted (white) AIFs. The respective 95% CIs are shown as error bars.

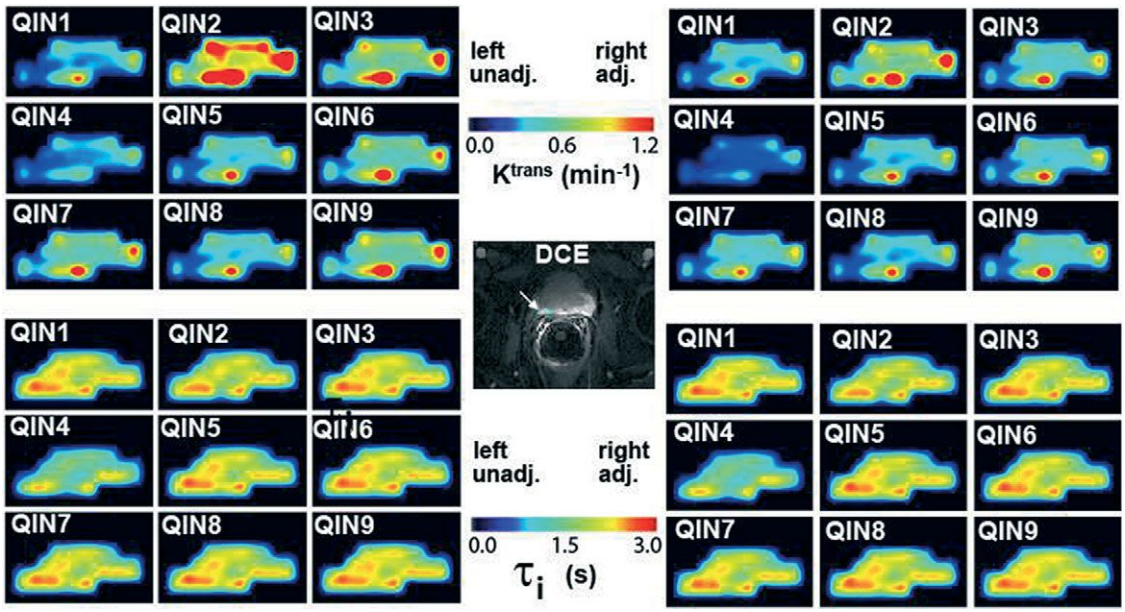


Figure 5. Voxel-based K^{trans} (top two panels) and τ_i (bottom two panels) parametric maps in a prostate tumor ROI, with each panel consisting of 9 maps corresponding to those obtained with AIFs measured by 9 QIN centers. The left and right two panels show the maps obtained with unadjusted and adjusted AIFs, respectively. The grayscale image at the center shows an axial postcontrast DCE-MRI image slice, with the arrow pointing to the cyan-colored prostate tumor ROI. The color scales of K^{trans} and τ_i are kept the same, respectively, for the unadjusted and adjusted AIF approaches.

min^{-1} and 0.012 seconds, respectively, for unadjusted AIFs, and 0.078 min^{-1} and 0.005 seconds, respectively, for adjusted AIFs. For K^{trans} and τ_i with the smallest CCC values (ie, the worst pair-wise agreements in the estimated K^{trans} and τ_i values), the means of parameter difference represented by the dotted lines are -0.56 min^{-1} and -0.18 seconds, respectively, for unadjusted AIFs, and -1.29 min^{-1} and -0.18 seconds, respectively, for adjusted AIFs. From unadjusted to adjusted AIFs, the decrease in the width of the 95% CI band is substantially greater for the K^{trans} parameter than that for the τ_i parameter; the average percent decrease (from the cases with the largest and smallest CCCs) is 37% for K^{trans} and 15% for τ_i. This indicates that the use of reference tissue-adjusted AIF has a stronger effect in improving parameter agreement in K^{trans} compared with τ_i. The same observation was made when comparing K^{trans} and k_{ep} (data not shown). In addition, in cases of poor K^{trans} agreement (bottom row of Figure 6A), there appears to be a correlation (linear bias) between the difference in K^{trans} and the mean of

Table 3. CCC Values for K^{trans}

	QIN1	QIN2	QIN3	QIN4	QIN5	QIN6	QIN7	QIN8	QIN9	GP
QIN1		0.239	0.702	0.683	0.914	0.846	0.921	0.838	0.790	0.005
QIN2	0.406		0.464	0.188	0.197	0.358	0.317	0.325	0.337	0.084
QIN3	0.836	0.642		0.440	0.666	0.937	0.825	0.857	0.639	0.159
QIN4	0.462	0.277	0.498		0.669	0.565	0.541	0.600	0.581	0.182
QIN5	0.960	0.409	0.840	0.643		0.820	0.864	0.747	0.718	0.089
QIN6	0.881	0.586	0.991	0.548	0.880		0.937	0.886	0.685	0.144
QIN7	0.990	0.447	0.862	0.562	0.969	0.906		0.780	0.800	0.045
QIN8	0.975	0.372	0.864	0.682	0.942	0.911	0.961		0.595	0.148
QIN9	0.977	0.488	0.866	0.620	0.938	0.895	0.981	0.931		0.057
GP	0.191	0.102	0.162	0.348	0.224	0.173	0.209	0.159	0.194	

CCC values for unadjusted (unadj.) AIFs are presented in the top right triangle and those for reference-tissue-adjusted (adj.) AIFs are presented in the bottom left triangle.

Table 4. CCC Values for τ_i

	QIN1	QIN2	QIN3	QIN4	QIN5	QIN6	QIN7	QIN8	QIN9	GP
QIN1		0.858	0.937	0.821	0.947	0.977	0.972	0.933	0.953	0.583
QIN2	0.920		0.935	0.835	0.895	0.855	0.869	0.881	0.882	0.577
QIN3	0.945	0.976		0.849	0.974	0.899	0.908	0.949	0.920	0.594
QIN4	0.803	0.845	0.859		0.860	0.842	0.864	0.840	0.872	0.773
QIN5	0.938	0.955	0.995	0.849		0.922	0.925	0.964	0.941	0.600
QIN6	0.997	0.906	0.938	0.806	0.937		0.993	0.949	0.973	0.619
QIN7	0.989	0.920	0.945	0.844	0.941	0.990		0.943	0.965	0.662
QIN8	0.974	0.916	0.960	0.815	0.957	0.971	0.965		0.954	0.617
QIN9	0.978	0.929	0.965	0.830	0.962	0.979	0.971	0.992		0.558
GP	0.702	0.640	0.658	0.840	0.653	0.714	0.764	0.675	0.677	

CCC values for unadjusted (unadj.) AIFs are presented in the top right triangle and those for reference-tissue-adjusted (adj.) AIFs are presented in the bottom left triangle.

K^{trans}; the larger the parameter value, the larger the difference in the parameter value between the two measurements. No clear correlation is observed for τ_i, even in cases of poor agreement (bottom row of Figure 6B).

DISCUSSION

In this part II of a multicenter data analysis challenge to evaluate the effect of variations in AIF determination on estimated PK parameters from prostate DCE-MRI data, the SSM was used for PK modeling of the DCE-MRI data. All other aspects in the data analysis were kept the same as those in part I (9) of the challenge where the standard two-parameter (K^{trans} and v_e) Tofts model was used. For example, quality control measures such as fixed tumor ROI definition, fixed tumor T_{10}, and central data analysis with a single SSM software package were adopted to ensure that PK parameter variations are mainly due to variations in only AIF. Compared with challenge part I (9), where the effect of AIF uncertainty was evaluated on parameters of K^{trans}, v_e, and k_{ep}, one additional parameter, τ_i, was included in this part II study.

Consistent with results from challenge part I (9), substantial variations in the estimated PK parameters were observed in this study owing to variations in AIF quantification by 9 QIN centers using site-specific methods (9), especially in K^{trans} and k_{ep}. Among the four parameters derived with the SSM using unadjusted AIFs, K^{trans} shows the largest AIF-caused variation with a wCV value of 0.58, while v_e and τ_i show the smallest variations

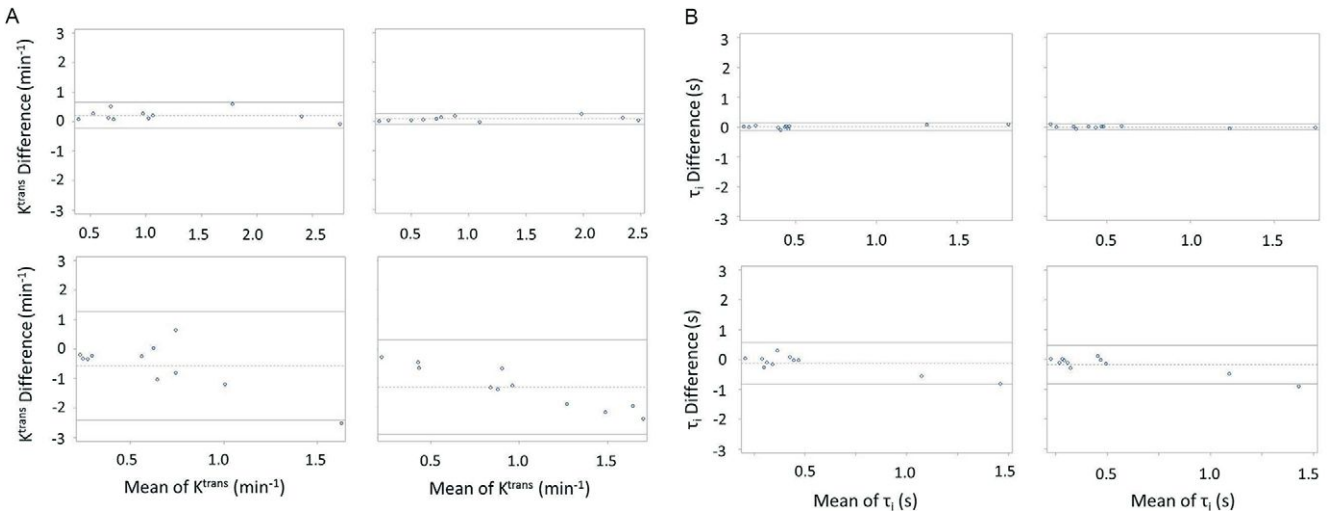


Figure 6. Bland–Altman plots showing agreements in K^{trans} (A) and τ_i (B) for AIF pairs with the largest (top row in A and B) and smallest (bottom row in A and B) CCC values under the conditions of unadjusted (left column in A and B) and adjusted (right column in A and B) AIFs. The two solid horizontal lines represent the upper and lower limits of the 95% CI, while the dotted horizontal line represents the mean value of K^{trans} (A) and τ_i (B) difference between the paired measurements.

with nearly equal wCV values of 0.27 and 0.24, respectively. Although higher than v_e and τ_i, k_{ep} has a lower AIF-caused variation than K^{trans}, with a wCV value of 0.42. Our findings are in agreement with a recent study comparing fully automated and semiautomated AIF determination approaches for prostate DCE-MRI data analysis (7), showing that K^{trans} variation owing to AIF uncertainty is the most prominent compared with other PK parameters. A similar conclusion was drawn in a brain DCE-MRI study (8) that investigated PK parameter variations caused by the use of AIFs measured from different vessels.

As shown by this study using the SSM, as well as part I of the challenge (9) using the standard Tofts model, adjusting the amplitudes of individually measured AIFs with a reference-tissue method (15, 24) by placing the reference ROI in the adjacent normal muscle region can decrease AIF variance (Table 1) and, as a result, reduce parameter variations. For example, the wCV values were decreased from 0.58 to 0.50 and from 0.27 to 0.10 for K^{trans} and v_e, respectively, when the reference tissue-adjusted AIFs replaced the unadjusted AIFs in the SSM analysis. The effect of AIF amplitude adjustment was smaller, however, on k_{ep} (wCV: 0.42 to 0.39) and τ_i (wCV: 0.24 to 0.22) parameters. These observations are consistent with the results from a simulation study using the SSM (16), which found significantly lower sensitivity of k_{ep} and τ_i to a 30% change in AIF amplitude compared with K^{trans} and v_e. Interestingly, the aforementioned brain DCE-MRI study (8) using the extended Tofts model (18) also showed lower variation of k_{ep} in response to different AIF sources compared with K^{trans}. Because k_{ep}, like K^{trans}, is also a measure of perfusion and permeability, the low sensitivity of k_{ep} to AIF amplitude uncertainty suggests that k_{ep} could be a more robust and reproducible imaging biomarker than K^{trans} for DCE-MRI characterization of tissue microvasculature (37) when consistent and accurate AIF quantification is difficult.

In pair-wise assessment of agreement in parameter values obtained with two different AIFs, the worst agreements (or the smallest CCC values) generally occurred when a measured AIF (from acquired DCE-MRI data) was paired with the literature population-averaged GP AIF, for any parameter and under the condition of either unadjusted or adjusted AIFs. It is important to note that, in addition to amplitude, the AIF curve shape also influences the estimation of the PK parameters (3, 9). Although the methods used by the 9 QIN centers to measure the AIFs were quite different (9), the individually measured AIFs captured the actual AIF curve shapes from the DCE-MRI data. The curve shape is specific to data acquisition details and data sampled for AIF quantification. This may not be well represented by the GP AIF, which is modeled on the basis of data from the aorta or iliac arteries acquired with different pulse sequence parameters at a different field strength. Such differences between the measured AIFs and GP AIF are probably a central reason why any pair-wise comparison of the GP AIF with a measured AIF resulted in large differences in estimated PK parameter values. Therefore, whenever possible, an individually measured AIF should be used for PK analysis of DCE-MRI data instead of a generic population-averaged AIF, which may be unrelated to a specific study. This conclusion is based on the results from this study, as well as on those from part I of this data analysis challenge (9), which were obtained from a single time-point pretreatment prostate DCE-MRI data sets. For longitudinal DCE-MRI studies of cancer response to treatment, percent changes in parameter values (rather than absolute values) are generally used to assess therapy response, and high parameter repeatability is crucial. The use of a fixed population-averaged AIF may have advantages over individually measured AIFs because of the likely randomness of AIF measurement errors in the latter approach across multiple studies over a period of time. A recent DCE-MRI study of 13 patients with abdominal metastases by Rata et al. (38) shows that the highest parameter repeatability in a baseline test–retest study was achieved with a population-averaged AIF in comparison with three approaches of direct AIF measurement from acquired imaging data, and that, as a result, parameters derived with the population-averaged AIF have the highest sensitivity to treatment-induced changes. Further investigations with larger patient cohorts and data collected from different organs are needed before clear recommendations can be made in terms of direct AIF measurement versus fixed population-averaged AIF (39) for longitudinal DCE-MRI evaluation of cancer therapy response.

The representative parametric maps of K^{trans} and τ_i (Figure 5) indicate that DCE-MRI parameter variations caused by AIF variations are mostly systematic. Despite differences in absolute voxel parameter values owing to different AIFs used in SSM analysis, it can be seen that the pattern of voxel parameter distribution largely remains the same for all the K^{trans} or τ_i maps, that is, there are no visible changes in the spatial locations of the parameter "hot" and "cold" spots when different AIFs were used. This is also observed in the voxel-based k_{ep} and v_e parametric maps (not shown). Therefore, the assessment of PK parameter spatial heterogeneity using texture analysis of parametric maps may not be affected by variations in AIF determination. However, quantitative texture feature analysis needs to be conducted to test this hypothesis.

The τ_i parameter is unique to the SSM, and its reciprocal, $1/\tau_i$, is a measure of the rate of water cycling across cell membrane. Previous studies (27-31) indicate that τ_i is an imaging biomarker of cellular metabolic activity, specifically the activity of Na^+-K^+-ATPase, which consumes ATP and drives active water cycling. The relationship between τ_i and Na^+-K^+-ATPase was recently validated by a study of breast cancer cell lines using magnetic resonance and immunofluorescence measurements (40). The present multicenter data analysis challenge shows that τ_i (along with v_e) not only has the smallest AIF-caused variance among the PK parameter but is also (along with k_{ep}) the least sensitive to changes in AIF amplitude. Therefore, the inclusion of the τ_i parameter (or the use of the SSM) in DCE-MRI studies could be advantageous, especially for studies of therapeutic monitoring when random errors of AIF measurement in multiple exams over time could lead to low accuracy and precision in parameters such as K^{trans} and consequently either over- or underestimation of true response to treatment.

Similar to challenge part I (9), this multicenter study has several limitations. The study cohort size is small (11 patients) and the results should be validated with a larger cohort size. Due to the lack of data for R_{10} measurement in the shared data sets,

a fixed R_{10} value was used for PK analysis of all voxel data across all 11 patients. Although this approach eliminated the contamination of R_{10} variation in the evaluation of the effect of AIF variation on SSM parameters, the use of a uniformed presumed R_{10} value most likely reduced the accuracy in the estimated parameter values, as well as in the assessments of intra- and intertumoral heterogeneity. The AIF determination methods used by the 9 QIN centers are constrained to direct measurement from the imaging data. Other AIF quantification methods were not evaluated in this study. It would be interesting to investigate if AIF variations from a method like blinded estimation (10) will have similar effects on PK parameter variance. Because the shared prostate DCE-MRI data sets were all acquired before treatment, it was not possible to assess the effects of AIF variation on DCE-MRI assessment of prostate cancer response to treatment, particularly the comparison of the individually measured AIFs with the population-averaged GP AIF.

CONCLUSION

The results from this part II of a multicenter DCE-MRI data analysis challenge using the SSM are generally consistent with those obtained using the standard Tofts model (9). Variations in AIF quantification result in considerable variance in the estimated PK parameters. Among the three conventional PK parameters (ie, K^{trans}, v_e, and k_{ep}), the AIF-caused parameter variation is the highest in K^{trans} and the lowest in v_e. The SSM-specific τ_i parameter has low AIF-caused variation, similar to v_e. Use of the reference tissue method to adjust the amplitude of measured AIF can improve agreement in AIF and reduce variations in K^{trans} and v_e, but it has little effect on k_{ep} and τ_i. k_{ep} may be a more robust and reproducible marker of prostate microvasculature than K^{trans} because of its lower sensitivity to AIF uncertainty. Because τ_i is the least sensitive among the four parameters to AIF variation and has the potential of being an imaging biomarker of metabolic activity, the SSM could be the better choice for PK analysis of DCE-MRI data acquired with sufficient sensitivity to the water-exchange kinetics (41), especially those acquired in longitudinal studies to assess cancer response to treatment. In multicenter quantitative DCE-MRI studies, central data analysis with a fixed AIF determination method should be adopted to minimize parameter variations due to inconsistency in AIF determination by each local site. If local PK data analysis is required, the AIFs used by the local sites need to be consistent: either individually measured from acquired data or a population-averaged AIF, but not both. Furthermore, the reference tissue-adjusted AIF should be used in data modeling to reduce AIF-caused parameter variations.

ACKNOWLEDGMENTS

This study was supported by National Institutes of Health grants U01-CA154602, U01-CA151261, U01-CA183848, U01-CA154601, U01-CA148131, U01-CA176110, U01-CA172320, U01-CA142565, U01-CA166104, and U01-CA140230, and Circle of Giving award from Oregon Health & Science University Center for Women's Health.

Disclosures: No disclosures to report.

Conflicts of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. Yankeelov TE, Mankoff DA, Schwartz LH, Lieberman FS, Buatti JM, Mountz JM, Erickson BJ, Fennessy FMM, Huang W, Kalpathy-Cramer J, Wahl RL, Linden HM, Kinahan PE, Zhao B, Hylton NM, Gillies RJ, Clarke L, Nordstrom R, Rubin DL. Quantitative imaging in cancer clinical trials. Clin Cancer Res. 2016;22:284–290.
2. O'Connor JPB, Jackson A, Parker GJM, Roberts C, Jayson GC. Dynamic contrast-enhanced MRI in clinical trials of antivascular therapies. Nat Rev Clin Oncol. 2012;9:167–177.
3. Khalifa F, Soliman A, El-Baz A, El-Ghar MA, El-Diasty T, Gimel'farb G, Ouseph R, Dwyer AC. Models and methods for analyzing DCE-MRI: a review. Med Phys. 2014;41:124301.
4. Leach MO, Morgan B, Tofts PS, Buckley DL, Huang W, Horsfield MA, Chenevert TL, Collins DJ, Jackson A, Lomas D, Whitcher B, Clarke L, Plummer R, Judson I, Jones R, Alonzi R, Brunner T, Koh DM, Murphy P, Waterton JC, Parker G, Graves MJ, Scheenen TW, Redpath TW, Orton M, Karczmar G, Huisman H, Barentsz J, Padhani A; Experimental Cancer Medicine Centres Imaging Network Steering Committee. Imaging vascular function for early stage clinical trials using dynamic contrast-enhanced magnetic resonance imaging. Eur Radiol. 2012;22:1451–1464.
5. Huang W, Li X, Chen Y, Li X, Chang MC, Oborski MJ, Malyarenko DI, Muzi M, Jajamovich GH, Fedorov A, Tudorica A, Gupta SN, Laymon CM, Marro KI, Dyvorne HA, Miller JV, Barbodiak DP, Chenevert TL, Yankeelov TE, Mountz JM, Kinahan PE, Kikinis R, Taouli B, Fennessy F, Kalpathy-Cramer J. Variations of dynamic contrast-enhanced magnetic resonance imaging in evaluation of breast cancer therapy response: a multicenter data analysis challenge. Transl Oncol. 2014;7:153–166.
6. Schabel MC, Morrell GR. Uncertainty in T(1) mapping using the variable flip angle method with two flip angles. Phys Med Biol 2009 Jan 7;54:N1–N8.
7. Fedorov A, Fluckiger J, Ayers GD, Li X, Gupta SN, Tempany C, Mulkern R, Yankeelov TE, Fennessy FM. A comparison of two methods for estimating DCE-MRI parameters via individual and cohort based AIFs in prostate cancer: a step towards practical implementation. Magn Reson Imaging. 2014 May;32:321–329.
8. Keil VC, Madler B, Gieseke J, Fimmers R, Hattingen E, Schild HH, Hadizadeh DR. Effects of arterial input function selection on kinetic parameters in brain dynamic contrast-enhanced MRI. Magn Reson Imaging. 2017;40:83–90.
9. Huang W, Chen Y, Fedorov A, Li X, Jajamovich GH, Malyarenko DI, Aryal MP, LaViolette PS, Oborski MJ, O'Sullivan F, Abramson RG, Jafari-Khouzani K, Afzal A, Tudorica A, Moloney B, Gupta SN, Besa C, Kalpathy-Cramer J, Mountz JM, Laymon CM, Muzi M, Kinahan PE, Schmainda K, Cao Y, Chenevert TL, Taouli B, Yankeelov TE, Fennessy FMM, Li X. The impact of arterial input function determination variations on prostate dynamic contrast-enhanced magnetic resonance imaging pharmacokinetic modeling: a multicenter data analysis challenge. Tomography. 2016;2:56–66.
10. Schabel MC, Fluckiger JU, DiBella EV. A model-constrained Monte Carlo method for blind arterial input function estimation in dynamic contrast-enhanced MRI: I. Simulations. Phys Med Biol. 2010;55:4783–4806.
11. Yankeelov TE, Luci JJ, Lepage M, Li R, Debusk L, Lin PC, Price RR, Gore JC. Quantitative pharmacokinetic analysis of DCE-MRI data without an arterial input function: a reference region model. Magn Reson Imaging. 2005;23:519–529.
12. Yang C, Karczmar GS, Medved M, Stadler WM. Estimating the arterial input function using two reference tissues in dynamic contrast-enhanced MRI studies: fundamental concepts and simulations. Magn Reson Med. 2004;52:1110–1117.
13. Kovar DA, Lewis M, Karczmar GS. A new method for imaging perfusion and contrast extraction fraction: input functions derived from reference tissues. J Magn Reson Imaging. 1998;8:1126–1134.
14. Parker GJ, Roberts C, Macdonald A, Buonaccorsi GA, Cheung S, Buckley DL, Jackson A, Watson Y, Davies K, Jayson GC. Experimentally-derived functional form for a population-averaged high-temporal-resolution arterial input function for dynamic contrast-enhanced MRI. Magn Reson Med. 2006;56:993–1000.
15. Li X, Priest RA, Woodward WJ, Siddiqui F, Beer TM, Garzotto MG, Rooney WD, Springer CS Jr. Cell membrane water exchange effects in prostate DCE-MRI. J Magn Reson. 2012;218:77–85.

16. Li X, Cai Y, Moloney B, Chen Y, Huang W, Woods M, Coakley FV, Rooney WD, Garzotto MG, Springer CS. Relative sensitivities of DCE-MRI pharmacokinetic parameters to arterial input function (AIF) scaling. J Magn Reson. 2016; 269:104–112.
17. Tofts PS, Kermode AG. Measurement of the blood-brain barrier permeability and leakage space using dynamic MR imaging. Magn Reson Med. 1991;17:357–367.
18. Tofts PS, Brix G, Buckley DL, Evelhoch JL, Henderson E, Knopp MV, Larsson HB, Lee TY, Mayr NA, Parker GJ, Port RE, Taylor J, Weisskoff RM. Estimating kinetic parameters from dynamic contrast-enhanced T1-weighted MRI of a diffusable tracer: standardized quantities and symbols. J Magn Reson Imaging. 1999;10:223–232.
19. Azahaf M, Haberley M, Betrouni N, Ernst O, Behal H, Duhamel A, Ouzzane A, Puech P. Impact of arterial input function selection on the accuracy of dynamic contrast-enhanced MRI quantitative analysis for diagnosis of clinically significant prostate cancer. J Magn Reson Imaging. 2016;43:737–749.
20. Yankeelov TE, Rooney WD, Li X, Springer CS. Variation of the relaxographic "Shutter-Speed" for transcytolemmal water exchange affects the CR bolus-tracking curve shape. Magn Reson Med. 2003;50:1151–1169.
21. Li X, Rooney WD, Springer CS. A unified pharmacokinetic theory for intravascular and extracellular contrast agents. Magn Reson Med 2005;54:1351–1359. [Erratum. Magn Reson Med 2006;55:1217.
22. Huang W, Tudorica LA, Li X, Thakur SB, Chen Y, Morris EA, Tagge IJ, Korenblit M, Rooney WD, Koutcher JA, Springer CS. Discrimination of benign and malignant breast lesions by using shutter-speed dynamic contrast-enhanced MR imaging. Radiology. 2011;261:394–403.
23. Tudorica LA, OH KY, Roy N, Kettler MD, Chen Y, Hemmingson SL, Afzal A, Grinstead JW, Laub G, Li X, Huang W. A feasible high spatiotemporal resolution breast DCE-MRI protocol for clinical settings. Magn Reson Imaging. 2012;30: 1257–1267.
24. Li X, Priest RA, Woodward WJ, Tagge IJ, Siddiqui F, Huang W, Rooney WD, Beer TM, Garzotto MG, Springer CS. Feasibility of Shutter-Speed DCE-MRI for improved prostate cancer detection. Magn Reson Med. 2013;69:171–178.
25. Tudorica A, OH KY, Chui SYC, Roy N, Troxell ML, Naik A, Kemmer K, Chen Y, Holtorf ML, Afzal A, Springer CS, Li X, Huang W. Early Prediction and Evaluation of Breast Cancer Response to Neoadjuvant Chemotherapy Using Quantitative DCE-MRI. Transl Oncol. 2016;9:8–17.
26. Chawla S, Loevner LA, Kim SG, Hwang WT, Wang S, Verma G, Mohan S, LiVolsi V, Quon H, Poptani H. Dynamic contrast-enhanced MRI-derived intracellular water lifetime (τ_i): a prognostic marker for patients with head and neck squamous cell carcinomas. Am J Neuroradiol. 2018;39:138–144.
27. Springer CS, Li X, Tudorica LA, OH KY, Roy N, Chui SYC, Naik AM, Holtorf ML, Afzal A, Rooney WD, Huang W. Intratumor mapping of intracellular water lifetime: metabolic images of breast cancer? NMR Biomed. 2014;27:760–773.
28. Zhang Y, Poirier-Quinot M, Springer CS, Balschi JA. Active trans-plasma membrane water cycling in yeast is revealed by NMR. Biophys J. 2011;101:2833–2842.
29. Zhang Y, Balschi JA. Water exchange kinetics in the isolated heart correlate with Na + /K + ATPase activity: potentially high saptiotemporal resolution. Proc Intl Soc Mag Reson Med. 2013;21:4045.
30. Sampath S, Parimal SA, Huang W, Mazlan I, Croft G, Totman T, Yvonne TWZ, Manigbas E, Chang MML, Qiu A, Klimas M, Evelhoch JL, de Kleijn DPV, Chin CL. Quantitative MRI measurement of the interplay between myocardial function, perfusion, structure and metabolism during acute and chronic remodeling in a porcine model of myocardial infarction. Proc Intl Soc Magn Reson Med. 2017;25:3249.
31. Bai R, Springer CS, Plenz D, Basser PJ. Fast, Na^+/K^+ pump driven, steady-state transcytolemmal water exchange in neuronal tissue: a study of rat brain cortical cultures. Magn Reson Med. 2018;79:3207–3217.
32. Hegde JV, Mulkern RV, Panych LP, Fennessy FM, Fedorov A, Maier SE, Tempany CM. Multiparametric MRI of prostate cancer: an update on state-of-the-art techniques and their performance in detecting and localizing prostate cancer. J Magn Reson Imaging. 2013;37:1035–1054.
33. Gupta R. A new look at the method of variable nutation angle for the measurement of spin-lattice relaxation time using Fourier transform NMR. J Magn Reson. 1977;25:231–235.
34. Lu H, Clingman C, Golay X, van Zijl PC. Determining the longitudinal relaxation time (T1) of blood at 3.0 Tesla. Magn Reson Med. 2004;52:679–682.
35. Padhani AR, Hayes C, Landau S, Leach MO. Reproducibility of quantitative dynamic MRI of normal human tissues. NMR Biomed. 2002;15:143–153.
36. Eye AV, Mun EY. Analyzing Rater Agreement: Manifest Variable Methods. Mahwah, NJ: Laurence Erlbaum Associates, Publishers; 2005.
37. Li X, Abramson RG, Arlinghaus LR, Kang H, Chakravarthy AB, Abramson VG, Farley J, Mayer IA, Kelley MC, Meszoely IM, Means-Powell J, Grau AM, Sanders M, Yankeelov TE. Multiparametric magnetic resonance imaging for predicting pathological response after the first cycle of neoadjuvant chemotherapy in breast cancer. Invest Radiol. 2015;50:195–204.
38. Rata M, Collins DJ, Darcy J, Messiou C, Tunariu N, Desouza N, Young H, Leach MO, Orton MR. Assessment of repeatability and treatment response in early phase clinical trials using DCE-MRI: comparison of parametric analysis using MR- and CT-derived arterial input functions. Eur Radiol. 2016;26:1991–1998.
39. Li X, Welch EB, Arlinghaus LR, Chakravarthy AB, Xu L, Farley J, Loveless ME, Mayer IA, Kelley MC, Meszoely IM, Means-Powell JA, Abramson VG, Grau AM, Gore JC, Yankeelov TE. A novel AIF tracking method and comparison of DCE-MRI parameters using individual and population-based AIFs in human breast cancer. Phys Med Biol. 2011;56:5753–5769.
40. Ruggiero MR, Baroni S, Pezzana S, Ferrante G, Crich SG, Aime S. Evidence for role of intracellular water lifetime as a tumor biomarker obtained by in vivo field-cycling relaxometry. Angew Chem Int Ed. 2018;57:1–6.
41. Li X, Huang W, Rooney WD. Signal-to-noise ratio, contrast-to-noise ratio, and pharmacokinetic modeling considerations in dynamic-contrast-enhanced magnetic resonance imaging. Magn Reson Imaging. 2012;30:1313–1322.

TOMOGRAPHY® RESEARCH ARTICLE

Evaluating Multisite rCBV Consistency from DSC-MRI Imaging Protocols and Postprocessing Software Across the NCI Quantitative Imaging Network Sites Using a Digital Reference Object (DRO)

Laura C. Bell[1], Natenael Semmineh[1], Hongyu An[2], Cihat Eldeniz[2], Richard Wahl[2], Kathleen M. Schmainda[3], Melissa A. Prah[3], Bradley J. Erickson[4], Panagiotis Korfiatis[4], Chengyue Wu[5], Anna G. Sorace[5], Thomas E. Yankeelov[5], Neal Rutledge[5], Thomas L. Chenevert[6], Dariya Malyarenko[6], Yichu Liu[7], Andrew Brenner[7], Leland S. Hu[8], Yuxiang Zhou[8], Jerrold L. Boxerman[9,10], Yi-Fen Yen[11], Jayashree Kalpathy-Cramer[11], Andrew L. Beers[11], Mark Muzi[12], Ananth J. Madhuranthakam[13], Marco Pinho[13], Brian Johnson[13,14], and C. Chad Quarles[1]

[1]Division of Neuroimaging Research, Barrow Neurological Institute, Phoenix, AZ; [2]Mallinckrodt Institute of Radiology, Washington University in St. Louis, St. Louis, MO; [3]Departments of Radiology and Biophysics, Medical College of Wisconsin, Wauwatosa, WI; [4]Department of Radiology, Mayo Clinic, Rochester, MN; [5]Department of Diagnostic Medicine, University of Texas at Austin, Austin, TX; [6]Department of Radiology, University of Michigan, Ann Arbor, MI; [7]UT Health San Antonio, San Antonio, TX; [8]Department of Radiology, Mayo Clinic, Scottsdale, AZ; [9]Department of Diagnostic Imaging, Rhode Island Hospital, Providence, RI; [10]Alpert Medical School of Brown University, Providence, RI; [11]Department of Radiology, Massachusetts General Hospital, Boston, MA; [12]Department of Radiology, University of Washington, Seattle, Washington; [13]UT Southwestern Medical Center, Dallas, TX; and [14]Philips Healthcare, Gainesville, FL

Corresponding Author:
Laura C. Bell, PhD
Division of Neuroimaging Research,
Barrow Neurological Institute, Phoenix, AZ, 85013;
E-mail: laura.bell@barrowneuro.org

Key Words: DSC-MRI, relative cerebral blood volume, standardization, multisite consistency, reproducibility

Abbreviations: Relative cerebral blood volume (rCBV), postprocessing methods (PMs), imaging protocols (IPs), dynamic susceptibility contrast magnetic resonance imaging (DSC-MRI), Quantitative Imaging Network (QIN), American Society of Functional Neuroradiology (ASFNR), digital reference object (DRO), standard imaging protocol (SIP), intraclass correlation coefficient (ICC), limits of agreement (LOA), covariance (CV), echo time (TE), normal appearing white matter (NAWM)

ABSTRACT

Relative cerebral blood volume (rCBV) cannot be used as a response metric in clinical trials, in part, because of variations in biomarker consistency and associated interpretation across sites, stemming from differences in image acquisition and postprocessing methods (PMs). This study leveraged a dynamic susceptibility contrast magnetic resonance imaging digital reference object to characterize rCBV consistency across 12 sites participating in the Quantitative Imaging Network (QIN), specifically focusing on differences in site-specific imaging protocols (IPs; n = 17), and PMs (n = 19) and differences due to site-specific IPs and PMs (n = 25). Thus, high agreement across sites occurs when 1 managing center processes rCBV despite slight variations in the IP. This result is most likely supported by current initiatives to standardize IPs. However, marked intersite disagreement was observed when site-specific software was applied for rCBV measurements. This study's results have important implications for comparing rCBV values across sites and trials, where variability in PMs could confound the comparison of therapeutic effectiveness and/or any attempts to establish thresholds for categorical response to therapy. To overcome these challenges and ensure the successful use of rCBV as a clinical trial biomarker, we recommend the establishment of qualifying and validating site- and trial-specific criteria for scanners and acquisition methods (eg, using a validated phantom) and the software tools used for dynamic susceptibility contrast magnetic resonance imaging analysis (eg, using a digital reference object where the ground truth is known).

INTRODUCTION

The relative cerebral blood volume (rCBV), derived from dynamic susceptibility contrast magnetic resonance imaging (DSC-MRI), is an established biomarker of glioma status that can aid in diagnosis (1), detecting treatment response (2, 3), guiding biopsies (4, 5), and reliable differentiation of post-treatment radiation effects and tumor progression (6–10). It is also increasingly leveraged as a biomarker of early therapeutic response in

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2018.00041

clinical trials (11, 12). However, variations in image acquisition and postprocessing methods (PMs) can limit rCBV reproducibility, potentially diminishing its clinical utility. To promote rCBV reproducibility across institutions, many national initiatives are underway to standardize DSC-MRI acquisition and PMs, including National Cancer Institute's Quantitative Imaging Network (QIN), Radiological Society of North America's Quantitative Imaging Biomarkers Alliance (QIBA), and the National Brain Tumor Society's Jumpstarting Brain Tumor Drug Development Coalition. Recent imaging protocol (IP) recommendations by the American Society of Functional Neuroradiology (ASFNR) has served as the first step in standardizing DSC-MRI protocols for clinical applications (13).

To aid in this effort, 12 institutions within the QIN aimed to investigate and determine the current rCBV reproducibility using a recently developed and validated in silico digital reference object (DRO) that is representative of a wide range of possible glioma magnetic resonance signals (14). Leveraging this DRO enables us as a community to determine the multisite consistency in rCBV owing to varying permutations of imaging acquisition parameters and postprocessing steps. In specific, our goals are to characterize rCBV consistency under conditions where there exist: (1) variations in the site-specific imaging acquisition parameters (PMs held constant), (2) variations in only site-specific PMs (IP held constant), and (3) variations owing to site-specific imaging and postprocessing protocols. Results from this community-based challenge will help steer standardization of DSC-MRI rCBV protocols with the hope that it can be successfully translated to the clinical setting.

MATERIALS AND METHODS

This National Cancer Institute QIN DSC-DRO challenge project was proposed and organized by the investigators at Barrow Neurological Institute (BNI). Eleven centers participated in this project: BNI (the managing center), Brown University (BU), Massachusetts General Hospital (MGH), Mayo Clinic Arizona (Mayo AZ), Mayo Clinic Minnesota (Mayo MN), Medical College of Wisconsin (MCW), University of Michigan (UM1), The University of Texas Health at San Antonio (UTSA), University of Texas at Austin (UT), University of Texas Southwestern Medical Center at Dallas (UTSW), University of Washington (UW), and Washington University (WashU). Unless specifically named, these participating sites have been anonymized, in no particular order, and will be referred to as sites 01–12 as seen in Table 1.

This project comprised 3 phases, summarized in the last 3 columns of Table 1, to evaluate the influence of IPs and/or PMs on multisite consistency:

- Phase I ("site IP w/constant PM") involved each participating site to submit their current clinical DSC IP to the managing center. The managing center then simulated site-specific DROs reflecting the IP parameters provided. Some sites provided >1 IP owing to differences in field strengths (sites 01, 04, and 05), dosing schemes (sites 03 and 10), and acquisition method (site 04). In total, 19 different IPs were submitted. The managing center postprocessed (specific details below in "Site-specific IP and PM") rCBV maps of each of these submitted site-specific IP DROs to evaluate differences owing to the IP provided.
- Phase II ("constant IP w/site PM") involved analysis of a "standard imaging protocol" (SIP), which represents DSC-MRI data acquired using the IP recommended by ASFNR (13). Each site was asked to process DSC-MRI DRO data derived from the SIP. Some sites choose to use multiple commercially available software packages (site 03) and different rCBV definitions (sites 05, 06, 12), yielding a total of 17 submitted rCBV maps.
- Phase III ("site IP w/site PM") required each site to calculate rCBV maps using their PM of choice and the site-specific DRO data. Combining the possible permutations owing to choice of IP and PM from phases I–II, a total of 25 rCBV maps were submitted.

All sites but 1 completed all 3 phases of the challenge. Site 11 completed only phase I, and these results are included in this study.

DRO Simulations

The DSC-MRI signals for each IP were simulated using a recently developed and validated population-based DRO that was trained to generate realistic signals using in vivo data from >40 000 voxels derived from patient data (14). The resulting DRO, which contains 10 000 unique voxels, reflects the distribution of perfusion, permeability, precontrast T1, T2*, diffusion coefficients, and the vascular and cellular features found in patients with high-grade glioma. Using this DRO, the DSC-MRI signals and resulting rCBV values can be computed for any combination of preload dosing scheme, contrast agent choice (by varying T1 relaxivities specific to the contrast agent), pulse sequence parameters, and postprocessing protocol. For the purposes of this study, the DRO consisted of tumor voxels simulated under two blood-brain-barrier (BBB) conditions to recapitulate DSC-MRI signals from an intact-BBB ($K^{trans} = 0$) and a disrupted-BBB ($K^{trans} > 0$). In addition to the tumor voxels, normal appearing white matter (NAWM) voxels ($K^{trans} = 0$) were simulated to normalize CBV. For the purposes of comparing site-to-site consistency, the SIP that has been postprocessed by the managing center was considered the reference standard where necessary. In our recent study, focused on investigating the influence of IP on CBV fidelity (15), the SIP yielded CBV values, when corrected for contrast agent leakage, that were among the most accurate.

Site-Specific IP and PM Methods

Site-specific IP and PM methods are briefly listed in Table 1. Overall, IPs were similar across sites. Most sites submitted clinical DSC IPs for 3 T with 3 sites that also included a 1.5 T IP. Overall, the following were the imaging parameters [mode (min-max)]: repetition time = 1500 milliseconds (1300–2560 milliseconds), echo time (TE) = 30 milliseconds (18–71 milliseconds), flip angle = 60° (60°–90°), preload dose = 0.05 mmol/kg (0–0.1 mmol/kg), and injection dose = 0.1 mmol/kg (0.05–0.15 mmol/kg). Five different gadolinium contrast agents were used across the 12 sites: gadobenate (n = 5), gadobutrol (n = 3), gadoterate (n = 2), gadoteridol (n = 1), and gadopentetate (n = 1). For PMs, there was a mix of software options used, including in-house-based software scripts (n = 4), IB Neuro (n = 4), 3D Slicer (n = 1),

Table 1. Summary of Participating Teams' IPs and PMs

	Imaging Protocol (IP)									ID Tag for Analysis		
	Scan Protocol				Dose Protocol							
Site Number	Field Strength	TR (ms)	TE (ms)	Flip	Preload (mmol/kg)	Injection (mmol/kg)	Time Between (min)	CA	Processing Method (PM)	Site IP w/Constant PP	Constant IP w/Site PP	Site IP w/Site PP
01	01:3.0 T	1500	30	60	0.05	0.10	3	Gadobenate	01: In-house processing	S01_IP01	S01_PM01	S01_IP01_PM01
	02:1.5 T	1500	30	60	0.05	0.10	3	Gadobenate		S01_IP02		S01_IP02_PM01
02	01:3.0 T	1600	30	60	0	0.1	n/a	Gadobenate	01: IB Neuro	S02_IP01	S02_PM01	S02_IP01_PM01
03	01:3.0 T	1500	31	90	0.05	0.15	6.5	Gadoterate	01: 3DSlicer	S03_IP01	S03_PM01	S03_IP01_PM01
	02:3.0 T	1500	31	90	0.1	0.1	6.5	Gadoterate	02: nordicICE	S03_IP02	S03_PM02	S03_IP01_PM02
									03: PGUI		S03_PM03	S03_IP01_PM03
												S03_IP02_PM01
												S03_IP02_PM02
												S03_IP02_PM03
04	01:3.0 T	1500	30	80	0.10	0.10	5	Gadobutrol	01: IB Neuro	S04_IP01	S04_PM01	S04_IP01_PM01
	02:3.0 T	1500	2,35	80	0	0.10	n/a	Gadobutrol		S04_IP02		n/a
	03:1.5 T	1500	30	72	0.10	0.10	5	Gadobutrol		S04_IP03		S04_IP03_PM01
	04:1.5 T	1500	2,35	72	0	0.10	n/a	Gadobutrol		S04_IP04		n/a
05	01:3.0 T	1300	30	60	0.025	0.10	5	Gadobutrol	01: IB Neuro (Integration limits 1)	S05_IP01	S05_PM01	S05_IP01_PM01
	02:1.5 T	1300	30	60	0.025	0.10	5	Gadobutrol	02: IB Neuro (Integration limits 2)	S05_IP02	S05_PM02	S05_IP01_PM02
												S05_IP02_PM01
												S05_IP02_PM02
06	01:3.0 T	1500	30	75	0.10	0.10	5	Gadoteridol	01: PGUI (rCBV definition 1)	S06_IP01	S06_PM01	S06_IP01_PM01
									02: PGUI (rCBV definition 2)		S06_PM02	S06_IP01_PM01
07	01:3.0 T	1500	30	65	0.025	0.075	6	Gadobenate	01: In-house processing	S07_IP01	S07_PM01	S07_IP01_PM01
08	01:3.0 T	1500	21	60	0.10	0.05	6	Gadobenate	01: In-house processing	S08_IP01	S08_PM01	S08_IP01_PM01
09	01:3.0 T	1500	18	60	0.05	0.05	6	Gadobenate	01: IB Neuro	S09_IP01	S09_PM01	S09_IP01_PM01
10	01:3.0 T	1900	36	90	0	0.10	n/a	Gadoterate	01: In-house processing	S10_IP01	S10_PM01	S10_IP01_PM01
	02:3.0 T	1900	36	90	0.10	0.10	5	Gadoterate		S10_IP02		S10_IP02_PM01
11	01:3.0 T	2560	71	90	0.025	0.10	2	Gadopentetate	n/a	S11_IP01	n/a	n/a
12	01:3.0 T	1757	30	90	0.033	0.067	8	Gadobutrol	01: Philips ISP (rCBV definition 1)	S12_IP01	S12_PM01	S12_IP01_PM01
									02: Philips ISP (rCBV definition 2)		S12_PM02	S12_IP01_PM02
									03: Philips ISP (rCBV definition 3)		S12_PM03	S12_IP01_PM03
Standard Protocol	01:3.0 T	1500	30	60	0.10	0.10	5	Gadopentetate	n/a	SIP	n/a	n/a
Total[a]: 12										19	17	25

[a] Excludes the standard protocol.

nordicICE (n = 1), PGUI (n = 2), and Philips IntelliSpace Portal (ISP; Philips Healthcare, Best, the Netherlands) (n = 1).

For PM methods, most sites defined rCBV = $\int_0^t \Delta R^*_{2,tumor-BSW} / \int_0^t \Delta R^*_{2,NAWM}$ and used the Boxerman–Schmainda–Weiskoff (BSW) method for leakage correction (16). A few sites submitted results that deviated from this postprocessing convention by alternative rCBV definitions (S06, S12) and differences in integration limits as determined by the software (S05). These differences are highlighted in Table 1. Site 06 defined CBV by the area under the curve of the deconvolved residue function. This deconvolved residue function was determined by singular value decompositions (rCBV definition 1) and by oscillating singular value decompositions approach (rCBV definition 2). S12 used 3 different rCBV definitions within the Philips ISP platform: a "model-free" option that integrates the area underneath the signal intensity curve (rCBV definition 1) (17), a "γ-variate" option that integrates the area underneath the signal intensity curve that has been fit to a γ-variate function (rCBV definition 2), and a "leakage correction" option that integrates the area underneath the computed delta R2* curve after a modified BSW leakage correction method is applied (rCBV definition 3). To be clear, the first 2 options of the Philips ISP do not apply any sort of leakage correction algorithm to the data. S05 included CBV maps calculated using the default inte-

Table 2. Summary of Participating Teams' PMs

Site Number	Software	CBV Definition	Normalized to NAWM?	Integration Limits	Leakage Correction Method	Comments
01	01: In-house processing	AUC of the ΔR_2^* time course	No	Time points: 2 to 64 (93 sec)	BSW leakage correction method	Manual inspection of pre- and post- contrast points for rCBV integration
02	01: IB Neuro	AUC of the ΔR_2^* time course	Yes	automatically detected (default option)	BSW leakage correction method	Default IB Neuro settings for rCBV
03	01: 3DSlicer	AUC of the ΔR_2^* time course	No	118 seconds	BSW leakage correction method	No thresholding
	02: nordicICE	AUC of the ΔR_2^* time course	Yes	Time points: 2 to 121 (178.5 sec)	BSW leakage correction method	
	03: PGUI	AUC of the ΔR_2^* time course	No	Time points: 2 to 121 (178.5 sec)	BSW leakage correction method	No thresholding, but smoothing applied
04	01: IB Neuro	AUC of the ΔR_2^* time course	Yes	automatically detected (default option)	BSW leakage correction method	
05	01: IB Neuro (Integration limits 1)	AUC of the ΔR_2^* time course	Yes	automatically detected (default option)	BSW leakage correction method	
	02: IB Neuro (Integration limits 2)	AUC of the ΔR_2^* time course	Yes	180 seconds (all time points)	BSW leakage correction method	
06	01: PGUI (rCBV definition 1)	Deconvolution of the residue function (SVD)	No	Time points: 5 to 121 (174 sec)	BSW leakage correction method	
	02: PGUI (rCBV definition 2)	Deconvolution of the residue function (oSVD)	No	Time points: 5 to 121 (174 sec)	BSW leakage correction method	
07	01: In-house processing	AUC of the ΔR_2^* time course	No	automatically detected (default option)	BSW leakage correction method	
08	01: In-house processing	AUC of the ΔR_2^* time course	Yes	90 sec	BSW leakage correction method	
09	01: IB Neuro	AUC of the ΔR_2^* time course	Yes	automatically detected (default option)	BSW leakage correction method	Did not use the entire NAWM ROI - instead used a 6 mm × 6 mm (~225 pixels) ROI
10	01: In-house processing	AUC of the ΔR_2^* time course	No	171 sec	BSW leakage correction method	ΔR_2^* maps were smoothed with a 5 × 5 Gaussian window that had an FWHM value of 3 mm
11	n/a					
12	01: Philips ISP (rCBV definition 1)	AUC of the SI time course	No	Based on the characteristics of signal time curves	No leakage correction method	
	02: Philips ISP (rCBV definition 2)	AUC of the SI time course fitted to a gamma-variate	No	Based on the characteristics of signal time curves	No leakage correction method	
	03: Philips ISP (rCBV definition 3)	AUC of the ΔR_2^* time course	No	180 s	BSW leakage correction method	

gration limits set by IB Neuro (integration limits 1) and manually chose all time points in IB Neuro (integration limits 2). A little less than 50% of the submitted rCBV maps were normalized to the NAWM. To compare maps, the managing site normalized tumor CBV to the mean NAWM CBV of all pixels when necessary. Specifics on site-specific postprocessing steps are outlined in Table 2.

The managing center postprocessed the site-specific DROs with an in-house script by defining $\text{rCBV} = \int_0^{120\text{sec}} \Delta R^*_{2,tumor-BSW} / \int_0^{120\text{sec}} \Delta R^*_{2,NAWM}$, where the conventional ΔR_2^* curves in the tumor were corrected for leakage effects using the BSW method. In our recent study, the CBV was found to be the most accurate by using these specific PM steps, and thus was chosen to be used as the reference where applicable (15). No thresholding, smoothing, or quality assessment was done before rCBV calculations when analyzed by the managing center.

Statistics

To evaluate the consistency of rCBV across sites owing to differences between IP and PM, the intraclass correlation coefficient (ICC) was calculated. Furthermore, to evaluate the agreement of rCBV between sites and a reference (SIP), the 95% limits of agreement (LOA) were extracted from a Bland–Altman analysis. Variability of rCBV was assessed across a distribution of rCBV values by calculating the covariance (CV) across sites. Lastly, Lin's correlation coefficient was calculated for rCBV

Table 3. Intraclass Correlation Coefficient Results for Each Phase of the Study for Computed rCBV from the Simulated Intact-BBB and Disrupted-BBB DRO

	Site-Specific IP w/Constant PM	Constant IP w/Site-Specific PM	Site-Specific IP w/Site-Specific PM
Intact-BBB	0.970	0.690	0.641
Disrupted-BBB	0.879	0.439	0.380

between the intact-BBB and disrupted-BBB DROs for each permutation of IP and PM to determine the agreement of rCBV after leakage correction was applied. All statistical calculations were done in MATLAB R2018a (The MathWorks Inc., Natick, MA) by the managing center.

RESULTS

In general, the ICC decreases when $K^{trans} > 0$, that is, disrupted-BBB (Table 3) for all the 3 phases of this study. High agreement is observed across sites when a constant PM is applied to site-specific IP (ICC = 0.879). However, when site-specific PMs are applied to either a constant IP or to their site-specific IP, the agreement is quite poor (ICC = 0.439 and 0.380, respectively).

Figure 1 shows consistency in rCBV measurements for all 3 phases of this study when compared with the reference. For each site, the 95% LOA of both $K^{trans} = 0$ (gray lines, intact-BBB) and $K^{trans} > 0$ (black lines, disrupted-BBB) are indicated in comparison to the reference (see Table 1 for site ID descriptions). For phase I (Figure 1A), the 95% LOA are generally fairly narrow and centered around the mean rCBV of the reference for the $K^{trans} > 0$ case. A few exceptions (S02_IP01, S10_IP01, and S12_IP01) show larger 95% LOA and a negative bias compared with the other sites. The first 2 sites (S02 and S10) did not use a contrast agent preload unlike the other sites, while the third site (S12) used 1/3 standard dose for a preload. Sites S09_IP01 and S10_IP01, although centered around the reference's mean rCBV, also express wider ranges of 95% LOA compared with other sites. These 2 sites have markedly lower TE and use less than a full standard dose compared with the other sites. Much larger LOA are seen for phase II in Figure 1B) than for that in Figure 1A. Large 95% LOA are observed for even the $K^{trans} = 0$ case, where no leakage correction is applied during postprocessing. The analysis software that show the smallest 95% LOA with the reference are in-house processing scrips (S01_PM01, S08_PM01), IB Neuro (S02_PM01, S04_PM01, S05_PM01, S05_PM02, S09_PM01), nordicICE (S03_PM02), and the "model-free option" in Philips ISP (S12_PM01). For phase III (Figure 1C), 9 out of the 24 sites show a tight 95% LOA and relatively no bias when compared to the SP reference (S01_IP01_PM01, S01_IP01_PM02, S03_IP02_PM02, S04_IP01_PM01, S04_IP03_PM01, S05_IP01_PM01, S05_IP01_PM02, S05_IP02_PM01, S05_IP02_PM02) for the $K^{trans} > 0$ case. These 4 sites implemented nordicICE, IB Neuro, and an in-house postprocessing script.

Figure 2 illustrates the CV as a function of rCBV across DROs for all voxels. The covariance across DROs ($n_{Phase\ I} = 19$, $n_{Phase\ II} = 17$ $n_{Phase\ III} = 25$) was calculated in the 10 000 tumor voxels and plotted against the mean rCBV of each voxel across DROs. The DRO simulated with $K^{trans} = 0$ (gray circles) and $K^{trans} > 0$ (black circles) is plotted along with the mean CV (horizontal line plots) across all voxels. This figure does not assume a reference for calculations. In general, the CV increases for each phase when more freedom is allowed in the rCBV calculations for both IP and PM. For phase I (Figure 2A), the average CV is 4% and it remains fairly flat over the rCBV distribution for $K^{trans} = 0$. However, when $K^{trans} > 0$, the average CV rose to 17% and exponentially decreased from roughly 60% to 10% as rCBV increased. For phases II and III (Figure 2, B and C, respectively), the CV is observed to exponentially decrease for both K^{trans} cases. For phase II, the average CV is 18% and 30% for $K^{trans} = 0$ and > 0, respectively. As rCBV increases, the CV exponentially

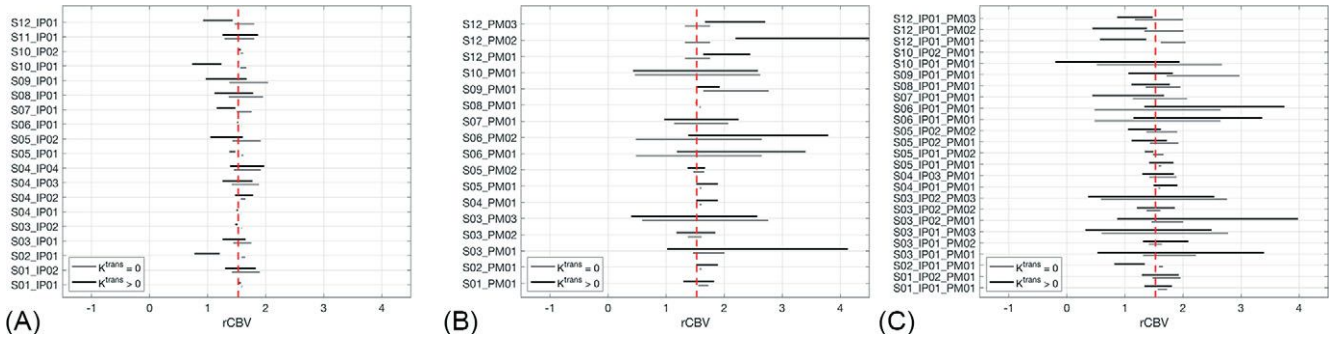


Figure 1. Bland–Altman limits of agreement (LOA) against the standard imaging protocol (SIP) plotted for site-specific IP w/constant postprocessing method (PM) (A), constant IP w/site-specific PM (B), and site-specific IP w/site-specific PM (C). The vertical dashed line is the mean rCBV across 10 000 voxels for the SIP.

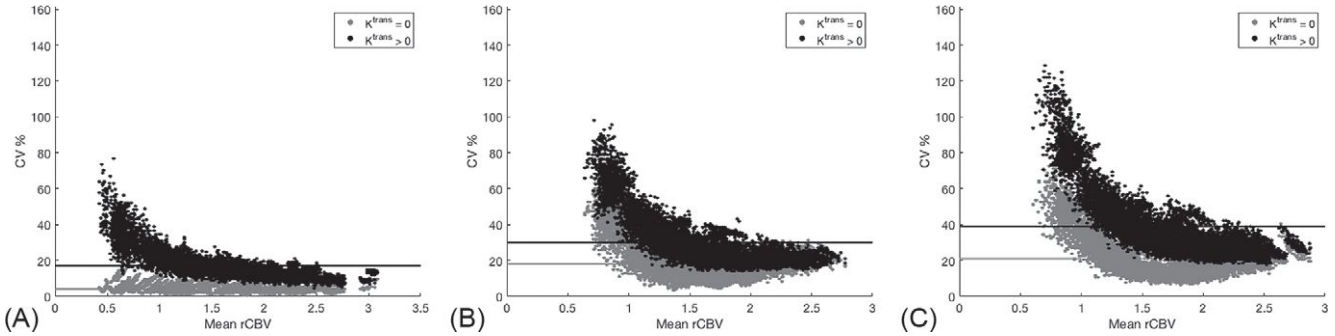


Figure 2. The covariance (CV%) across all relative cerebral blood volume (rCBV) maps for each of the 10 000 voxels plotted across the mean rCBV of the voxels for site-specific IP w/constant PM (A), constant IP w/site-specific PM (B), and site-specific IP w/site-specific PM (C). Results from the $K^{trans} = 0$ (light gray) and $K^{trans} > 0$ (black) are included with their mean CV% across all 10 000 voxels indicated for the horizontal lines. For all 3 phases of this study, the largest variation in rCBV occurs at the low rCBV range for $K^{trans} > 0$, and CV% increases when more freedom was introduced in the choice of IPs and PMs.

decreases from roughly 80% to 20% for both K^{trans} cases. For phase III, the average CV is 21% and 39% for $K^{trans} = 0$ and > 0, respectively. As rCBV increases, the CV exponentially decreases from roughly 120% to 35% for both K^{trans} cases.

Figure 3 examines the agreement between the intact-BBB ($K^{trans} = 0$) and disrupted-BBB ($K^{trans} > 0$) DRO for each processed rCBV map. The LCC for each analysis combination was sorted from the highest (perfect agreement = 1) to the lowest (no agreement = 0) for each of the 3 phases. A high agreement indicates that the processed CBV from the simulated disrupted-BBB DRO had high accuracy when compared to the simulated intact DRO where no leakage occurs. Site-specific IP with constant PM is shown by the black bars in the bar graph. Note that the third black bar is the SIP and has a high LCC value, which is consistent with previous results (15) and therefore used as reference in Figure 1. Here we observed that most of the sites' IPs are able to accurately compute CBV—most likely because these sites already use IPs similar to the SIP. Three site protocols had an LCCC < 0.8, indicating low rCBV accuracy when leakage effects are introduced: S02_IP01, S10_IP01, and S12_IP01. These protocols also resulted in large LOA and a negative bias as seen in Figure 1. These results indicate that the IP is highly sensitive to contrast agent leakage effects even when a leakage correction PM algorithm is applied. Constant IP with site-specific PM results are indicated in the dark gray bars in the bar graph. Here we observe 10 software programs that clearly show

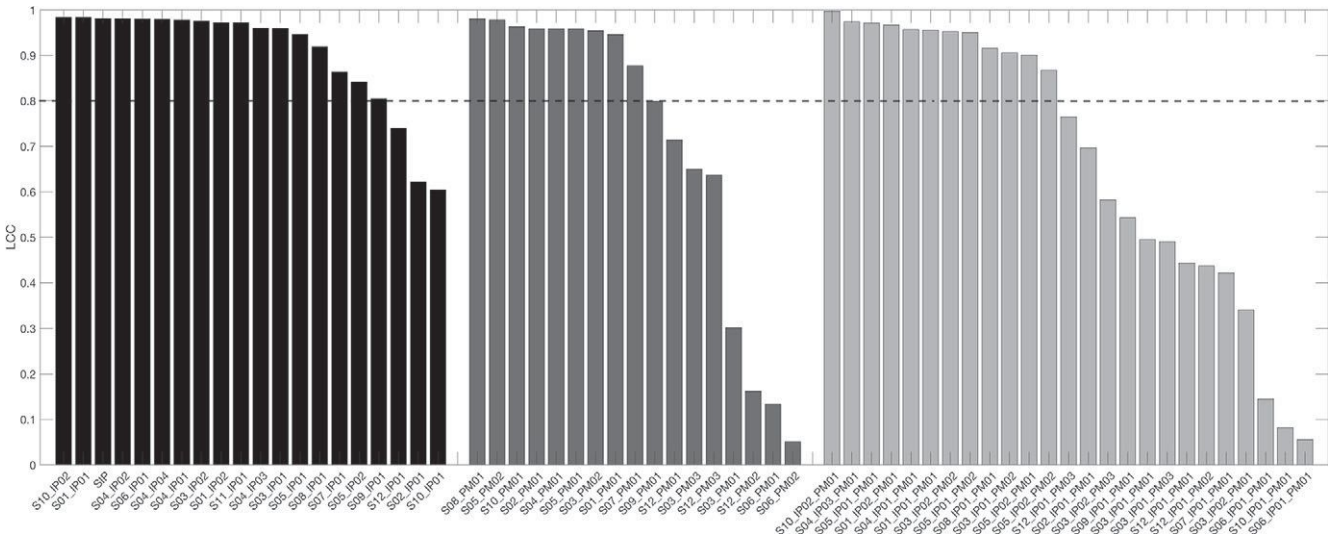


Figure 3. A bar plot of Lin's correlation coefficient (LCC) for each rCBV map for site-specific IP w/constant PM (black), constant IP w/site-specific PM (medium gray), and site-specific IP w/site-specific PM (light gray). Each phase is sorted by the resulting LCC from the highest to the lowest value. A horizontal bar at LCC = 0.8 is placed to evaluate agreement good agreement (LCC > 0.8).

high agreement: in-house scripts (n = 4), IB Neuro (n = 4), nordicICE (n = 1), and "model-free" option in the Philips ISP. Lastly, site-specific IP with site-specific PM resulted in 50% of the rCBV maps with LCC < 0.8, most likely owing to a combination of variations in IPs and postprocessing as deduced from the earlier 2 phases.

DISCUSSION AND CONCLUSION

Reproducibility in DSC-MRI rCBV is crucial for the success of multisite clinical trials. In this study, we have evaluated rCBV consistency owing to differences in both IPs and PMs across 12 QIN sites using a DRO. The results outlined in this manuscript show that standardization of both is warranted.

Our prior DRO investigation highlighted the significant influence of IPs (including preload dosing and pulse sequence parameters) on CBV accuracy (15). The findings of this study strongly indicate that differences in the PM can also confound multisite CBV consistency and accuracy. High agreement when site-specific IP were processed by the managing center most likely reflects the similarity of the IP parameters across all the sites owing to previous initiatives from the ASFNR that aimed to standardize IPs (13). However, it was observed that when no preload was used in the IP (S02_IP01 and S10_IP01), a systematic negative bias relative to the SIP occurs. Furthermore, a slight negative bias is observed for the sites that administered less than a full standard dose as the main injection (S07_IP01, S08_IP01, S09_IP01, S12_IP01). These 2 findings underscore potential challenges to comparing CBV changes in a clinical trial from sites that use dissimilar preload and bolus dosing protocols. Three sites (S01, S04, S05) provided clinical IPs for 1.5 T. Sites 01 and 05 used the same IP at both field strengths, and it was observed that the LOA did widen when compared to the 3.0 T protocol. Site 04 used a smaller flip angle at 1.5 T than at 3 T; however, a widening of LOA was still observed. Differences here may warrant further investigation into a standardized 1.5 T IP; however, for the scope of this paper, high agreement was observed when both field strengths were compared together.

When each site was asked to postprocess the SIP, agreement decreased substantially as indicated by the ICC and the 95% LOAs. Interestingly, the disagreement across sites is not isolated to differences in the leakage correction method, as poor agreement is also observed with the $K^{trans} = 0$ case. For the $K^{trans} = 0$ case, 1 potential source of disagreement in rCBV arises from whether smoothing is implemented in the software and the CBV definition. Methods 01 and 02 from S12 deviated from the traditional CBV definition, as these methods calculated CBV from the signal intensity curves, potentially losing the biophysics and kinetic properties. For the $K^{trans} > 0$ case, potential sources of disagreement in rCBV may be attributed to smoothing and the algorithms and/or implementation of algorithms used for leakage correction.

It is challenging to compare the results from this current study directly to prior ones since we performed a voxel-wise analysis across the DRO, whereas most other studies, like the recent DSC-MRI challenge (18), report comparisons between mean region of interest tumor values across data that likely exhibits patient-specific rCBV distributions. As seen in Figure 2, there is greater variation across platforms at low rCBV values. These differences most likely average out when hotspot types of analyses are performed. Although most likely sufficient for diagnosis of tumor grade, this might not be ideal for longitudinal assessment of treatment response where voxel-wise analysis and/or CBV difference quantification has shown to be more beneficial (11, 12). Despite this, our results indicating inconsistent CBV values as more freedom is allowed to the IP and processing methods is not surprising. Kelm et al. compared rCBV measurements using 3 software platforms (IB Neuro, FuncTool, and nordicICE) and also found significant variation in rCBV (19).

A limitation to our study is that the ROIs for brain tumor and NAWM have been clearly outlined and predetermined for analysis. In the context of patient data, allowing users to define ROIs would likely contribute to greater rCBV inconsistency. Schmainda et al. showed high mean CBV agreement when ROIs were predetermined (18). In addition, sites were not required to determine an AIF for the CBV calculations within this manuscript.

Results from this study and our prior DRO analysis, which focused on IP optimization (15), highlight the IPs and PMs that maximize rCBV accuracy and multisite consistency. First, IPs that yield the highest rCBV accuracy and multisite concordance utilize a full-dose contrast agent preload and a full-dose bolus injection, low (30°) or moderate (65°) flip angle, ~30 millisecond TE, and a ~1.5 millisecond TR. In both studies, the use of lower bolus dose injections (eg, 1/2 dose) were found to substantially reduce both consistency and accuracy, likely owing to the lower CNR. Second, the 2 studies further show that, even with optimized IPs, leakage correction should be applied to DSC-MRI data in brain tumors. Further, the correction algorithms should be based on the underlying biophysics and kinetics, such as the BSW correction, as they maximize both accuracy and precision. Generic leakage correction algorithms (like gamma variate fitting) that arbitrarily modify the shape of DSC-MRI data to remove T1 and/or T2* leakage effects are not recommended. It should be noted that in the IP optimization study (15), a low flip angle approach (30° with a 30 millisecond TE) with a full-dose bolus injection, no preload, and application of BSW leakage correction provided accuracy slightly less than that using the ideal protocol. Studies are currently underway to validate the clinical potential of this protocol as it could be a compelling single-dose option for routine surveillance scans and in clinical trials.

Although great efforts have been made to standardize DSC-MRI imaging acquisition protocols, this study highlights that poor CBV agreement can arise when there are variations in processing platforms. Highest agreement is observed when site-specific CBV maps are processed by 1 managing center, as might be expected in a clinical trial setting where acquisition and PMs are predetermined, and/or raw data are sent to a single site for analysis. However, differences in CBV, especially at low values, as would be expected with effective therapy, arise when different platforms are used. This finding has important implications for comparing CBV values across trials, where variability in trial-specific PMs could confound the comparison of therapeutic effectiveness and/or any attempts to establish thresholds for categorical response (eg, predetermined percent changes in

rCBV values that could be used to refine RANO criteria). To overcome these challenges and to ensure the successful use of rCBV as a clinical trial biomarker, it is critical that the DSC-MRI community establish qualifying and validating criteria, similar to that in the RSNA DCE-MRI Profile (20), for scanners and acquisition methods to be used in clinical trials (eg, using a validated phantom) and the software used for DSC-MRI analysis (eg, using a DRO where the ground truth is known).

ACKNOWLEDGMENTS

NIH/NCI R01CA213158 (LCB, NS, CCQ); NIH/NCI U01CA207091 (AJM, MCP); NIH/NCI U01CA166104 and P01CA085878 (DM, TLC); NIH/NCI U01CA142565 (CW, AGS, TEY, NR); NIH/NCI U01 CA176110 (KMS, MAP).

Disclosure: No disclosures to report.

Conflicts of Interest: The authors have no conflicts of interest to declare.

REFERENCES

1. Fink JR, Muzi M, Peck M, Krohn KA. Multimodality brain tumor imaging: MR imaging, PET, and PET/MR imaging. J Nucl Med. 2015;56:1554–1561.
2. Boxerman JL, Ellingson BM, Jeyapalan S, Elinzano H, Harris RJ, Rogg JM, Pope WB, Safran H. Longitudinal DSC-MRI for distinguishing tumor recurrence from pseudoprogression in patients with a high-grade glioma. Am J Clin Oncol. 2014; 0:1–7.
3. Ellingson BM, Zaw T, Cloughesy TF, Naeini KM, Lalezari S, Mong S, Lai A, Nghiemphu PL, Pope WB. Comparison between intensity normalization techniques for dynamic susceptibility contrast (DSC)-MRI estimates of cerebral blood volume (CBV) in human gliomas. J Magn Reson Imaging. 2012;35:1472–1477.
4. Chaskis C, Stadnik T, Michotte A, Van Rompaey K, D'Haens J. Prognostic value of perfusion-weighted imaging in brain glioma: a prospective study. Acta Neurochir (Wien). 2006;148:277–285.
5. Maia ACM, Malheiros SMF, da Rocha AJ, Stávale JN, Guimarães IF, Borges LR, Santos AJ, da Silva CJ, de Melo JG, Lanzoni OP, Gabbai AA, Ferraz FA. Stereotactic biopsy guidance in adults with supratentorial nonenhancing gliomas: role of perfusion-weighted magnetic resonance imaging. J Neurosurg. 2004;101: 970–976.
6. Danchaivijitr N, Waldman AD, Tozer DJ, Benton CE, Brasil Caseiras G, Tofts PS, Rees JH, Jäger HR. Low-grade gliomas: do changes in rCBV measurements at longitudinal perfusion-weighted MR imaging predict malignant transformation? Radiology. 2008;247:170–178.
7. Rossi Espagnet MC, Romano A, Mancuso V, Cicone F, Napolitano A, Scaringi C, Minniti G, Bozzao A. Multiparametric evaluation of low grade gliomas at follow-up: comparison between diffusion and perfusion MR with ^{18}F-FDOPA PET. Br J Radiol. 2016;89:20160476.
8. Boxerman JL, Paulson ES, Prah MA, Schmainda KM. The effect of pulse sequence parameters and contrast agent dose on percentage signal recovery in DSC-MRI: implications for clinical applications. AJNR Am J Neuroradiol. 2013;34:1364–1369.
9. Hu LS, Baxter LC, Smith KA, Feuerstein BG, Karis JP, Eschbacher JM, Coons SW, Nakaji P, Yeh RF, Debbins J, Heiserman JE. Relative cerebral blood volume values to differentiate high-grade glioma recurrence from posttreatment radiation effect: direct correlation between image-guided tissue histopathology and localized dynamic susceptibility-weighted contrast-enhanced perfusion MR imaging measurements. Am J Neuroradiol. 2009;30:552–558.
10. Barajas RF, Chang JS, Segal MR, Parsa AT, McDermott MW, Berger MS, Cha S. Differentiation of recurrent glioblastoma multiforme from radiation necrosis after external beam radiation therapy with dynamic susceptibility-weighted contrast-enhanced perfusion MR imaging. Radiology. 2009;253:486–496.
11. Galbán CJ, Lemasson B, Hoff BA, Johnson TD, Sundgren PC, Tsien C, Chenevert TL, Ross BD. Development of a multiparametric voxel-based magnetic resonance imaging biomarker for early cancer therapeutic response assessment. Tomography. 2015;1:44–52.
12. Boxerman JL, Zhang Z, Safriel Y, Larvie M, Snyder BS, Jain R, Chi TL, Sorensen AG, Gilbert MR, Barboriak DP. Early post-bevacizumab progression on contrast-enhanced MRI as a prognostic marker for overall survival in recurrent glioblastoma: results from the ACRIN 6677/RTOG 0625 Central Reader Study. Neuro Oncol. 2013;15:945–954.
13. Welker K, Boxerman J, Kalnin A, Kaufmann T, Shiroishi M, Wintermark M; American Society of Functional Neuroradiology MR Perfusion Standards and Practice Subcommittee of the ASFNR Clinical Practice Committee. ASFNR recommendations for clinical performance of MR dynamic susceptibility contrast perfusion imaging of the brain. AJNR. Am J Neuroradiol. 2015;36:E41–E51.
14. Semmineh NB, Stokes AM, Bell LC, Boxerman JL, Quarles CC. A population-based digital reference object (DRO) for optimizing dynamic susceptibility contrast (DSC)-MRI methods for clinical trials. Tomography. 2017;3:41–49.
15. Semmineh N, Bell L, Stokes A, Hu L, Boxerman J, Quarles C. Optimization of acquisition and analysis methods for clinical dynamic susceptibility contrast (DSC) MRI using a population-based digital reference object. AJNR Am J Neuroradiol. 2018;39:1981–1988.
16. Boxerman JL, Schmainda KM, Weisskoff RM. Relative cerebral blood volume maps corrected for contrast agent extravasation significantly correlate with glioma tumor grade, whereas uncorrected maps do not. AJNR Am J Neuroradiol. 2006;27:859–867.
17. Meyer-Baese A, Lange O, Wismueller A, Hurdal MK. Analysis of dynamic susceptibility contrast MRI time series based on unsupervised clustering methods. IEEE Trans Inf Technol Biomed. 2007;11:563–573.
18. Schmainda KM, Prah MA, Rand SD, Liu Y, Logan B, Muzi M, Rane SD, Da X, Yen YF, Kalpathy-Cramer J, Chenevert TL, Hoff B, Ross B, Cao Y, Aryal MP, Erickson B, Korfiatis P, Dondlinger T, Bell L, Hu L, Kinahan PE, Quarles CC. Multisite concordance of DSC-MRI analysis for brain tumors: results of a National Cancer Institute Quantitative Imaging Network Collaborative Project. AJNR Am J Neuroradiol. 2018;39:1008–1016.
19. Kelm ZS, Korfiatis PD, Lingineni RK, et al. Variability and accuracy of different software packages for dynamic susceptibility contrast magnetic resonance imaging for distinguishing glioblastoma progression from pseudoprogression. J Med imaging (Bellingham). 2015;2:26001.
20. DCE MRI Technical Committee. DCE MRI Quantification Profile, Quantitative Imaging Biomarkers Alliance. 2012; Version 1.0. Available from: http://rsna.org/QIBA_.aspx.

TOMOGRAPHY® RESEARCH ARTICLE

Developing a Pipeline for Multiparametric MRI-Guided Radiation Therapy: Initial Results from a Phase II Clinical Trial in Newly Diagnosed Glioblastoma

Michelle M. Kim[1], Hemant A. Parmar[2], Madhava P. Aryal[1], Charles S. Mayo[1], James M. Balter[1], Theodore S. Lawrence[1], and Yue Cao[1]

Departments of [1]Radiation Oncology and [2]Radiology, University of Michigan, Ann Arbor, MI

Corresponding Author:
Michelle Kim, MD
University of Michigan Medical Center, Department of Radiation Oncology, 1500 E. Medical Center Dr., Ann Arbor, MI 48109-0010;
E-mail: michekim@med.umich.edu

Key Words: pipeline, workflow, multiparametric, MRI, glioblastoma

Abbreviations: Dynamic contrast-enhanced (DCE), diffusion-weighted (DW), magnetic resonance imaging (MRI), glioblastoma (GBM), T1 gadolinium (T1-Gd), T2-weighted fluid attenuated inversion recovery (T2-FLAIR), magnetic resonance imaging (MRI), relative cerebral blood volume (rCBV), apparent diffusion coefficient (ADC), computed tomography (CT), in-house functional image analysis tool (imFIAT), arterial input function (AIF), 2-dimensional (2D), echo time (TE), repetition time (TR), hypercellular volume (HCV), volume of interest (VOI), high CBV (hCBV), white matter (WM), gray matter (GM), dynamic susceptibility contrast (DSC)

ABSTRACT

Quantitative mapping of hyperperfused and hypercellular regions of glioblastoma has been proposed to improve definition of tumor regions at risk for local recurrence following conventional radiation therapy. As the processing of the multiparametric dynamic contrast-enhanced (DCE-) and diffusion-weighted (DW-) magnetic resonance imaging (MRI) data for delineation of these subvolumes requires additional steps that go beyond the standard practices of target definition, we sought to devise a workflow to support the timely planning and treatment of patients. A phase II study implementing a multiparametric imaging biomarker for tumor hyperperfusion and hypercellularity consisting of DCE-MRI and high b-value DW-MRI to guide intensified (75 Gy/30 fractions) radiation therapy (RT) in patients with newly diagnosed glioblastoma was launched. In this report, the workflow and the initial imaging outcomes of the first 12 patients are described. Among all the first 12 patients, treatment was initiated within 6 weeks of surgery and within 2 weeks of simulation. On average, the combined hypercellular volume and high cerebral blood volume/tumor perfusion volume were 1.8 times smaller than the T1 gadolinium abnormality and 10 times smaller than the FLAIR abnormality. Hypercellular volume and high cerebral blood volume/tumor perfusion volume each identified largely distinct regions and showed 57% overlap with the enhancing abnormality, and minimal-to-no extension outside of the FLAIR. These results show the feasibility of implementing a workflow for multiparametric magnetic resonance-guided radiation therapy into clinical trials with a coordinated multidisciplinary team, and the unique and complementary tumor subregions identified by the combination of high b-value DW-MRI and DCE-MRI.

INTRODUCTION

Conventional therapies for glioblastoma (GBM) continue to rely on anatomic imaging modalities for both surgery and radiation therapy (RT), including T1 gadolinium- (T1-Gd) enhanced and T2-weighted fluid attenuated inversion recovery (T2-FLAIR) sequences that do not provide biological information about the underlying disease. Multiple studies have shown the prognostic value of physiological magnetic resonance imaging (MRI) techniques such as proton spectroscopy and perfusion and diffusion MRI and measures such as progression-free survival (PFS) and overall survival (OS) in predicting treatment response in patients with GBM (1-9). These imaging techniques may show abnormal tumor infiltration beyond the contrast-enhanced or nonenhanced areas conventionally targeted by surgery and radiation, and these may potentially be used to guide radiation treatment, reduce tumor recurrence, and improve patient outcome (10).

Dynamic contrast-enhanced (DCE)-MRI assesses relative cerebral blood volume (rCBV), cerebral blood flow, and vascular permeability, which are associated with neovascularization and tumor growth and predict PFS and OS in patients with GBM (1, 2, 5, 11). While regions of elevated rCBV often overlap with regions of contrast enhancement, the nonenhancing, infiltrating tumor beyond this region may potentially be underestimated with perfusion MRI (12). In contrast, diffusion-weighted (DW) MRI may identify tumor phenotype by estimating water mobility in the tissue microenvironment as an indicator of tumor

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2018.00035

cellularity (13). Apparent diffusion coefficient (ADC) is inversely correlated with cellularity but it may be unreliable in regions of highly cellular tumor, normal brain tissue, edema, and micronecrosis, yielding elevated ADC compared with normal tissue using standard b-values of 0–1000 s/mm^2 (10). At our center, we developed a novel DW-MRI technique using high b-value (b = 3000 s/mm^2) to selectively isolate solid, often nonenhancing, tumor that is predictive of PFS and often extends beyond the high-dose radiation target (14). We have shown that a combination of these imaging techniques (DCE-MRI and high b-value DW-MRI) into a multiparametric imaging signature predicts PFS with spatial correspondence with patterns of failure, representing biologically high-risk tumor subvolumes identifiable before therapy (12).

Based on these findings, we wished to develop a phase II study to evaluate the feasibility and efficacy of using a multiparametric hypervascular/hypercellular MRI signature to identify areas at highest risk of failure before radiation treatment in patients with newly diagnosed GBM (NCT02805179). Building on a prior phase I/II study showing the safety and efficacy of radiation dose-escalation with concurrent temozolomide (15), this multiparametric advanced imaging technique was used to prospectively guide the boost volume for dose-intensified radiation. To conduct this trial, the development of a workflow was required to permit the integration of an advanced, multiparametric imaging technique into the radiation treatment planning process. Here, we report the workflow and imaging characteristics of the initial patients treated on this prospective clinical trial.

METHODOLOGY

Patient Population

Adult patients of ≥18 years of age with newly diagnosed, pathologically confirmed suptratentorial GBM following any extent of resection were enrolled on this University of Michigan IRB-approved clinical trial following study-specific informed consent. Research was conducted in compliance with the World Medical Association Declaration of Helsinki-Ethical Principles for Medical Research Involving Human Subjects. Eligibility included Karnofsky performance status ≥70, minimal life expectancy of 12 weeks, adequate organ function, and maximal contiguous volume of tumor based on advanced imaging-defined boost volume of <1/3 volume of brain. Patients unable to undergo MRI scans or with prior overlapping radiation therapy were excluded. All patients were treated with standard concurrent daily (75 mg/m^2) and adjuvant monthly (150–200 mg/m^2) temozolomide.

MRI and Computed Tomography Simulation

All patients underwent an MRI simulation and computed tomography (CT) simulation after surgery for radiation planning, within 14 days of commencing chemoradiation. Rigid alignment of the T1-weighted contrast-enhanced and T2 FLAIR MRI to the CT image volumes in the Eclipse image registration workspace was performed by the medical physicist and verified by the radiation oncologist.

Commissioning of Hardware and Software QA

All MRI scans were performed on a 3 T scanner (Skyra, Siemens Healthineers, Erlangen, Germany) in the Radiation Oncology Department. Routine quality assurance of this scanner consists of daily checks of intensity uniformity as well as weekly checks following the ACR phantom accreditation scanning protocol (16). T1 mapping is a critical element of DCE-MRI analysis. To assess the accuracy, repeatability, and interplatform reproducibility of T1 quantification from variable flip angles, we scanned a National Institute of Standards and Technology (NIST) T1 water phantom on our system, provided by our participation in an NCI Quantitative Imaging Network (QIN) multicenter collaborative project (17). We used the extended Tofts model to quantify DCE-MRI, which was implemented in an in-house functional image analysis tool (imFIAT) (18). The performance of our implementation of the extended Tofts model was evaluated using digital reference objects, that is, synthesized DCE phantoms with and without noise, which was fully reported previously (19). In addition, we participated in an NCI QIN multicenter arterial input function (AIF) challenge to validate and compare our AIF delineation procedure to others' (20). On the basis of these evaluations and validations, imFIAT has been granted a level-2 benchmark by the NCI QIN (21).

MRI Acquisition

All images were acquired on a 3 T scanner (Skyra) using a 20-channel head coil. Conventional images, such as 2-dimensional (2D) T2-FLAIR images, and 3-dimensional pre- and post-contrast T1-weighted images, were acquired. In addition, physiological image acquisitions are described in the following subsections.

Diffusion-Weighted Imaging. DW images were acquired using a 2D RESOLVE pulse sequence with diffusion weighting in 3 orthogonal directions and b-values of 0 and 3000 s/mm^2 (1 and 4, respectively) to reduce geometric distortion required for radiation treatment planning. RESOLVE is a multishot technique that uses 2D navigator correction with readout-segmented echo planar imaging (22). Thirty slices were acquired to cover the whole brain with echo time (TE)/repetition time (TR) = 81/7650 milliseconds, matrix size = 160 × 160, and slice thickness/gap = 4.0/1.2 mm for ~4.23 minutes. DW images acquired with b = 3000 s/mm^2 were used for target definition. In addition, DW images were acquired by a 2D echo planar spin echo pulse sequence with diffusion weighting in 3 orthogonal directions and 11 b-values from 0 to 2500 s/mm^2 as a backup scan. Thirty slices were acquired with TE/TR = 93/8200 milliseconds, matrix size = 192 × 192, slice thickness/gap = 4.0/1.2 mm, parallel imaging factor of 4 and a single average for 5 minutes. Parallel imaging factor of 4 was used to reduce the echo training time and thereby reduce geometric distortion. A full characterization of geometric accuracy of DW images with these acquisition parameters was previously reported (14).

DCE Imaging. DCE images were acquired by a 3D gradient echo pulse sequence, called TWIST, in the sagittal orientation to avoid the in-flow effect and ensure artery coverage for an input function delineation. To cover the whole brain, a field of view of 250 × 256 × 187 mm^3 was used with a matrix of 128 × 128 × 104 to obtain an isotropic voxel size of ~1.9, which allows reformatting of the images in an axial plane or other planes as desired. Other acquisition parameters included flip angle = ~10°, TE/TR = ~0.95/2.65 milliseconds, temporal resolution = ~3 s, dynamic phase volumes = 60, and total acquisition

time = 3 minutes. Contrast was injected after 5 dynamic image volumes to achieve sufficient baseline data points.

Acquisition for T1 Quantification. 3D gradient echo images with 4 flip angles (3°, 7°, 12°, and 16°), TE/TR = 2.27/5.34 milliseconds, a voxel size = ~2 mm, and total acquisition time = 1:45 minutes before contrast injection were acquired to quantify native T1. Low spatial resolution B1 maps were acquired to correct flip angle errors in T1 quantification, with an acquisition time of ~12 seconds.

Target Volume Definition and Data Transfer

Physician-defined volumes were delineated in the Eclipse treatment planning system (Varian Medical Systems) directly on the MRI scans. T2/FLAIR abnormality was defined on the FLAIR MRI (FLAIR^GTV). The surgical cavity (Cavity^GTV), residual contrast enhancement (Gd^GTV), and combination of cavity and contrast enhancement (GTV_Low) were delineated on the T1-Gd MRI. Volumes were then exported from the treatment planning system to the image analysis software (functional image analysis tool or imFIAT), for creation of DCE-MRI tumor volumes (high CBV [hCBV]), and high b-value DW-MRI tumor volumes (hypercellular volume [HCV]).

Image Analysis

DCE Analysis. Three-parameter Tofts model was used to quantify the fractional plasma volume (V_p), transfer constant of contrast (K^{trans}), and the fractional volume of extravascular extracellular space (v_e) (23). The model was programed using C++ with a GUI in a functional image analysis tool (imFIAT). A full characterization of performance of software using digital reference objects with a large range of physiological parameters, acquisition parameters, and added Gaussian noise has been previously published (19).

In brief, we used the general assumption that

$$C_t \propto \Delta R_1 \tag{1}$$

where C_t is a contrast concentration in a voxel, and ΔR_1 is a change in longitudinal relaxation rates after and before the contrast injection. If TR × R1 ≪1, which is generally satisfactory for brain normal tissue and tumors,

$$\Delta R_1 \approx \frac{\Delta S}{S_{baselline}} R_{10} \tag{2}$$

where ΔS is a change in gradient echo intensities after and before the contrast injection, $S_{baseline}$ is the averaged baseline gradient echo intensity before contrast injection, and R_{10} is the longitudinal relaxation rate before contrast injection. To obtain an AIF, 20 voxels with maximum intensity differences in a dynamic frame that was 1–2 time frames (~4–7 seconds) before the enhancement peak were delineated. We participated in an NCI QIN challenge project of AIF delineation using our approach and our software (24). The parameters quantified from our AIF were well correlated with others (24).

T1 Calculation. T1 maps were derived by fitting

$$S = S_0 \frac{\sin(\alpha)}{1 - \cos(\alpha)\exp\left(-\frac{TR}{T_{10}}\right)}(1 - \exp(-\mathrm{TR} \times T_{10})) \tag{3}$$

where α is a flip angle, TR is repetition time, and S_0 is margination amplitude, to the 4-flip angle T1-weighted images using Simplex algorithm.

Using equations (1) and (2), AIF, T_{10} and the 3-parameter Tofts model, V_p maps were calculated. Then, hematocrit of 0.45 was used to convert V_p to CBV as CBV = $[V_p/(1 - 0.45) \times 100$ (ml/100 g)], where blood density (1 mL/g) was used.

HCV Delineation

HCV was determined on DW images with b = 3000 s/mm^2. A threshold was used from the normal tissue volume of interest (VOI) that was most contralateral to GBM. To obtain the normal brain VOI, an automated process was used to first extract the brain surface and find the middle line near the central fissure of the brain on T2-weighted images (b = 0). Then, the FLAIR abnormality volume was mirrored to the opposite side of the brain surface through the middle line. To remove CSF influence on the signals from the normal VOI, we remove all voxels with strong CSF signals by classifying CSF on T2-weighted images (b = 0) using fuzzy c-means. The VOI was eroded at least 5 mm from GTV^FLAIR, and had ~600 pixels per slice. Then, voxels within the GTV^FLAIR on each slice were thresholded using mean + 2SD of the intensities in the VOI on the slice to account for DW intensity variations across slices. All these processes are fully automated. If a visual inspection of the normal brain VOI indicated the VOI inadequate, HCV could be recreated after adjusting the normal brain VOI by physician coauthors.

hCBV Delineation

hCBV was delineated on the CBV images with a threshold that was established in a previous study (12). Because normal white matter (WM) and gray matter (GM) have intrinsically different CBV values, the threshold value obtained from an uninvolved contralateral volume would vary depending on the ratio of GM to WM in that volume. We therefore segmented uninvolved contralateral GM in the frontal lobe (which has a higher CBV than uninvolved WM) and defined the hCBV tumor volume as the volume of tumor with CBV >1 SD above GM. This definition resulted in hCBV tumor volumes that predicted PFS and OS (12). Therefore, we used this definition in this pilot study of the clinical trial. This threshold was applied to GTV_Gd with 0–3 mm extension on CBV maps.

Volume Review

HCV and hCBV volumes were reviewed by the physician, neuroradiologist, and MRI physicist. Volumes were slightly edited during central review to remove components outside of the brain parenchyma for both HCV and hCBV, or overlap with blood vessels rather than parenchymal tumor volume for hCBV. Finalized tumor volumes were then imported from imFIAT back into the treatment planning system as binary image volumes associated with the HCV and hCBV image set. Images were automatically registered in the Eclipse Image Registration workspace to the original T1-post-Gd scan and checked by the clinical physicist. The clinical physicist then copied the HCV and hCBV volumes to the CT data set in the treatment planning system.

Image and Volume Registration and Delineation

The physician reviewed the HCV and hCBV volumes in the treatment planning system. CTV and PTV structures were then created as follows: CTV_Low was defined as a 1.7 cm expansion

Table 1. Baseline Patient Characteristics

Clinical Characteristic	No (%)
Median age (range)	65–(51-77)
Male	8 (67%)
Extent of resection	
Gross total resection	6 (50%)
Subtotal resection	4 (33%)
Biopsy	2 (17%)
MGMT methylation status	
Positive	3 (25%)
Negative	9 (75%)
Tumor location	
Frontal lobe	4 (33%)
Temporal lobe	5 (42%)
Parietal lobe	2 (17%)
Occipital lobe	1 (8%)

from GTV_Low delimited by normal anatomic boundaries. PTV_Low was defined 0.3 cm (0.2 cm positioning uncertainty with daily CBCT plus 0.1 cm MRI to CT registration uncertainty) as per institutional standard. For the advanced MRI (HCV and hCBV) boost target volumes, no CTV margin was used. PTV_High was defined as a 0.5 cm expansion from the HCV/hCBV volumes (0.2 cm positioning uncertainty + 0.2 cm RESOLVE DWI uncertainty + 0.1 cm MR to CT registration uncertainty).

PTV_Low was prescribed 60 Gy in 30 fractions and PTV_High was prescribed 75 Gy in 30 fractions using a simultaneous integrated boost technique. Volumetric modulated arc therapy using the Eclipse treatment planning system was used in all cases. The goal was to cover 100% of the target volumes with 95% of the prescribed dose, while maintaining conventional dose limits as utilized on cooperative group trials for high-grade glioma. This included maintaining optic chiasm and optic nerves ≤54 Gy, brainstem surface (ventral 3 mm of brainstem) ≤64 Gy, and brainstem core ≤55 Gy.

RESULTS

Patient Characteristics and Imaging Subvolumes

The initial 12 patients enrolled between September 2016 and June 2017 were included in this analysis. Baseline characteristics of patients are described in Table 1. All patients had IDH1 wild-type tumors by immunohistochemistry. Fifty percent of patients underwent gross total resection, 33% underwent subtotal resection, and the remainder underwent biopsy alone. The workflow for image acquisition, volume delineation and data transfer, and treatment planning is depicted in Figure 1. All patients initiated radiation within 6 weeks of surgery and within 2 weeks of simulation. Advanced volume processing was generally done within 24–36 hours of simulation.

Characteristics and distributions of CBV in normal frontal GM, normal WM, and the hCBV tumor volumes are shown in

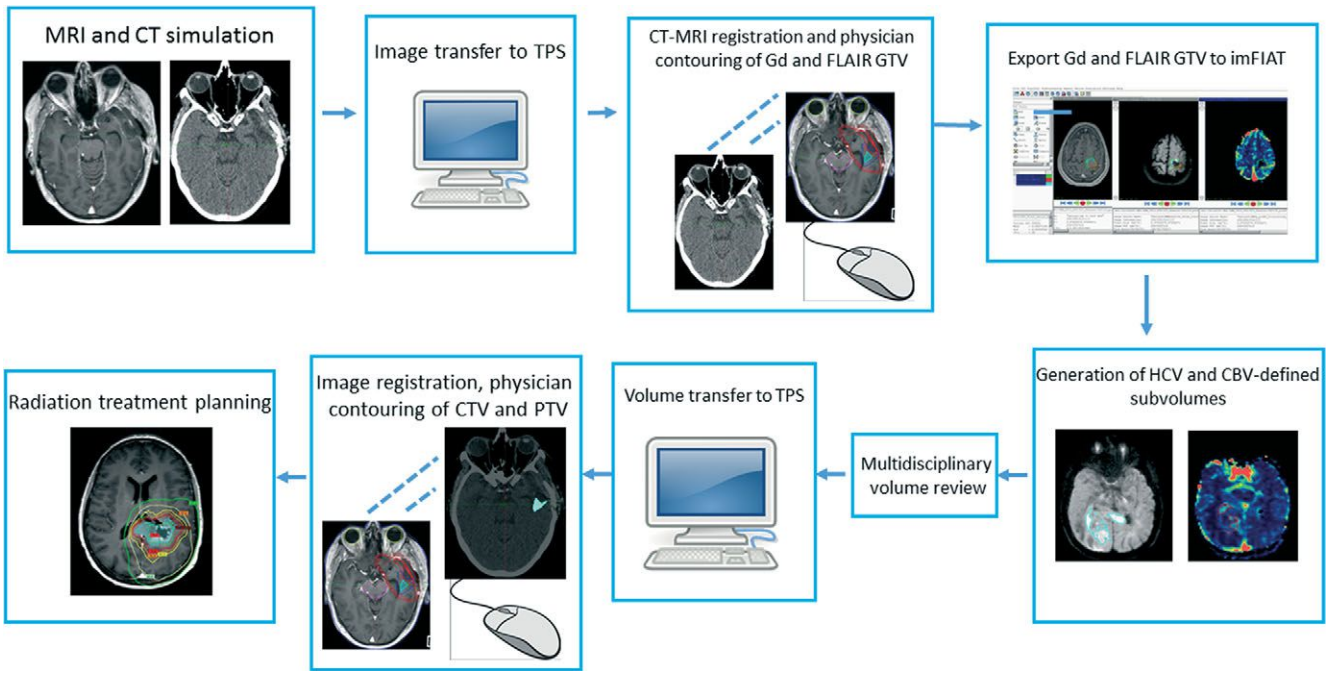


Figure 1. Integrated workflow diagram for the implementation of an advanced dynamic contrast-enhanced (DCE)- and high b-value magnetic resonance (MR) imaging signature to guide dose-intensified radiotherapy. TPS = treatment planning system; Gd = T1-Gd-enhanced MRI; GTV = gross tumor volume; CTV = clinical target volume; PTV = planning target volume.

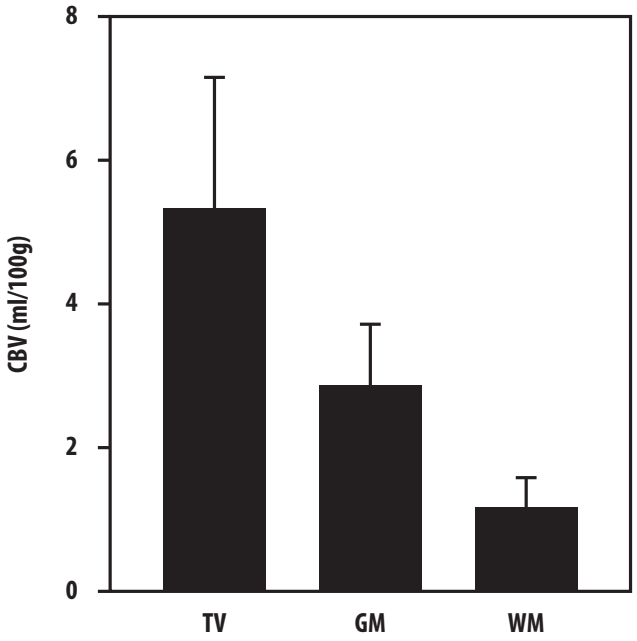


Figure 2. The averaged CBV values in the hCBV tumor volumes, normal frontal white matter (WM) and normal frontal gray matter (GM) in the 12 study patients. The error bars depict the averaged standard deviations of cerebral blood volume (CBV) in the 3 volumes of interest across the 12 patients. Note that the mean CBV value + 2SDs in the frontal WM is smaller than the mean value in the frontal GM, and thus, it cannot be used as a threshold value to define the elevated CBV in the tumor volume.

Figure 2. Note that the CBV value + 2SD in the frontal WM was below the mean value in the normal frontal GM, and thus, it cannot be used to define the elevated CBV in the tumor volume.

Characteristics of conventional and advanced MRI target volumes are listed in Table 2. On average, the Gd-enhanced volume was >3 times larger than those of either HCV or hCBV, and 1.8 times larger than the union (combination) of HCV and hCBV. HCV and hCBV identified largely distinct volumes with only ~1 cc of overlap between the 2 (range, 0.002–6.8 cc). The enhancing component of the union of HCV and hCBV was only 57% (range, 0.3–0.9) of the volume, with an average of 4.1 cc (range, 1.4–7.9) extending outside of the enhancing region. Only 2 patients showed minimal extension of the union HCV and hCBV volume beyond FLAIR (0.06 cc and 0.04 cc, respectively), and FLAIR volumes were ~10 times larger than the union of HCV and hCBV. An example of HCV and hCBV volumes that are largely nonoverlapping overlaid on the corresponding T1-Gd MRI is shown in Figure 3. Two examples of representative radiation plans for 2 different patients are shown in Figure 4. For comparison, example plans without the advanced MRI boost are also depicted. As showed, the advanced

Table 2. Volume and Overlap of Conventional and Advanced Imaging Subvolumes

Target	Mean Volume (cc)	Range
GTV^Gd	23.9	3.9–49.9
GTV^FLAIR	128.9	39.2–248.5
GTV^HCV	7.5	1.7–20.4
GTV^hCBV	6.6	0.5–18.2
Union of HCV and hCBV	13.1	2.3–31.8
Overlap of HCV and hCBV	0.9	0.002–6.8
Overlap Gd and HCV	5.3	0.4–17.4
Overlap of Gd and hCBV	4.5	0.3–15.0
Overlap Gd and Union	8.9	0.9–26.0
Overlap FLAIR and HCV	7.5	1.7–20.4
Overlap FLAIR and hCBV	6.5	0.5–18.2
Overlap FLAIR and Union	13.1	2.3–31.7
HCV outside of Gd	2.2	0.8–3.6
hCBV outside of Gd	2.0	0.0–6.1
Union outside of Gd	4.1	1.4–7.9

GTV^Gd = Gadolinium enhanced target volume; GTV^FLAIR = FLAIR target volume; GTV^HCV = Hypercellular high b-value DW-MRI target volume; GTV^hCBV = Hyperperfused DCE-MRI target volume.

MRI boost volume was generally a smaller tumor subregion contained within the conventional target volume.

DISCUSSION

While the limitations of anatomic MRI for radiation therapy have been reinforced by multiple studies showing that tumor identified by advanced MRI techniques extending outside of conventionally defined volumes predicts patient prognosis independent of T1-Gd, T2-FLAIR, and other clinical factors, advanced imaging techniques have not been incorporated into routine radiation planning (10, 25). In this initial report of a prospective, single-arm phase II trial for patients with newly diagnosed GBM from a single institution, we describe the end-to-end process of delivery of dose-intensified RT to predicted, high-risk tumor subregions identified by multiparametric MRI. The advanced hypercellular and hyperperfusion tumor subvolumes were significantly smaller than the conventionally defined T1-Gd and T2/FLAIR abnormalities standardly targeted for radiation treatment planning, and identified distinct regions that were frequently nonenhancing and therefore excluded from standard radiation boost volume definition. Using physician-defined volumes on conventional T1-Gd and T2-FLAIR images, the semiautomated creation of advanced MR boost volumes was accomplished for real-time planning, yielding successful delivery of advanced imaging-defined dose-intensified RT in all patients beginning within 2 weeks of simulation.

Given the known limitations of conventional MRI for defining tumor extent and predicting outcome in patients with GBM, the use of advanced imaging including perfusion and

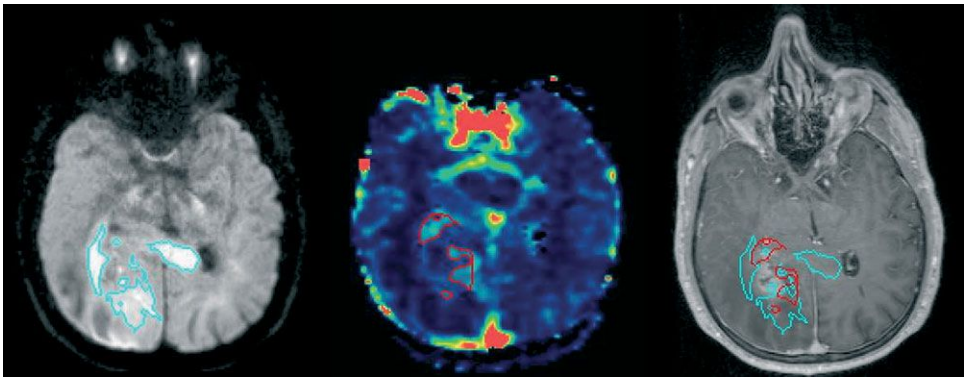


Figure 3. An example of a patient with largely nonoverlapping hypercellular tumor regions (TV_{HCV}) identified by high b-value diffusion-weighted (DW)-magnetic resonance imaging (MRI) (cyan, left panel) and hyperperfused tumor regions (TV_{HCBV}) identified by DCE-MRI (red, middle panel). Significant extension of TV_{HCV} is showed beyond the T1 Gd-enhanced region (overlay on T1 Gd-enhanced image, right panel).

diffusion-weighted MRI has been studied for more than a decade to assess physiologic phenotypes of prognostic significance in this disease. Dynamic susceptibility contrast (DSC) and DCE-MRI permit quantitative estimation of parameters reflective of tumor neovascularization that has been associated with tumor growth in GBM, including CBV, cerebral blood flow, and K^{trans} (23, 26). Maximum CBV and pathologically verified tumor vascularity are correlated, and elevated mean relative CBV (rCBV)

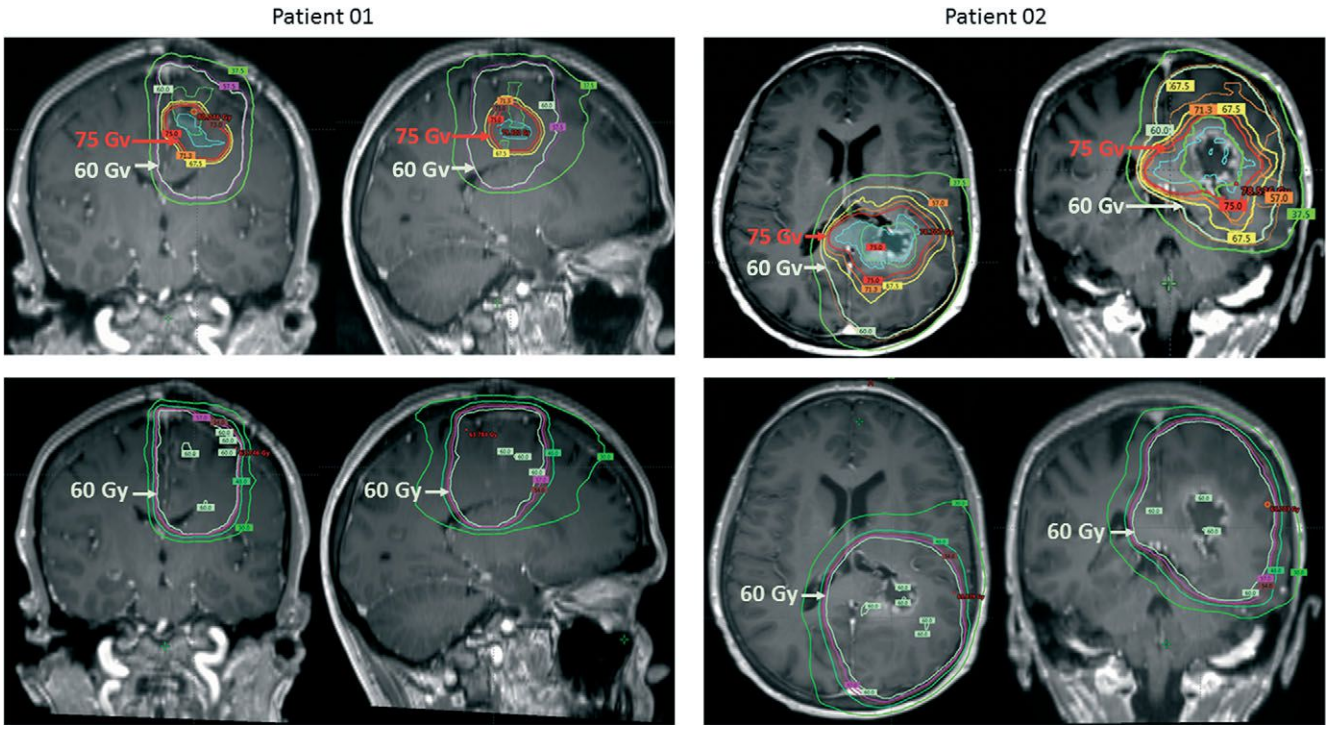


Figure 4. Representative images from radiation plans from 2 different patients. The top row depicts images of radiation plans using advanced MRI to boost tumor subregions to 75 Gy. High-risk tumor targets are identified by advanced MRI (cyan) beyond the abnormal regions seen on T1 Gd-enhanced conventional MRI (green). The conformal 75-Gy isodose line targeting the advanced imaging tumor volume is depicted in red, and the larger 60-Gy isodose line targeting the anatomic T1-Gd-enhanced region with standard clinical margins is depicted in gray-white. The bottom row depicts images of comparison standard radiation plans for the same patients based on anatomic T1-Gd-enhanced MRI with standard clinical margins prescribed to 60 Gy. As showed, advanced MRI-identified boost regions prescribed to 75 Gy were often contained within standard anatomic MRI regions prescribed to 60 Gy.

>1.75 in patients with low- and high-grade gliomas is associated with shorter time to progression (2). Additionally, these perfusion parameters are predictive of overall survival in patients with malignant gliomas (10, 11).

Given the well-known limitation in geometric accuracy of DSC images, we selected T1-weighted DCE images to estimate CBV for the purpose of radiation boost target definition. Several previous studies as well as ours have directly compared CBV values estimated by DSC and DCE images (27-31). An early study reported good correlation of median values in tumors between 2 CBV estimates ($r = 0.67$) and excellent pixel-by-pixel correlation between the 2 estimates in normal brain tissues ($r = 0.96$) in 9 patients with intraaxial cerebral tumors (27). Another study of 32 patients with high-grade glioma showed a weak-but-significant correlation between the 2 estimates in the enhancing tumor volumes in a pixel-by-pixel comparison (28). Another study including 17 healthy subjects and 9 patients with glioblastoma reported excellent correlations of the 2 CBV estimates in normal GM and WM ($r = 0.9$ and $r = 0.89$, respectively) and a good and significant correlation in the tumor ($r = 0.67$) (29). A recent study examined the diagnostic accuracy of glioma grades in 26 patients using the median tumor values of the 2 CBV estimates, which achieved a similar diagnostic accuracy (30). We have performed similar analysis in 20 patients with brain metastases, in which both DCE and DSC images were acquired in a single session with 2 contrast injections. We found good correlation between the 2 CBV estimates in both normal brain tissue and brain tumor volumes ($r = 0.66–0.71$) (unpublished data). These similarities and discrepancies depend upon several factors. The 2 imaging methods rely on considerably different contrast mechanisms and model theories, which can be affected by different physical and physiological conditions. Also, different acquisition parameters and modeling implementations, for example, correcting T1 and vascular leakage effects in DSC analysis, can affect the results.

Additional physiologic properties of malignant gliomas may be assessed with DW-MRI, which has been used to assess the mobility of water molecules in the tissue microenvironment as a surrogate for tumor cellularity, and is a method for therapeutic response assessment as first shown in patients with glioma (6, 7, 13). An inverse correlation is observed between ADC and brain tumor cellularity in preclinical studies (13). However, known limitations of this approach for isolating tumor cellularity from normal brain tissue, edema, and micronecrosis in the heterogeneous GBM microenvironment may lead to unpredictably elevated ADC compared with normal tissue. This limitation may be mitigated through the use of high b-value DW-MRI (3000–4000 s/mm^2) versus 0 and 800–1000 s/mm^2 to attenuate signals due to edema (14, 32-34). We investigated the prognostic value of this approach, showing that the hypercellular tumor region (HCV) identified before RT using high b-value DW-MRI correlates with worse PFS in patients with newly diagnosed GBM treated with standard chemoradiation (14). We determined that in contrast to DCE-MRI alone, the combined use of high b-value MRI with DCE-MRI identifies largely spatially distinct regions with mean overall of only 21% (12). Moreover, the combination of these modalities correlated with patterns of failure and progression, and therefore, these are rationally targeted for intensified radiation treatment (12).

A limited number of studies are prospectively evaluating the incorporation of advanced imaging for the radiation treatment of patients with GBM. These include ongoing studies using proton MR spectroscopic imaging and amino acid positron emission tomography to guide radiation treatment in patients with GBM. Proton MR spectroscopic imaging detects chemical compounds reflective of cellular turnover and proliferation and correlates with histologic tumor cell density and survival in patients with GBM (35-38), although its use for radiation treatment has been limited to select centers with imaging expertise. Amino acid positron emission tomography including ^{11}C-Methionine and ^{18}F-radiolabeled 3,4-dihydroxy-6-[^{18}F]fluoro-L-phenylalanine [^{18}F]F-DOPA tracers is also under evaluation for targeting of potentially aggressive tumor regions beyond conventional MRI in ongoing trials in the United States and Europe. Studies have shown significant correlation in the standard uptake values of ^{11}C-MET and ^{18}F-FDOPA with nearly identical patterns of spatial uptake (39); both have been shown to be prognostic for survival and recurrence and potentially complementary to MRI, although not widely adopted and limited to research centers with expertise in complex radiotracer synthesis or on-site cyclotrons (40-42).

Our study represents the first report of the prospective implementation of a multiparametric imaging signature that is integrated in the RT workflow to guide intensified RT against distinct, poor prognosis phenotypes in patients with GBM. Initial implementation of the real-time use of a multiparametric MR signature for radiation planning involved QA and commissioning of DW- and DCE-MRI for clinical usage before clinical implementation. Implementation of an advanced MR biomarker for radiation treatment requires close coordination between the radiation oncologist, imaging physics, and clinical physics teams to delineate tumor volumes, process and transfer data between treatment planning and advanced imaging software, and ensure timely initiation of treatment. Limitations of this approach include the phenotypic and biologic diversity of GBM, and whether a multiparametric signature is sufficient to characterize this heterogeneity and guide treatment in this disease. To address this, we are acquiring and correlating other physiological imaging modalities with advanced MRI, as well as acquiring longitudinal imaging to evaluate whether temporal changes in advanced imaging features may be used to predict outcome and further tailor therapy.

In this report, we show the feasibility of the real-time use of a multiparametric MR signature to guide radiation treatment against prognostic, unique tumor subregions that substantially differ from the T1-Gd-enhancing high-risk boost volumes standardly defined by conventional MRI. Survival outcomes are awaited from this study, and future directions include translation of this workflow to a second site to validate the generalizability of this novel radiotherapeutic approach for patients with GBM.

ACKNOWLEDGMENTS

This study was funded in part by NIH/NCI grant U01CA183848.

Disclosures: No disclosures to report.

Conflict of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. Cao Y, Tsien CI, Nagesh V, Junck L, Ten Haken R, Ross BD, Chenevert TL, Lawrence TS. Survival prediction in high-grade gliomas by MRI perfusion before and during early stage of RT [corrected]. Int J Radiat Oncol Biol Phys. 2006;64:876–885.
2. Law M, Young RJ, Babb JS, Peccerelli N, Chheang S, Gruber ML, Miller DC, Golfinos JG, Zagzag D, Johnson G. Gliomas: predicting time to progression or survival with cerebral blood volume measurements at dynamic susceptibility-weighted contrast-enhanced perfusion MR imaging. Radiology. 2008;247:490–498.
3. Li X, Jin H, Lu Y, OH J, Chang S, Nelson SJ. Identification of MRI and 1H MRSI parameters that may predict survival for patients with malignant gliomas. NMR Biomed. 2004;17:10–20.
4. Li Y, Lupo JM, Parvataneni R, Lamborn KR, Cha S, Chang SM, Nelson SJ. Survival analysis in patients with newly diagnosed glioblastoma using pre- and postradiotherapy MR spectroscopic imaging. Neuro Oncol. 2013;15:607–617.
5. Hirai T, Murakami R, Nakamura H, Kitajima M, Fukuoka H, Sasao A, Akter M, Hayashida Y, Toya R, Oya N, Awai K, Iyama K, Kuratsu JI, Yamashita Y. Prognostic value of perfusion MR imaging of high-grade astrocytomas: long-term follow-up study. AJNR Am J Neuroradiol. 2008;29:1505–1510.
6. Galban CJ, Hoff BA, Chenevert TL, Ross BD. Diffusion MRI in early cancer therapeutic response assessment. NMR Biomed. 2017;30.
7. Chenevert TL, Stegman LD, Taylor JM, Robertson PL, Greenberg HS, Rehemtulla A, Ross BD. Diffusion magnetic resonance imaging: an early surrogate marker of therapeutic efficacy in brain tumors. J Natl Cancer Inst. 2000;92:2029–2036.
8. Moffat BA, Chenevert TL, Lawrence TS, Meyer CR, Johnson TD, Dong Q, Tsien C, Mukherji S, Quint DJ, Gebarski SS, Robertson PL, Junck LR, Rehemtulla A, Ross BD. Functional diffusion map: a noninvasive MRI biomarker for early stratification of clinical brain tumor response. Proc Natl Acad Sci U S A. 2005;102:5524–5529.
9. Galbán CJ, Chenevert TL, Meyer CR, Tsien C, Lawrence TS, Hamstra DA, Junck L, Sundgren PC, Johnson TD, Ross DJ, Rehemtulla A, Ross BD. The parametric response map is an imaging biomarker for early cancer treatment outcome. Nat Med. 2009;15:572–576.
10. Cao Y, Tseng CL, Balter JM, Teng F, Parmar HA, Sahgal A. MR-guided radiation therapy: transformative technology and its role in the central nervous system. Neuro Oncol. 2017;19(*Suppl_2*):ii16–ii29.
11. Cao Y, Nagesh V, Hamstra D, Tsien CI, Ross BD, Chenevert TL, Junck L, Lawrence TS. The extent and severity of vascular leakage as evidence of tumor aggressiveness in high-grade gliomas. Cancer Res. 2006;66:8912–8917.
12. Wahl DR, Kim MM, Aryal MP, Hartman H, Lawrence TS, Schipper MJ, et al. Combining perfusion and high B-value diffusion MRI to inform prognosis and predict failure patterns in glioblastoma. Int J Radiat Oncol Biol Phys. 2018;102:757–764.
13. Sugahara T, Korogi Y, Kochi M, Ikushima I, Shigematu Y, Hirai T, Okuda T, Liang L, Ge Y, Komohara Y, Ushio Y, Takahashi M. Usefulness of diffusion-weighted MRI with echo-planar technique in the evaluation of cellularity in gliomas. J Magn Reson Imaging. 1999;9:53–60.
14. Pramanik PP, Parmar HA, Mammoser AG, Junck LR, Kim MM, Tsien CI, et al. Hypercellularity components of glioblastoma identified by high b-value diffusion-weighted imaging. Int J Radiat Oncol Biol Phys. 201515;92:811–819.
15. Tsien CI, Brown D, Normolle D, Schipper M, Piert M, Junck L, Heth J, Gomez-Hassan D, Ten Haken RK, Chenevert T, Cao Y, Lawrence T. Concurrent temozolomide and dose-escalated intensity-modulated radiation therapy in newly diagnosed glioblastoma. Clin Cancer Res. 2012;18:273–279.
16. Phantom test guidance for the ACR MRI accreditation program. American College of Radiology; Reston, VA. 1998.
17. Bane O, Hectors SJ, Wagner M, Arlinghaus LL, Aryal MP, Cao Y, Chenevert TL, Fennessy F, Huang W, Hylton NM, Kalpathy-Cramer J, Keenan KE, Malyarenko D, Mulkern RV, Newitt DC, Russek SE, Stupic KF, Tudorica A, Wilmes LJ, Yankeelov TE, Yen YF, Boss MA, Taouli B. Accuracy, repeatability, and interplatform reproducibility of T1 quantification methods used for DCE-MRI: Results from a multicenter phantom study. Magn Reson Med. 2018;79:2564–2575.
18. Cao Y. WE-D-T-6C-03: Development of image software tools for radiation therapy assessment. Med Phys. 2005;32:2136–2136.
19. Cao Y, Li D, Shen Z, Normolle D. Sensitivity of quantitative metrics derived from DCE MRI and a pharmacokinetic model to image quality and acquisition parameters. Acad Radiol. 2010;17:468–478.
20. Huang W, Chen Y, Fedorov A, Li X, Jajamovich GH, Malyarenko DI, Aryal MP, LaViolette PS, Oborski MJ, O'Sullivan F, Abramson RG, Jafari-Khouzani K, Afzal A, Tudorica A, Moloney B, Gupta SN, Besa C, Kalpathy-Cramer J, Mountz JM, Laymon CM, Muzi M, Schmainda K, Cao Y, Chenevert TL, Taouli B, Yankeelov TE, Fennessy F, Li X. The impact of arterial input function determination variations on prostate dynamic contrast-enhanced magnetic resonance imaging pharmacokinetic modeling: a multicenter data analysis challenge. Tomography. 2016;2:56–66.
21. Farahani K, Kalpathy-Cramer J, Chenevert TL, Rubin DL, Sunderland JJ, Nordstrom RJ, Buatti J, Hylton N. Computational challenges and collaborative projects in the NCI Quantitative Imaging Network. Tomography. 2016;2:242–249.
22. Porter DA, Heidemann RM. High resolution diffusion-weighted imaging using readout-segmented echo-planar imaging, parallel imaging and a two-dimensional navigator-based reacquisition. Magn Reson Med. 2009;62:468–475.
23. Tofts PS, Brix G, Buckley DL, Evelhoch JL, Henderson E, Knopp MV, Larsson HB, Lee TY, Mayr NA, Parker GJ, Port RE, Taylor J, Weisskoff RM. Estimating kinetic parameters from dynamic contrast-enhanced T(1)-weighted MRI of a diffusable tracer: standardized quantities and symbols. J Magn Reson Imaging. 1999;10:223–232.
24. Huang W, Chen Y, Fedorov A, Li X, Jajamovich GH, Malyarenko DI, Aryal MP, LaViolette PS, Oborski MJ, O'Sullivan F, Abramson RG, Jafari-Khouzani K, Afzal A, Tudorica A, Moloney B, Gupta SN, Besa C, Kalpathy-Cramer J, Mountz JM, Laymon CM, Muzi M, Schmainda K, Cao Y, Chenevert TL, Taouli B, Yankeelov TE, Fennessy F, Li X. The impact of arterial input function determination variations on prostate dynamic contrast-enhanced magnetic resonance imaging pharmacokinetic modeling: a multicenter data analysis challenge. Tomography. 2016;2:56–66.
25. Cao Y, Sundgren PC, Tsien CI, Chenevert TT, Junck L. Physiologic and metabolic magnetic resonance imaging in gliomas. J Clin Oncol. 2006;24:1228–1235.
26. Rosen BR, Belliveau JW, Aronen HJ, Kennedy D, Buchbinder BR, Fischman A, Gruber M, Glas J, Weisskoff RM, Cohen MS, et al. Susceptibility contrast imaging of cerebral blood volume: human experience. Magn Reson Med. 1991;22:293–299; discussion 300–303.
27. Haroon H, Patankar T, Zhu X, Li K, Thacker N, Scott M, Jackson A. Comparison of cerebral blood volume maps generated from T2* and T1 weighted MRI data in intra-axial cerebral tumours. Br J Radiol. 2014;80:161–168.
28. Alcaide-Leon P, Pareto D, Martinez-Saez E, Auger C, Bharatha A, Rovira A. Pixel-by-pixel comparison of volume transfer constant and estimates of cerebral blood volume from dynamic contrast-enhanced and dynamic susceptibility contrast-enhanced MR imaging in high-grade gliomas. AJNR Am J Neuroradiol. 2015;36:871–876.
29. Artzi M, Liberman G, Nadav G, Vitinshtein F, Blumenthal DT, Bokstein F, Aizenstein O, Ben Bashat D. Human cerebral blood volume measurements using dynamic contrast enhancement in comparison to dynamic susceptibility contrast MRI. Neuroradiology. 2015;57:671–678.
30. Santarosa C, Castellano A, Conte GM, Cadioli M, Iadanza A, Terreni MR, Franzin A, Bello L, Caulo M, Falini A, Anzalone N. Dynamic contrast-enhanced and dynamic susceptibility contrast perfusion MR imaging for glioma grading: preliminary comparison of vessel compartment and permeability parameters using hotspot and histogram analysis. Eur J Radiol. 2016;85:1147–1156.
31. Bazyar S, Ramalho J, Eldeniz C, An H, Lee YZ. Comparison of cerebral blood volume and plasma volume in untreated intracranial tumors. PloS One. 2016;11:e0161807.
32. Mardor Y, Roth Y, Ochershvilli A, Spiegelmann R, Tichler T, Daniels D, Maier SE, Nissim O, Ram Z, Baram J, Orenstein A, Pfeffer R. Pretreatment prediction of brain tumors' response to radiation therapy using high b-value diffusion-weighted MRI. Neoplasia. 2004;6:136–142.
33. Mardor Y, Pfeffer R, Spiegelmann R, Roth Y, Maier SE, Nissim O, Berger R, Glicksman A, Baram J, Orenstein A, Cohen JS, Tichler T. Early detection of response to radiation therapy in patients with brain malignancies using conventional and high b-value diffusion-weighted magnetic resonance imaging. J Clin Oncol. 2003;21:1094–1100.

34. Chu HH, Choi SH, Ryoo I, Kim SC, Yeom JA, Shin H, Jung SC, Lee AL, Yoon TJ, Kim TM, Lee SH, Park CK, Kim JH, Sohn CH, Park SH, Kim IH. Differentiation of true progression from pseudoprogression in glioblastoma treated with radiation therapy and concomitant temozolomide: comparison study of standard and high-b-value diffusion-weighted imaging. Radiology. 2013;269: 831–840.
35. McKnight TR, von dem Bussche MH, Vigneron DB, Lu Y, Berger MS, McDermott MW, Dillon WP, Graves EE, Pirzkall A, Nelson SJ. Histopathological validation of a three-dimensional magnetic resonance spectroscopy index as a predictor of tumor presence. J Neurosurg. 2002;97:794–802.
36. Vigneron D, Bollen A, McDermott M, Wald L, Day M, Moyher-Noworolski S, Henry R, Chang S, Berger M, Dillon W, Nelson S. Three-dimensional magnetic resonance spectroscopic imaging of histologically confirmed brain tumors. Magn Reson Imaging. 2001;19:89–101.
37. Dowling C, Bollen AW, Noworolski SM, McDermott MW, Barbaro NM, Day MR, Henry RG, Chang SM, Dillon WP, Nelson SJ, Vigneron DB. Preoperative proton MR spectroscopic imaging of brain tumors: correlation with histopathologic analysis of resection specimens. AJNR Am J Neuroradiol. 2001;22: 604–612.
38. Croteau D, Scarpace L, Hearshen D, Gutierrez J, Fisher JL, Rock JP, Mikkelsen T. Correlation between magnetic resonance spectroscopy imaging and image-guided biopsies: semiquantitative and qualitative histopathological analyses of patients with untreated glioma. Neurosurgery. 2001;49:823–829.
39. Becherer A, Karanikas G, Szabo M, Zettinig G, Asenbaum S, Marosi C, Henk C, Wunderbaldinger P, Czech T, Wadsak W, Kletter K. Brain tumour imaging with PET: a comparison between [18F]fluorodopa and [11C]methionine. Eur J Nucl Med Mol Imaging. 2003;30:1561–1567.
40. Bell C, Dowson N, Puttick S, Gal Y, Thomas P, Fay M, Smith J, Rose S. Increasing feasibility and utility of (18)F-FDOPA PET for the management of glioma. Nucl Med Biol. 2015;42:788–795.
41. Lee IH, Piert M, Gomez-Hassan D, Junck L, Rogers L, Hayman J, Ten Haken RK, Lawrence TS, Cao Y, Tsien C. Association of 11C-methionine PET uptake with site of failure after concurrent temozolomide and radiation for primary glioblastoma multiforme. Int J Radiat Oncol Biol Phys. 2009;73:479–485.
42. Albert NL, Weller M, Suchorska B, Galldiks N, Soffietti R, Kim MM, et al. Response assessment in neuro-oncology working group and European Association for Neuro-Oncology recommendations for the clinical use of PET imaging in gliomas. Neuro Oncol. 2016;18:1199–1208.

Gleason Probability Maps: A Radiomics Tool for Mapping Prostate Cancer Likelihood in MRI Space

Sean D. McGarry[1], John D. Bukowy[1], Kenneth A. Iczkowski[2], Jackson G. Unteriner[1], Petar Duvnjak[1], Allison K. Lowman[1], Kenneth Jacobsohn[3], Mark Hohenwalter[1], Michael O. Griffin[1], Alex W. Barrington[1], Halle E. Foss[1], Tucker Keuter[4], Sarah L. Hurrell[1], William A. See[3], Marja T. Nevalainen[5,6], Anjishnu Banerjee[4], and Peter S. LaViolette[1,7]

Departments of [1]Radiology, [2]Pathology, [3]Urological Surgery, [4]Biostatistics, [5]Radiation Oncology, [6]Pharmacology, and [7]Biomedical Engineering, Medical College of Wisconsin, Milwaukee, WI

Corresponding Author
Peter S. LaViolette, PhD
Department of Radiology and Biomedical Engineering, Medical College of Wisconsin, 8701 Watertown Plank Rd., Milwaukee, WI 53226;
E-mail: plaviole@mcw.edu.

Key Words: prostate Cancer, radiomics, rad-path, radio-pathomics
Abbreviations: Magnetic resonance imaging (MRI), prostate-specific antigen (PSA), multiparametric MRI (MP-MRI), dynamic contrast-enhanced (DCE), field of view (FOV), receiver operator characteristic (ROC), area under the curve (AUC)

ABSTRACT

Prostate cancer is the most common noncutaneous cancer in men in the United States. The current paradigm for screening and diagnosis is imperfect, with relatively low specificity, high cost, and high morbidity. This study aims to generate new image contrasts by learning a distribution of unique image signatures associated with prostate cancer. In total, 48 patients were prospectively recruited for this institutional review board–approved study. Patients underwent multiparametric magnetic resonance imaging 2 weeks before surgery. Post-surgical tissues were annotated by a pathologist and aligned to the in vivo imaging. Radiomic profiles were generated by linearly combining 4 image contrasts (T2, apparent diffusion coefficient [ADC] 0-1000, ADC 50-2000, and dynamic contrast-enhanced) segmented using global thresholds. The distribution of radiomic profiles in high-grade cancer, low-grade cancer, and normal tissues was recorded, and the generated probability values were applied to a naive test set. The resulting Gleason probability maps were stable regardless of training cohort, functioned independent of prostate zone, and outperformed conventional clinical imaging (area under the curve [AUC] = 0.79). Extensive overlap was seen in the most common image signatures associated with high- and low-grade cancer, indicating that low- and high-grade tumors present similarly on conventional imaging.

INTRODUCTION

Prostate cancer is the most frequently diagnosed noncutaneous cancer in men in the United States, accounting for ~1 in 5 new cancer diagnoses (1). Increased screening efforts, early aggressive therapy for high-risk disease, and the relatively indolent nature of the disease in most patients have resulted in an overall 5-year survival rate of 99% for organ-confined prostate cancer (1). The current paradigm for prostate cancer diagnosis centers on obtaining tissue diagnosis before definitive therapy, either through conventional 12-core systematic transrectal ultrasound-guided biopsy systems or newer magnetic resonance imaging (MRI)-fusion targeted biopsies. These data are typically combined with clinical information (ie, prostate-specific antigen [PSA], PSA density, and clinical T stage) and implemented into various nomograms to predict disease "risk" status.

While PSA screening has been shown to reduce mortality (2-4), PSA alone has relatively low specificity for prostate cancer diagnosis and is insufficient in stratifying disease risk status, leading to an abundance of low-risk patients undergoing an invasive biopsy (5). The conventional paradigm for prostate cancer diagnosis and staging has been challenged in recent years, with data showing that nontargeted biopsies can lead to under-sampling, inaccurate risk stratification, or missing the target cancer all together (6, 7). As a result, noninvasive imaging with multiparametric MRI (MP-MRI) of the prostate is increasingly being used as a tool for prostate cancer detection, preoperative staging, active surveillance, targeted biopsy, and guidance for definitive focal therapy. Several recent prospective trials have shown that using MP-MRI in the prebiopsy setting to identify target lesions for targeted biopsy outperforms systematic 12-core biopsy, leading to a higher rate of diagnosis for clinically significant cancers and a fewer clinically insignificant cancers (8, 9). The incorporation of MR-guidance, however, requires a radiologist to identify and label potential targets.

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2018.00033

With the clinical standard of care shifting toward image-guided biopsies, an increased burden will be placed on radiologists to correctly and efficiently identify prostate tumors for biopsy.

A typical clinical prostate MRI protocol includes T2-weighted imaging to delineate structure and zone anatomy, multi b-value diffusion-weighted imaging to identify areas of diffusion restriction, and dynamic contrast-enhanced (DCE) imaging to identify early or contemporaneous focal enhancement. The exams are interpreted by a radiologist according to the PI-RADS v2 (10) that assesses the probability of clinically significant cancer. The lack of specificity inherent to PSA screening and MP-MRI means that a definitive diagnosis still requires a biopsy procedure which often leads to the overtreatment of low-risk prostate cancer (11). Patients who do not necessarily have cancer undergo invasive biopsy procedures to mitigate the uncertainty in the screening tests. The clinical barrier to overcome is to appropriately stratify patients near the boundary between intervention and active surveillance before biopsy. The line between active surveillance and treatment is histologically Gleason 3 vs Gleason 4, which roughly correlates to PI-RADS 3 vs PI-RADS 4 on MP-MRI.

Computer-aided diagnostic tools informed by postoperative tissue may provide an opportunity to address this clinical barrier (12-14). Predictive models made from aligned whole mount tissue and in vivo imaging provide opportunity to bring additional information into image space, increasing the overall specificity of a nonspecific test. Radiomics and machine-learning based approaches have been a great success over the past decade (15-18) and improved sensitivity yet decreasing the specificity. However, there is a need for further improvement of this technology.

Radiomics provides a framework for quantifying tumor microenvironment by analyzing images as a mineable database. In addition, by creating a database of aligned pathology or genetics with clinical imaging, it becomes feasible to find radiologic patterns of tumor phenotype which may provide critical predictive information (19-21). Radiomics-based approaches have seen success over the past decade, proving successful across modalities (22, 23) and organ systems, (24-26) by providing a useful means for engineering features amenable to machine learning approaches.

This study uses an aligned rad-path data set to determine whether unique imaging signatures predict the presence of pathologist-identified prostate cancer. We present a method which learns a distribution of unique image characteristics associated with histologic annotations to create voxel-wise predictive maps on a naïve test set.

METHODS

Patient Population

Forty-eight patients were recruited prospectively for this institutional review board (IRB)–approved study between June 2014 and February 2017. Written consent was obtained from all patients. Patients' age ranged from 45 to 71 years (mean, 60 years). Inclusion criteria for this study included a scheduled radical prostatectomy and clinical imaging with additional high *b* value DWI 2 weeks before surgery.

Imaging

MP-MRI was acquired on a 3 T MRI scanner (General Electric, Waukesha, WI) using an endorectal coil. MP-MRI included field of view (FOV)– optimized and –constrained undistorted single shot (FOCUS) DWI, DCE imaging, and T2-weighted imaging. T2 acquisition parameters were as follows: repetition time (TR) = 3370 milliseconds, FOV = 120 mm, voxel dimensions = 0.23 × 0.23 × 3 mm, acquisition matrix = 512, and slices = 26. Diffusion images were collected with 10 b-values (b = 0, 10, 25, 50, 80, 100, 200, 500, 1000, 2000). The DCE images were collected during injection of a gadolinium contrast agent with acquisition matrix = 256, slices = 25, and FOV = 120 mm. All image contrasts used in this study were acquired axially.

MRI Preprocessing

The T2-weighted images were intensity-normalized to the standard deviation within a manually drawn prostate mask (26-29). The B = 0 image was aligned to the T2 using FLIRT (30, 31) and corrected manually if necessary using a freesurfer tool, tkregister2 (surfer.nmr.mgh.harvard.edu). ADC was calculated from 2 combinations of b-value for the purposes of this study, 0 and 1000 and 50 and 2000 (32). The DCE volume with maximal contrast influx was identified using a custom algorithm programmed in Matlab (MathWorks Inc., Natick, MA) and manually aligned to the T2-weighted image using tkregister2. The DCE was intensity-normalized as described above for the T2-weighted images.

Tissue Processing

Following surgery, prostate samples were sectioned using patient-specific tissue slicing molds created from the presurgical T2 images, as previously published (29, 32). Surface models were created from the manually drawn prostate mask using 3D slicer and subtracted from a template-slicing mold matching the T2 slice spacing using Blender. The slicing molds were then 3D-printed using a fifth-generation Makerbot. Tissue sections were paraffin-embedded, whole-mounted, and stained with hematoxylin-eosin. Slides were digitally scanned using a microscope equipped with an automated stage (Nikon Metrology, Brighton MI). The digitized histology was annotated by a fellowship-trained urologic pathologist (KAI) using color codes corresponding to the Gleason grading system. Annotations were drawn on a Microsoft Surface Pro 4 using a predefined color palette. The annotations were saved as a mask overlaid on the high-resolution histology.

Tissue Alignment

Digitized samples were aligned to the T2-weighted images using custom software previously published (29, 32). Control points were manually placed on analogous points on both the histology and the MRI. A nonlinear transform was then calculated to warp the histology into T2 space using the imwarp command in the Matlab image processing toolbox (MathWorks Inc.). The annotations from our pathologist were likewise transformed into T2 space using the same transform and a nearest-neighbor interpolation. The pathologist annotations in MRI space are referred to as "deep annotations" throughout the manuscript.

Stratification by Tumor Volume

A 3-fold cross-validation approach was chosen to test generalizability. A custom sampling algorithm was required to distribute patients into cohorts with balanced tumor burden. Patient-wise tumor burden was calculated as the sum of the number of pathologist-annotated high grade (Gleason 4-5) and low grade (Gleason 3) voxels. Stratification by high grade and low grade was chosen in lieu of individual Gleason grade owing to the relatively limited amount of grade 5 tumor in the data set. The ideal cohort tumor burden was calculated as the total tumor burden divided by the number of cohorts. Random permutations of patients were created, and the difference between the randomly generated tumor burden and the ideal tumor burden was calculated. A separate error metric is calculated for each cohort (n) and then summed to produce a single error score for that permutation as seen in equation 1:

$$Error = \sum_{n=1}^{3} \left| 1 - \frac{Actual\ Tumor\ Burden}{Ideal\ Tumor\ Burden} \right| \times Cost \qquad (1)$$

The cost function was empirically set at 3:1, favoring high-grade tumors; this results in a much larger error score if the high-grade tumor volume is unbalanced. There were a larger number of low-grade annotations than high-grade ones in the data set, and balancing high-grade tumor volume was deemed more important than balancing low grade. The permutation producing the lowest error in 10,000 iterations was used as the group assignment for this study. There are approximately 10^{21} possible combinations, thus sampling 10,000 limits bias but provides relatively balanced cohorts.

Global Thresholding and Segmentation

Each of the 3 cohorts was used as a test set for an algorithm trained on the other 2 cohorts. Three sets of global thresholds were established; for each cohort a global threshold was established using the 2 unused cohorts (32 patients). The contrasts used in this study were ADC (b = 0 and b = 1000), ADC (b = 50 and b = 2000), T2, and DCE. Global thresholds were created using Otsu's method calculated on all voxels in the manually drawn prostate masks for the entire training cohort of 32 patients; thresholds applied to the test set were not generated from these data (33). Global thresholding tests the assumption that all cohorts represent the same probability density function and additionally removes the constraint that each patient expresses each unique image characteristic. Images were segmented using the calculated thresholds into dark, intermediate, and bright intensities represented by values of 1, 2, and 3 respectively.

Generation of Radiomic Profiles

A unique code was created for each voxel by linearly combining the segmented image contrasts. Radiomic profiles were created by multiplying the segmented contrasts by ascending powers of 10 such that each digit represents the segmentation value of that individual contrast. With 4 image contrasts with 3 color values each a total of 81 radiomic profiles are possible; a voxel encoded with 1133 contains dark ADC_{short}, dark ADC_{long}, bright T2, and bright DCE. A schematic demonstrating the generation of the radiomic profiles is shown in Figure 1 (26).

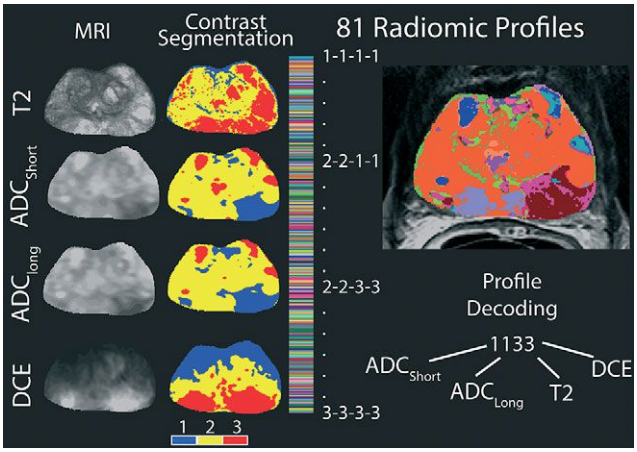


Figure 1. Generation of Radiomic profiles. Left: The 4 contrasts included in this study and the resulting segmentations created using Otsu's method. Right: 81 unique image characteristics created by linearly combining the segmented image contrasts. Each voxel receives a 4-digit code representative of the segmented image contrasts. Code 1133 indicates dark ADCshort, dark ADClong, bright T2, and bright DCE.

Training Set Independence

To determine the training set independence, imaging from an additional 5 patients not included in the previous analysis was processed. Three sets of radiomic profiles were generated using each of the 3 sets of thresholds calculated prior. Each pixel within the prostate mask was analyzed to quantify the overlap among the 3 sets of images. Pixels where the radiomic profile was identical across all images were labelled 3, and pixels where only 2 images matched were labelled as 2. A high overlap score (a large percentage of voxels with a value of 2 or 3) would indicate stable performance regardless of training set.

Gleason Probability Maps Generation

A probability table was generated by analyzing the distribution of each unique image signature within the pathologist-annotated regions. The distribution of each profile in the training set is recorded for low grade, high grade, and benign atrophy (ie, profile 1111 contains benign atrophy 75% of the time). The probability distribution is then propagated to the test set, where each profile is replaced by its respective percentage value, creating 4 maps depicting low grade, high grade, benign, and cancer likelihood. Figure 2 shows the generation of the probability table and Gleason probability maps.

Zone Dependence

The imaging signatures of prostate cancer are known to be zone-dependent. Lesions are evaluated via the PI-RADS scale using primarily the T2-weighted images in the transition zone and DWI in the peripheral zone. To test the robustness of the Gleason probability maps to tumor location, additional probability tables were created stratified by zone (ie, profile

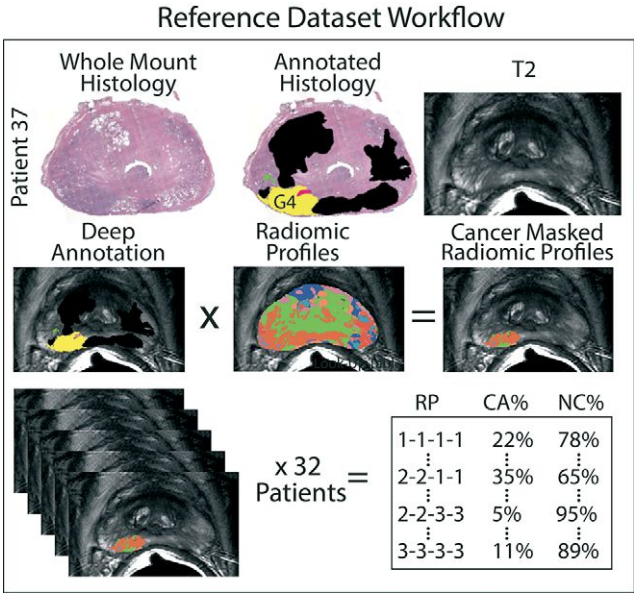


Figure 2. Generation of Gleason probability maps from a training data set. Top: Hematoxylin and eosin stained whole mount histology and the corresponding pathologist annotations and T2 slice. Middle: Radiomic profiles are masked by the pathologist annotations and the distribution of the radiomic profiles. Bottom: The distribution of radiomic profiles within high grade, low grade, and benign regions are analyzed over 32 patients. The resulting probability values are applied to the radiomic profiling images on naïve data to create Gleason probability maps.

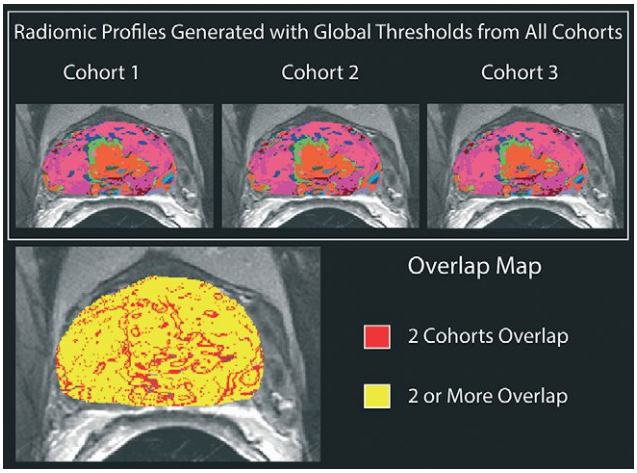


Figure 3. Top: Radiomic profiles generated with global thresholds calculated on 3 different training sets, applied to a patient not otherwise included in the analysis. Bottom: Overlap map: yellow pixels have an identical image signature across all cohorts, red *and yellow* pixels have identical image signatures in 2 cohorts. Blank pixels have no overlap.

3333 contains high-grade tumor in 3% of the voxels located in the peripheral zone and 0% of the voxels in the transition zone) and applied to the test set. These new maps were evaluated using a receiver operator characteristic (ROC) and compared to the nonzone-dependent maps.

Evaluation of Gleason Probability Maps and Comparison to Clinical Imaging

The Gleason probability maps were evaluated lesion-wise using a ROC analysis. High grade was compared to all other tissue, cancer to all other tissue, and high grade to low grade. In addition, the intensity normalized image contrasts were evaluated for their ability to distinguish high-grade cancer from all other tissue and high grade from low grade.

RESULTS

Patient Stratification

Patients were stratified pseudorandomly, attempting to match an empirically determined ideal tumor burden using a custom sampling algorithm. The high-grade tumor and low-grade tumor burdens were, on average, 7.3% and 19.4% off the calculated ideal tumor burden.

Training Set Independence

Radiomic profile maps generated using thresholds from 3 patients cohorts were compared for overlap on data from 5 patients not included in the study. The individual radiomic profile maps and the overlap map can be seen in Figure 3. At least 2 cohorts had identical radiomic profile values on 98.6% of the pixels in the additional subject (red and yellow areas), and all 3 cohorts agreed on 76.3% of voxels (yellow only).

Zone Independence

The zone dependence of the technique was tested using a ROC. The resulting AUCs can be seen in Table 1. The algorithms' performance was nearly identical when zone information was included; however, the addition of zone information requires manually drawn zone masks. Figure 4 shows the high-grade Gleason probability map on the same patient with and without zone information included.

Table 1. Comparison of ROC AUC in Gleason Probability Maps Made With and Without the Inclusion of Zone Information in the Probability Table

	Zones	No Zones
High Grade vs All	0.76	0.77
Cancer vs All	0.79	0.77
Normal vs All	0.79	0.79

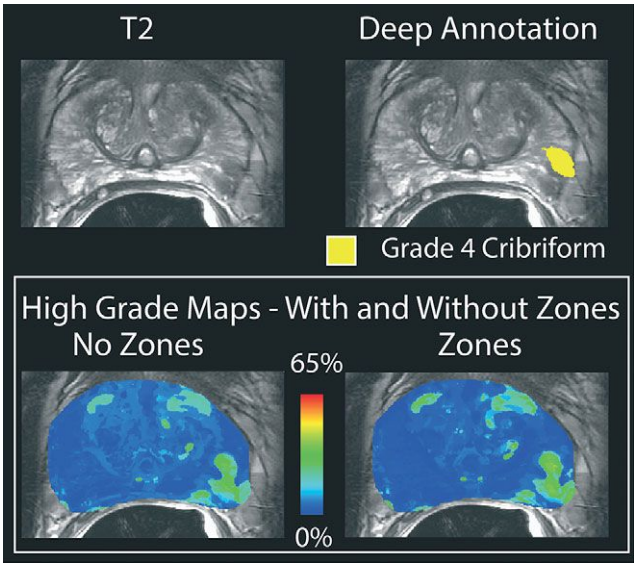


Figure 4. Top: T2-weighted image and deep annotation overlaid on the same slice. A grade 4 cribriform tumor is shown in yellow. Bottom: Gleason probability maps created with and without the inclusion of zone information in the training data set. The images are nearly identical and the tumor is clearly visible.

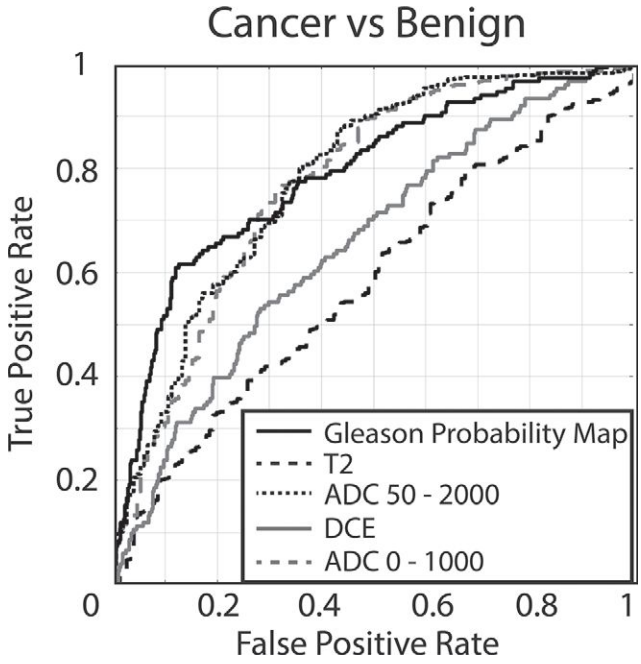


Figure 5. Receiver operator characteristic (ROC) evaluating the performance of the 4 raw image contrasts compared to Gleason probability maps (cancer probability). ADC 50-2000 = 0.78, ADC 0-1000 = 0.77, DCE = 0.65, T2 = 0.57, Gleason probability map = 0.79.

Receiver Operator Characteristic

The resulting ROCs can be seen in Table 2 along with the difference between the 2 conditions. The AUC of the clinical contrasts alone ranged from 0.55 to 0.78. The Gleason probability maps achieved an AUC of 0.79, distinguishing cancer from benign atrophy. Figure 5 shows ROCs comparing the Gleason probability maps to the raw image contrasts used in the analysis. Figure 6 shows the resulting Gleason probability maps on 3 true-positive and 1 true-negative case.

Image Signatures Unique to Low- and High-Grade Tumors

The 10 most common profiles by volume seen in both high- and low-grade cancer can be seen in Table 3. Seven of the top 10 most frequently seen profiles in high-grade cancer also frequently appear in low-grade cancer. Approximately 13,000 voxels containing high-grade tumor are explained by profile 1122, making it the most frequently seen high-grade image signature. That profile is the fifth most common low-grade profile, but the volume is nearly identical at 12,000 voxels, resulting in a similar probability that a voxel containing 1122 in the test set is high grade or low grade.

The similarity in image characteristics between high- and low-grade tumors occurs independent of lesion size. Normal and benign regions, on average, exhibit image signature 2222. Lesions less than 200 contiguous voxels (<12.5 mm^2 in-plane) exhibit an identical image signature—these lesions are indistinguishable from normal tissue. Large lesions exhibit profile 1122 regardless of final Gleason grade.

Table 2. Comparison of ROC AUC in Gleason Probability Maps and Clinical Image Contrasts

	Cancer vs Benign	High Grade vs Low Grade
T2	0.58	0.53
ADC 0-1000	0.77	0.58
ADC 50-2000	0.78	0.60
DCE	0.65	0.51
Gleason Probability Map	0.79	0.56

Both cancer versus benign and high grade vs low grade were tested.

DISCUSSION

This study translated a technique developed as a risk stratification model in glioblastoma (26) to identify unique image signatures associated with prostate cancer. The technique outperforms the diagnostic capacity of each of the clinical images individually (Figure 5) and brings histologic data in the form of a learned probability distribution of unique image signatures into image space on naive data. Patients were successfully stratified pseudo-randomly into cohorts with roughly equivalent volumes of high- and low-grade prostate cancer. Gleason probability mapping produces nearly identical results independent of training cohort and functions without requiring zone information.

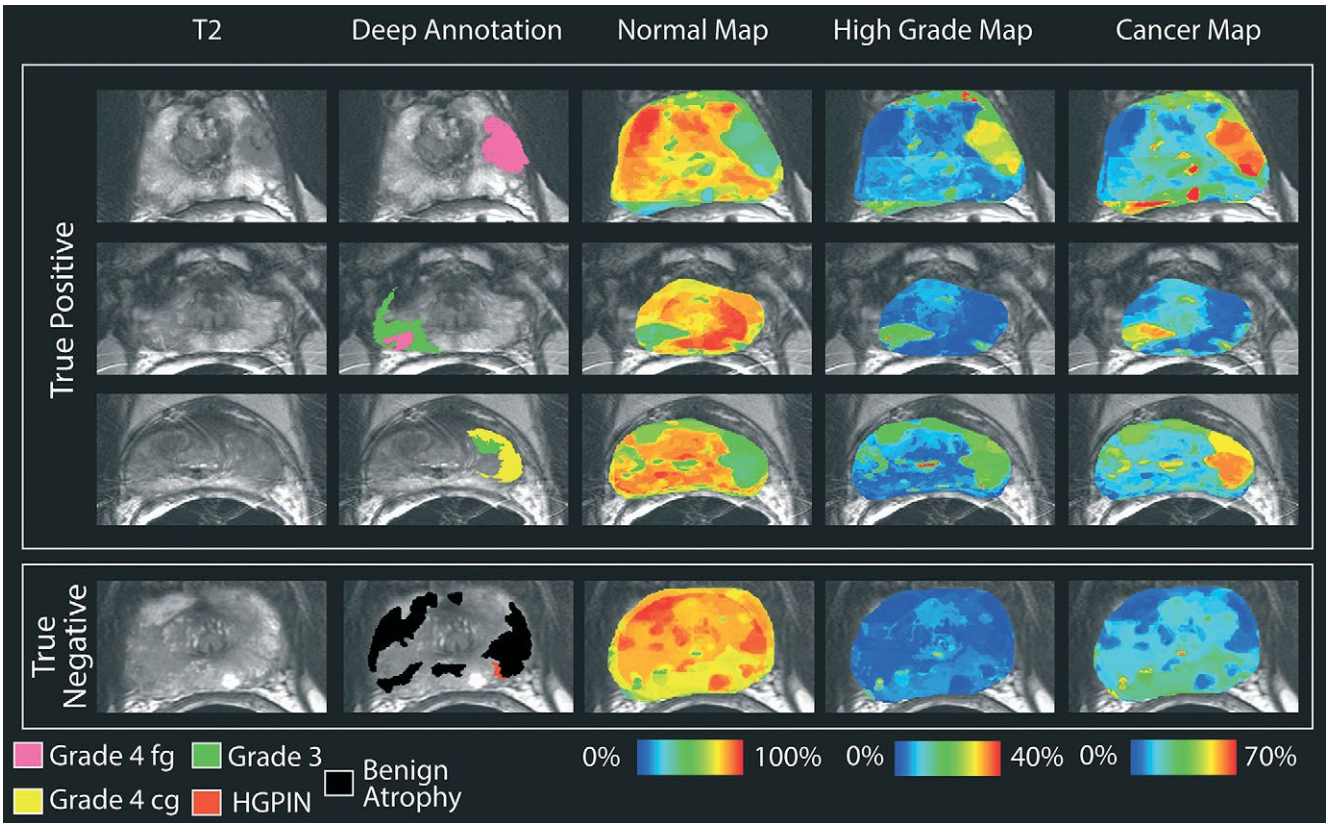


Figure 6. Gleason probability maps. Top: True-positive cases. High-grade tumors are shown on the deep annotation in pink (cribriform) and yellow (not cribriform). Low-grade tumors are shown in green. Images are scaled to reflect the maximum probability in the training data set. Bottom: True negative. The displayed slide has only benign atrophy, and thus, no hot spots occur in the Gleason probability maps.

Table 3. Top 10 Most Common Radiomic Profiles in High- and Low-Grade Lesions Ordered by Volume

Low Grade		High Grade	
Volume	**Profile[a]**	**Volume**	**Profile**
22 179	2222	13 572	1122
17 900	*1123*	12 146	1112
16 302	*1112*	9871	1132
13 609	*2212*	8470	1123
13 060	*1122*	8356	1111
12 651	*2223*	8323	1121
9159	*1111*	7800	2211
8972	1121	5873	2221
8780	2221	5452	1133
8712	*2211*	5157	2222

[a] Profiles that are common between the two are shown in italics on the low-grade profile list.

Other prostate radiomics approaches have seen success in detecting prostate cancer using MP-MRI (34, 35). These tools rely either on confirmed pathologically radiologist ROIs or aligned, annotated whole mount slides and frequently combine first-order histogram features with texture and volume features to create a single risk score rather than applying global thresholds. Many volume- and histogram-based features require an *a priori* ROI in the test set, whereas pixel-wise approaches can provide both disease severity and detection. These tools have trended toward aligned whole mount histology which allows voxel-wise predictions and eliminates the need for a radiologist to manually draw ROIs. Reported AUC varies by technique and ranges from 0.76 to 0.99; however, all published techniques distinguish cancer (G3+) from benign or healthy tissue. Notably, to the best of our knowledge, no technique has been published to date that is capable of distinguishing G3 and G4 patterned lesions, which is where the clinical decision is often made, as it pertains to choosing active surveillance versus definitive therapy. The method introduced in this study, Gleason probability mapping, likewise performs poorly at distinguishing high-grade tumor from low-grade tumor, likely because of the limitations of the imaging techniques themselves. Future studies focused specifically on differentiating G3 from G4+ need to occur.

Transition zone lesions are identified primarily by the T2-weighted images in PI-RADS v2 because benign prostatic hyperplasia nodules often exhibit diffusion restriction and early/contemporaneous enhancement (mimicking cancer) but appear morphologically different from significant cancers, which may not be captured fully by a segmentation-based method. Currently, no prostate CAD tool reads a prostate exam like a radiologist—techniques are either contrast-based and well suited to identify peripheral zone lesions or texture-based and well suited to identify transition zone lesions. It is plausible that the most effective method of identifying prostate tumors distinguished by zone and uses vastly different techniques depending on the lesions location.

There are known sources of error associated with rad-path studies that may reduce accuracy. Our sample includes patients with cancer: other confounding diseases are unlabeled and may thus contribute to error. While we have previously validated our control point–warping technique, there is still error involved in the process. This technique used global thresholds generated from 1 set of acquisitions on similar magnets. Future studies should validate these thresholds on acquisitions from magnets by other manufacturers. This study is limited to endorectal coil images, but future studies may quantify the generalizability to a population imaged without an endorectal coil.

CONCLUSIONS

Gleason probability mapping stratifies cancer tissue from normal prostate tissue independent of zone and training set. The technique performs better than traditional image contrasts alone and provides a voxelwise map which may be potentially useful for biopsy guidance and reading clinical scans. Additional research is necessary to further classify regions of tumor among the different Gleason patterns.

ACKNOWLEDGMENTS

We would like to thank the patients who graciously participated in our study and Eugenia Westfall for helpful conversations.

This study was funded by The State of Wisconsin Tax Check-off Program for Prostate Cancer Research, RO1CA218144, R01CA113580, National Center for Advancing Translational Sciences, NIH UL1TR001436 and TL1TR001437, 1R21CA231892-01.

Disclosures: No disclosures to report.

Conflict of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J Clin. 2018; 68:7–30.
2. Hugosson J, Carlsson S, Aus G, Bergdahl S, Khatami A, Lodding P, Pihl CG, Stranne J, Holmberg E, Lilja H. Mortality results from the Goteborg randomised population-based prostate-cancer screening trial. Lancet Oncol. 2010;11:725–732.
3. Schroder FH, Roobol MJ. Defining the optimal prostate-specific antigen threshold for the diagnosis of prostate cancer. Curr Opin Urol. 2009;19:227–231.
4. Schroder FH, van den Bergh RC, Wolters T, van Leeuwen PJ, Bangma CH, van der Kwast TH, Roobol MJ. Eleven-year outcome of patients with prostate cancers diagnosed during screening after initial negative sextant biopsies. Eur Urol. 2010;57:256–266.
5. Loeb S, Bruinsma SM, Nicholson J, Briganti A, Pickles T, Kakehi Y, Carlsson SV, Roobol MJ. Active surveillance for prostate cancer: a systematic review of clinicopathologic variables and biomarkers for risk stratification. Eur Urol. 2015;67: 619–626.
6. Mufarrij P, Sankin A, Godoy G, Lepor H. Pathologic outcomes of candidates for active surveillance undergoing radical prostatectomy. Urology. 2010;76:689–692.
7. Serefoglu EC, Altinova S, Ugras NS, Akincioglu E, Asil E, Balbay MD. How reliable is 12-core prostate biopsy procedure in the detection of prostate cancer? Can Urol Assoc J. 2013;7:E293–E298.
8. Ahmed HU, El-Shater Bosaily A, Brown LC, Gabe R, Kaplan R, Parmar MK, Collaco-Moraes Y, Ward K, Hindley RG, Freeman A, Kirkham AP, Oldroyd R, Parker C, Emberton M; PROMIS study group. Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. Lancet. 2017;389:815–822.
9. Kasivisvanathan V, Emberton M, Moore CM. MRI-targeted biopsy for prostate-cancer diagnosis. N Engl J Med. 2018;379:589–590.
10. Vargas HA, Hotker AM, Goldman DA, Moskowitz CS, Gondo T, Matsumoto K, Ehdaie B, Woo S, Fine SW, Reuter VE, Sala E, Hricak H. Updated prostate imaging reporting and data system (PIRADS v2) recommendations for the detection of clinically significant prostate cancer using multiparametric MRI: critical evaluation using whole-mount pathology as standard of reference. Eur Radiol. 2016;26: 1606–1612.
11. Rosenkrantz AB, Taneja SS. Prostate MRI can reduce overdiagnosis and overtreatment of prostate cancer. Acad Radiol. 2015;22:1000–1006.
12. Chatterjee A, Bourne RM, Wang S, Devaraj A, Gallan AJ, Antic T, Karczmar GS, Oto A. Diagnosis of prostate cancer with noninvasive estimation of prostate tissue composition by using hybrid multidimensional MR imaging: a feasibility study. Radiology. 2018;287:864–873.
13. Litjens GJ, Elliott R, Shih NN, Feldman MD, Kobus T, Hulsbergen-van de Kaa C, Barentsz JO, Huisman HJ, Madabhushi A. Computer-extracted features can distinguish noncancerous confounding disease from prostatic adenocarcinoma at multiparametric MR imaging. Radiology. 2016;278:135–145.
14. Ozer S, Langer DL, Liu X, Haider MA, van der Kwast TH, Evans AJ, Yang Y, Wernick MN, Yetik IS. Supervised and unsupervised methods for prostate cancer segmentation with multispectral MRI. Med Phys. 2010;37:1873–1883.
15. Niaf E, Lartizien C, Bratan F, Roche L, Rabilloud M, Mege-Lechevallier F, Rouviere O. Prostate focal peripheral zone lesions: characterization at multiparametric MR imaging–influence of a computer-aided diagnosis system. Radiology. 2014;271:761–769.
16. Viswanath S, Bloch BN, Rosen M, Chappelow J, Toth R, Rofsky N, Lenkinski R, Genega E, Kalyanpur A, Madabhushi A. Integrating structural and functional imaging for computer assisted detection of prostate cancer on multi-protocol *in vivo* 3 Tesla MRI. Proc SPIE Int Soc Opt Eng. 2009;7260:72603I.
17. Vos PC, Hambrock T, Hulsbergen-van de Kaa CA, Futterer JJ, Barentsz JO, Huisman HJ. Computerized analysis of prostate lesions in the peripheral zone using dynamic contrast enhanced MRI. Med Phys. 2008;35:888–899.
18. Viswanath S, Bloch BN, Chappelow J, Patel P, Rofsky N, Lenkinski R, Genega E, Madabhushi A. Enhanced multi-protocol analysis via intelligent supervised embedding (EMPrAvISE): detecting prostate cancer on multi-parametric MRI. Proc SPIE Int Soc Opt Eng. 2011;7963:79630U.
19. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, Bussink J, Monshouwer R, Haibe-Kains B, Rietveld D, Hoebers F, Rietbergen MM, Leemans CR, Dekker A, Quackenbush J, Gillies RJ, Lambin P. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:4006.
20. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology. 2016;278:563–577.
21. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, Zegers CM, Gillies R, Boellard R, Dekker A, Aerts HJ. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48:441–446.
22. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJ. Machine learning methods for quantitative radiomic biomarkers. Sci Rep. 2015;5:13087.

23. Vallieres M, Freeman CR, Skamene SR, El Naqa I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. Phys Med Biol. 2015;60:5471–5496.
24. Cameron A, Khalvati F, Haider MA, Wong A. MAPS: a quantitative radiomics approach for prostate cancer detection. IEEE Trans Biomed Eng. 2016;63:1145–1156.
25. Kickingereder P, Burth S, Wick A, Gotz M, Eidel O, Schlemmer HP, Maier-Hein KH, Wick W, Bendszus M, Radbruch A, Bonekamp D. Radiomic profiling of glioblastoma: identifying an imaging predictor of patient survival with improved performance over established clinical and radiologic risk models. Radiology. 2016; 280:880–889.
26. McGarry SD, Hurrell SL, Kaczmarowski AL, Cochran EJ, Connelly J, Rand SD, Schmainda KM, LaViolette PS. Magnetic resonance imaging-based radiomic profiles predict patient prognosis in newly diagnosed glioblastoma before therapy. Tomography. 2016;2:223–228.
27. Ellingson BM, Kim HJ, Woodworth DC, Pope WB, Cloughesy JN, Harris RJ, Lai A, Nghiemphu PL, Cloughesy TF. Recurrent glioblastoma treated with bevacizumab: contrast-enhanced T1-weighted subtraction maps improve tumor delineation and aid prediction of survival in a multicenter clinical trial. Radiology. 2014; 271:200–210.
28. Ellingson BM, Zaw T, Cloughesy TF, Naeini KM, Lalezari S, Mong S, Lai A, Nghiemphu PL, Pope WB. Comparison between intensity normalization techniques for dynamic susceptibility contrast (DSC)-MRI estimates of cerebral blood volume (CBV) in human gliomas. J Magn Reson Imaging. 2012;35:1472–1477.
29. McGarry SD, Hurrell SL, Iczkowski KA, Hall W, Kaczmarowski AL, Banerjee A, Keuter T, Jacobsohn K, Bukowy JD, Nevalainen MT, Hohenwalter MD, See WA, LaViolette PS. Radio-pathomic maps of epithelium and lumen density predict the location of high-grade prostate cancer. Int J Radiat Oncol Biol Phys. 2018;101: 1179–1187.
30. Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage. 2002;17:825–841.
31. Jenkinson M, Smith S. A global optimisation method for robust affine registration of brain images. Med Image Anal. 2001;5:143–156.
32. Hurrell SL, McGarry SD, Kaczmarowski A, Iczkowski KA, Jacobsohn K, Hohenwalter MD, Hall WA, See WA, Banerjee A, Charles DK, Nevalainen MT, Mackinnon AC, LaViolette PS. Optimized b-value selection for the discrimination of prostate cancer grades, including the cribriform pattern, using diffusion weighted imaging. J Med Imaging (Bellingham). 2018;5:011004.
33. Otsu N. A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern. 1979;9:62–66.
34. Smith CP, Czarniecki M, Mehralivand S, Stoyanova R, Choyke PL, Harmon S, Turkbey B. Radiomics and radiogenomics of prostate cancer. Abdom Radiol. 2018. [Epub ahead of print].
35. Stoyanova R, Takhar M, Tschudi Y, Ford JC, Solorzano G, Erho N, Balaguruna-than Y, Punnen S, Davicioni E, Gillies RJ, Pollack A. Prostate cancer radiomics and the promise of radiogenomics. Transl Cancer Res. 2016;5:432–447.

Multiparameter MRI Predictors of Long-Term Survival in Glioblastoma Multiforme

Olya Stringfield[1], John A. Arrington[2,7], Sandra K. Johnston[5,6], Nicolas G. Rognin[3], Noah C. Peeri[4], Yoganand Balagurunathan[3], Pamela R. Jackson[5], Kamala R. Clark-Swanson[5], Kristin R. Swanson[5], Kathleen M. Egan[4,7], Robert A. Gatenby[2,7], and Natarajan Raghunand[3,7]

[1]IRAT Shared Service, Departments of [2]Diagnostic & Interventional Radiology, [3]Cancer Physiology, and [4]Cancer Epidemiology, Moffitt Cancer Center, Tampa, FL; [5]Mathematical NeuroOncology Lab, Precision Neurotherapeutics Innovation Program, Mayo Clinic, Phoenix, AZ; [6]Department of Radiology, University of Washington, Seattle, WA; and [7]Department of Oncologic Sciences, University of S Florida, Tampa, FL

Corresponding Author:
Natarajan Raghunand, PhD
Moffitt Cancer Center, 12902 Magnolia Drive,
Tampa, FL 33612, USA;
E-mail: Natarajan.Raghunand@moffitt.org

Key Words: glioblastoma, survival, MRI, habitats, cancer evolution

Abbreviations: Magnetic resonance imaging (MRI), glioblastoma multiforme (GBM), fluid-attenuated inversion-recovery (FLAIR), contrast-enhancing (CE), nonenhancing (NE), dynamic susceptibility contrast-enhanced (DSC), unenhanced T1-weighted (T1W), contrast-enhanced T1-weighted (T1W-CE), long-term survival (LTS), short-term survival (STS), T2-weighted (T2W), volume of interest (VOI), relative cerebral blood volume (rCBV), T1W-CE − T1W difference volume (ΔT1W)

ABSTRACT

Standard-of-care multiparameter magnetic resonance imaging (MRI) scans of the brain were used to objectively subdivide glioblastoma multiforme (GBM) tumors into regions that correspond to variations in blood flow, interstitial edema, and cellular density. We hypothesized that the distribution of these distinct tumor ecological "habitats" at the time of presentation will impact the course of the disease. We retrospectively analyzed initial MRI scans in 2 groups of patients diagnosed with GBM, a long-term survival group comprising subjects who survived >36 month postdiagnosis, and a short-term survival group comprising subjects who survived ≤19 month postdiagnosis. The single-institution discovery cohort contained 22 subjects in each group, while the multi-institution validation cohort contained 15 subjects per group. MRI voxel intensities were calibrated, and tumor voxels clustered on contrast-enhanced T1-weighted and fluid-attenuated inversion-recovery (FLAIR) images into 6 distinct "habitats" based on low- to medium- to high-contrast enhancement and low–high signal on FLAIR scans. Habitat 6 (high signal on calibrated contrast-enhanced T1-weighted and FLAIR sequences) comprised a significantly higher volume fraction of tumors in the long-term survival group (discovery cohort, 35% ± 6.5%; validation cohort, 34% ± 4.8%) compared with tumors in the short-term survival group (discovery cohort, 17% ± 4.5%, $P < .03$; validation cohort, 16 ± 4.0%, $P < .007$). Of the 6 distinct MRI-defined habitats, the fractional tumor volume of habitat 6 at diagnosis was significantly predictive of long- or short-term survival. We discuss a possible mechanistic basis for this association and implications for habitat-driven adaptive therapy of GBM.

INTRODUCTION

Glioblastoma multiforme (GBM) typically exhibits substantial intratumoral heterogeneity at both microscopic and radiological spatial scales (1). Analysis of genomic patterns from The Cancer Genome Atlas (TCGA) database led to a general molecular model that identified 4 distinct "species" of GBM: proneural, neural, classical, and mesenchymal (2). However, more recent studies (3) found substantial spatial variations, so that, in some cases, all 4 species could be observed in different regions of the same tumor. Canoll et al. used RNA-sequencing and histological analysis of image-guided biopsies to show differences in cellular and molecular markers between tissue taken from the contrast-enhancing (CE) core and that from the nonenhancing (NE) margins of GBM tumors (4). Characteristic metabolic differences between the CE and NE regions in GBM have also been identified by 1H magnetic resonance spectroscopy (5). Machine learning on patterns in standard brain magnetic resonance imaging (MRI) images, and parameter maps from diffusion tensor imaging, and dynamic susceptibility contrast-enhanced (DSC)-MRI have been reported to correlate with molecular subtype and survival in newly diagnosed patients with GBM (6). Radiogenomic analysis informed by spatially localized biopsies has identified spatially complex distributions of molecularly distinct subpopulations in GBMs (7). Although such spatial variations in expression of molecular and pathologic markers, metabolism, and radiologic imaging patterns are known to exist in all solid tumors, the origin and the clinical significance of this heterogeneity remain subjects of investigation.

Heterogeneity within tumors may drive resistance to both untargeted and targeted therapies (8). Reliance on conventional

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2018.00052

maximum tolerated dose–based treatment regimens may accelerate the unopposed proliferation of resistant populations by eliminating the susceptible populations and the attendant competition for space and substrate. Enriquez-Navas et al. recently showed that an evolution-based adaptive therapeutic strategy that exploits such competition between subpopulations of tumor cells could prolong progression-free survival in preclinical models of breast cancer (9). An ongoing clinical trial in prostate cancer (10) has shown that evolutionary dynamics can be successfully integrated into clinical cancer treatment protocols, and it highlighted the unmet need for noninvasive metrics of intratumoral subpopulation changes during treatment.

In the present work, we build upon a conceptual model of GBMs as spatially heterogeneous complex adaptive systems in which tumor growth and response to therapy are governed by eco-evolutionary interactions between the tumor microenvironment and phenotypic properties of local cellular populations. This model posits an explicit and predictable link between macroscopic tumor features observed radiologically and the molecular-, cellular-, and tissue-scale properties of the underlying cancer cell populations. In this model, we hypothesize that radiologically apparent spatial heterogeneity within each GBM can be quantified by some combination of a small number of distinct eco-evolutionary "habitats," each of which may have different patterns of growth and invasion and may respond differently to therapy (11). Our approach builds upon methods developed in landscape ecology to bridge spatial scales. For example, field biologists are often tasked with estimating species distribution within a large area such as a county or state. Methods developed in landscape ecology typically begin with an analysis of satellite imagery of the region. By combining image channels containing nonoverlapping information (RADAR, infrared and visible light, for example), the biologist can divide the whole region into a patchwork collection of distinct habitats. By sampling the species distribution within each distinct habitat, the geographic distribution of each species over the entire region can be estimated (12, 13).

Multispectral clustering on MRI images has been used before to quantify spatial variations within tumors. Vannier et al. recognized the analogy between multispectral remote-sensing satellite imagery and multiparametric MRI and showed that signatures for "scene components" in the radiologic images could be computed (14-16). This approach can be used to further objectively subdivide the tumor itself into spatially distinct subregions ("habitats") that harbor distinct subpopulations of tumor cells (11, 17, 18). Spatial heterogeneity of GBMs at radiological scales presents as regional variations in contrast enhancement and edema, and we have used multispectral clustering to decompose each glioma into a small number of distinct "habitats" based on their intensity on different MRI sequences. Tumor voxels were clustered by the calibrated signal intensities on contrast-enhanced T1-weighted (T1W-CE) and fluid-attenuated inversion-recovery (FLAIR) sequences into 6 distinct "habitats" based on low- to medium- to high-contrast enhancement and low–high signal on FLAIR scans. The long-term survival (LTS) cohort (>36 months postdiagnosis) were found to have a significantly higher fraction of habitat 6 (high CE and high FLAIR signal intensity) compared with the short-term survival (STS) cohort (≤ 19 months postdiagnosis) in both the discovery and validation cohorts. We discuss a possible mechanistic basis for this association between habitat 6 and survival in GBM, and implications for habitats-driven adaptive therapy of GBM.

MATERIALS AND METHODS

Discovery Cohort

In this work, we have used the terms "discovery" (or training) and "validation" as they are understood in the field of machine learning, namely, to refer to the specific steps of training–validation–test in model development (19). Following IRB approval, patients with pathologically confirmed primary GBM and available preoperative T2-weighted (T2W), FLAIR, unenhanced T1W, and T1W-CE scans were identified retrospectively from a single participating institution. Median survival in glioblastoma is reported to be between 12 and 18 months postdiagnosis (20, 21). Recent estimates of 5-year survival rates for patients receiving maximal safe resection, concurrent radiotherapy and chemotherapy, and adjuvant chemotherapy are ~10% (22). Our original intent was to investigate MRI habitats in high-grade gliomas from subjects who survived >5 years postdiagnosis. However, after application of the additional requirement that certain MRI scans be available at diagnosis, we had to downgrade this criterion to >3-year survival postdiagnosis of GBM so as to form cohorts with reasonable numbers of subjects. Thus, an LTS group comprising 22 subjects who survived >36 months postdiagnosis (median survival, 62.6 months; range, 36–107 months) was created. A control STS group of 22 subjects who survived <19 months postdiagnosis (median survival, 11.6 months; range, 2.5–19 month) was created to individually match to LTS subjects on age and calendar year of diagnosis.

Validation Cohort

Following IRB approval, patients with pathologically confirmed primary GBM and available preoperative T2W, FLAIR, T1W, and T1W-CE MRI scans were identified retrospectively from a multiinstitutional database, matching on age and sex. The LTS group included 15 subjects who survived >36 months postdiagnosis (median survival, 86.6 months; range, 39–177 months), while the STS group included 15 subjects who survived ≤19 months postdiagnosis (median survival, 12.6 moths; range, 1.8–19 months).

Patient Population Statistics

Additional demographic and clinical covariates of relevance to this study are shown in Table 1.

Image Registration

For each patient, the FLAIR, T1W, and T1W-CE images were coregistered with the T2W images using in-house MATLAB (MathWorks, Natick, MA) software (top panel in Figure 1). As part of this process, the FLAIR, T1W, and T1W-CE images were resampled to match pixel dimensions and slice thicknesses with the reference T2W images. Spatial alignment was performed using a combination of rigid and affine geometrical transformations.

Tumor Segmentation

In this work, we restricted our analysis of intratumoral "habitats" to the CE portion of the tumor volumes. For this purpose, a

Table 1. Demographic and Clinical Characteristics of Patients in the Discovery and Validation Cohorts According to LTS and STS Status

Characteristics	LTS	STS
	Discovery Cohort	
	(N = 22)	**(N = 22)**
Median Age (years)	50.5 (range: 22–74)	50.5 (range: 28–72)
Percent Male	59.1	63.6
Percent College Graduate[a]	45.5	23.8
Median KPS Score[a]	90%	80%
Median Year Diagnosed	2010	2011
Percent Completed Stupp Protocol[b]	37	0
Median Survival (Months)	67.7 (range: 36–126)	11.5 (range: 2.5–19)
	Validation Cohort	
	(N = 15)	**(N = 15)**
Median Age (years)	50 (range: 23–68)	62 (range: 23–78)
Percent Male	67	60
Median Education (years)	Unknown	Unknown
Median KPS Score	90[c]	90[d]
Median Year Diagnosed	2009	2009
Percent Completed Stupp Protocol	66.7	26.7
Median Survival (months)	86.6 (range: 39–177)	12.6 (range: 1.8–19)

[a] 1 STS missing education; 3 LTS and 6 STS missing KPS score.
[b] As defined in PubMed PMID: 15758009. Results based on 20 LTS and 16 STS patients with complete information on receipt of the chemoradiation protocol. A total of 7 patients underwent biopsy as the only form of surgery (1 LTS and 6 STS).
[c] 10 missing values.
[d] 11 missing values.

contour was manually drawn to circumscribe the CE tumor in all applicable slices on postregistration T1W-CE images (middle panel in Figure 1).

Intensity Calibration

The next step in our image processing pipeline was intensity calibration (middle panel of Figure 1), the objective of which is to allow comparison of voxel intensities across patients on each given type of MRI scan. For this purpose, 2 reference normal tissue regions were automatically segmented as shown in Figure 2. In brief, intensities within the T1W-CE − T1W difference volume (ΔT1W) were clustered into low- and high-intensity classes using Otsu thresholding (23). Then, on T2W, voxels from the low-intensity class were subdivided further into low- and high-intensity clusters using Otsu thresholding. Voxels from the low cluster formed a volume of interest (VOI) that was applied to T1W, which was subdivided into low- and high-intensity clusters by Otsu thresholding, with the resulting voxels in the high-intensity class labeled as "normal white matter" (reference region 1). Voxels from the high T2W cluster formed a VOI mask that was applied to the FLAIR scan, and these were again subdivided into low- and high-intensity clusters using Otsu thresholding, and the low-intensity cluster was labeled as "CSF" (reference region 2). Voxel intensities on T2W, FLAIR, and precontrast T1W images were then linearly calibrated using "normal white matter" and "CSF" as reference tissues. The reference intensity values for these 2 tissues, respectively, were 81 and 183 on T2W, 587 and 464 on FLAIR, and 1099 and 748 on precontrast T1W, all in arbitrary units. These reference values were taken from the T2W, FLAIR, and precontrast T1W images of a patient chosen randomly from the discovery cohort, and do not carry any particular physiological meaning as such. Intensity calibration for T1W-CE was performed using the same linear transformation as computed for the associated precontrast T1W. Our input data comprise standard-of-care MRI images that were acquired with varying protocols per subject. Acquisition parameters such as the repetition time, echo time, and flip angle were not the same across all subjects for each scan type (T2W, FLAIR, T1W). Because MRI signal intensity is a nonlinear function of these acquisition parameters, linear calibration against 2 reference tissues may not necessarily be adequate for standardization of intensities per scan type. Fortunately, the range of excursions in these acquisition parameters across subjects was relatively small, and signal equation simulations indicated that calibration of raw signal intensity against 2 dissimilar reference tissues would provide satisfactory intensity calibration for other tissues with T1 and T2 values similar to or in-between those of the 2 reference tissues. The coefficient of variation of normal gray matter intensity across all patients was significantly smaller postcalibration as compared with precalibration on each of FLAIR, T1W, and T1W-CE images, and we took this to be evidence of successful intensity calibration (see online Supplemental Figure 1).

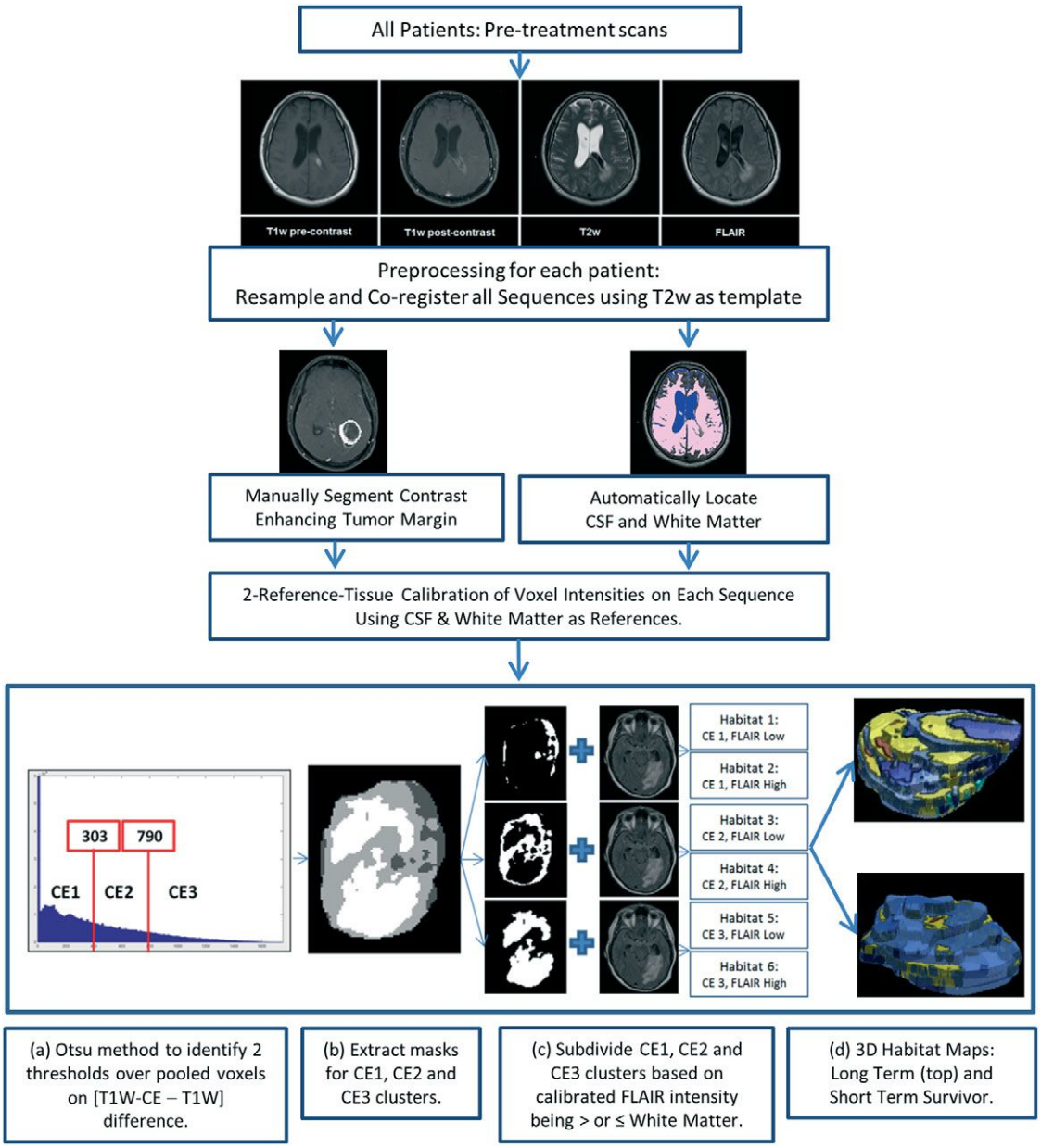


Figure 1. Fluid-attenuated inversion-recovery (FLAIR), T1-weighted (T1W), and contrast-enhanced T1-weighted (T1W-CE) images were coregistered with and resampled to match voxel dimensions in the reference T2W scans (top panel). A contour was manually drawn to circumscribe the CE tumor in all applicable slices on postregistration T1W-CE images (middle panel). Normal white matter and cerebral spinal fluid (CSF) were automatically segmented (middle panel, details in Figure 2). Voxel intensities were calibrated against white matter (WM) and CSF to permit cluster analysis of voxels pooled across patients on each type of magnetic resonance imaging (MRI) scan (middle panel). Pooled voxels from within the CE tumor mask were clustered into 6 habitats using the criteria listed in Table 1 (bottom panel). Also shown in the bottom panel is a 3D stack of maps of habitats 1–6 in an example tumor, for illustrative purposes.

Multispectral Clustering to Define Intratumoral Habitats

Calibration of intensities per MRI scan type allows us to pool voxels over multiple patients for combined cluster analysis. This series of steps is depicted in the bottom panel of Figure 1. In brief, the manually drawn CE tumor mask was applied to the calibrated ΔT1W difference volume of each patient in the discovery cohort, and the voxels within the mask were pooled over all subjects and clustered by Otsu thresholding into 3 levels of contrast enhancement: CE1 (low enhancement), CE2 (medium enhancement), and CE3 (high enhancement). The low-, medium- and high-contrast enhancement thresholds identified on the discovery cohort were refined on validation, specifically that the maximum value of ΔT1W difference intensity was capped at 5000 arbitrary units postcalibration before Otsu thresholding.

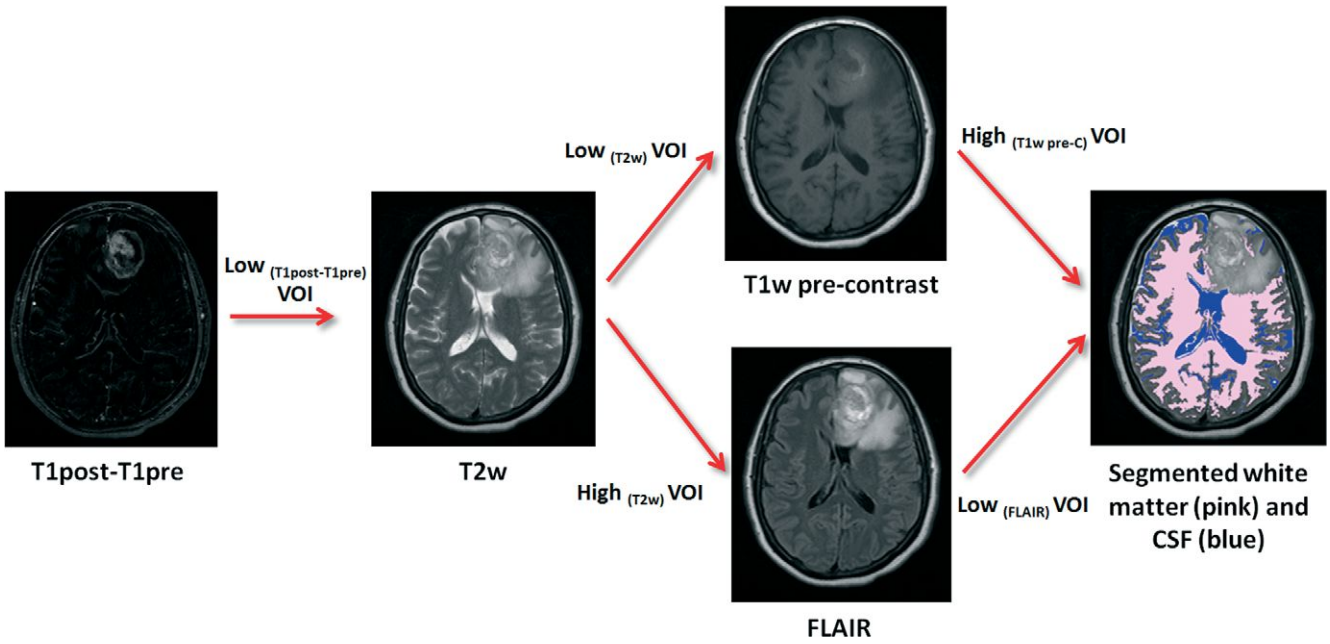


Figure 2. Automatic segmentation procedure to locate WM and CSF volumes within the brain for use in intensity calibration. Intensities within the T1W-CE − T1W difference volume (ΔT1W) of a given subject were clustered into low- and high-intensity classes by Otsu thresholding. A mask of voxels in the low-intensity class was applied to the T2W image and further subdivided into low- and high-intensity clusters by Otsu thresholding. The resulting mask of voxels in the low-intensity cluster was applied to the T1W image, which was again subdivided into low- and high-intensity clusters with the high-intensity class labeled as "normal white matter" (reference region 1). The mask of high-intensity voxels from the T2W image was applied to the FLAIR image, and it was again subdivided into low- and high-intensity clusters with the resulting low-intensity cluster labeled as "CSF" (reference region 2).

This was done to manage the skewing of the clustering process by a long 1-sided tail on the ΔT1W difference intensity histogram in some patients. Each of these 3 clusters was further subclustered into 2 classes around a calibrated value of 600 on FLAIR, a threshold value that is similar to the mean intensity of normal white matter over all subjects after calibration. The final habitat definitions are listed in Table 2.

Statistics and Survival Analyses

Absolute tumor volume, habitat volumes, and habitat volume fractions for each habitat were computed. Statistical analyses were performed using GraphPad Prism 7 (GraphPad Software, La Jolla, CA). Data normality was assessed using the D'Agostino–Pearson test, and significance of differences in habitat volumes between groups was assessed by 2-tailed unpaired *t*-tests. Survival analyses were performed using Kaplan–Meier survival curves, and statistical significance was computed using the log-rank test. For the Kaplan–Meier analysis, habitat volumes were dichotomized into 2 groups using the median score value.

Table 2. Intratumoral Habitats' Definitions on Calibrated FLAIR and ΔT1W Intensities

	Calibrated FLAIR Image Intensity	Calibrated ΔT1W Difference Intensity
Habitat 1	≤600	≤303
Habitat 2	>600	≤303
Habitat 3	≤600	303 < ΔT1W ≤ 790
Habitat 4	>600	303 < ΔT1W ≤ 790
Habitat 5	≤600	>790
Habitat 6	>600	>790

RESULTS

Mean tumor volumes at diagnosis were comparable between the LTS and STS groups in the discovery cohort (33 ± 6.6 cm^3 vs. 37 ± 6.1 cm^3, $P = .62$) (see online Supplemental Figure 2A). There was no statistically significant difference in mean tumor volumes at diagnosis between the LTS and STS groups in the validation cohort (33 ± 7.0 cm^3 vs. 17 ± 4.8 cm^3, $P = .075$), although there was a trend toward smaller tumor volumes in the STS group (see online Supplemental Figure 2B).

Figure 3 depicts differences in habitat 6 (high contrast enhancement and high FLAIR) content between a representative LTS subject (left; overall survival, 41+ months) and STS subject (right; overall survival, 3 months) at the time of tumor presentation before surgical intervention. In the discovery cohort habitat 6 comprised a significantly higher volume fraction ($P < .03$) of the tumor volume at diagnosis in long-term survivors

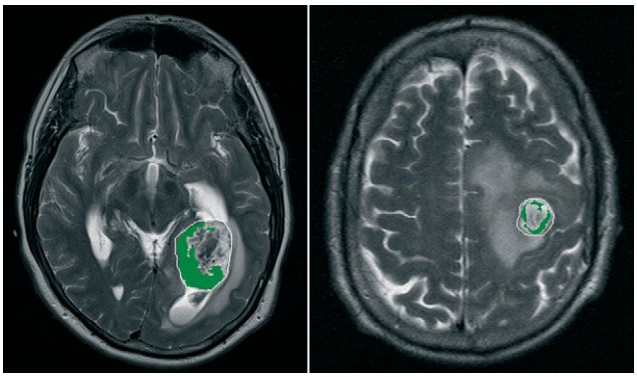


Figure 3. Habitat 6 (high enhancement and high FLAIR) on preoperative MRI comprises 23% of the tumor by volume in a long-term survivor (left, overall survival 41+ months) and 9% of the tumor by volume in a short-term survivor (right, overall survival 3 months).

(mean ± S.E.M. = 35% ± 6.5%; n = 22) compared with short-term survivors (mean ± S.E.M. = 17% ± 4.5%; n = 22) (Figure 4A). This finding was replicated in the validation cohort ($P < .007$), with habitat 6 comprising 34% ± 4.8% (n = 15) of the tumor volume in LTS subjects compared with 16% ± 4.0% (n = 15) of the tumor volume in STS subjects (Figure 4B).

Habitat 2 (low enhancement and high FLAIR) comprised a significantly lower volume fraction ($P = .0126$) of the tumor at diagnosis in long-term survivors (mean ± S.E.M. = 28 ± 5.7; n = 22) relative to short-term survivors (mean ± S.E.M. = 51 ± 6.8; n = 22) in the discovery cohort (Figure 5A), but this was not replicated in the validation cohort (Figure 5B). In parallel, habitat 1 (low enhancement and low FLAIR) was not found to be significantly different between LTS and STS subjects in the discovery cohort (Figure 5C) but comprised a significantly lower volume fraction ($P = .0279$) of the tumor at diagnosis in long-term survivors (mean ± S.E.M. = 3.2 ± 0.96; n = 15) relative to short-term survivors (mean ± S.E.M. = 12 ± 3.4; n = 15) in the validation cohort (Figure 5D). Minor inconsistencies in FLAIR intensity calibration across the patients may be the root cause of this variable finding, given that Habitats 1 and 2 belong to the low and high FLAIR clusters, respectively.

Habitat 3 (medium enhancement and low FLAIR), habitat 4 (medium enhancement and high FLAIR), and habitat 5 (high enhancement and low FLAIR) were not significantly different between the LTS and STS groups in either the discovery or validation cohorts (see online Supplemental Figure 3).

Median percent of tumor volume occupied by habitat 6 in the discovery cohort (5.77%) was used as a cutpoint to dichotomize patients into high and low habitat 6 fraction groups. Kaplan–Meier survival analyses were then carried out separately in the discovery and validation cohorts using the prespecified cutpoint (5.77%) established for the discovery cohort, as a stringent test of reproducibility. Based on the median cutpoint, low and high fractions of habitat 6 were not associated with overall survival in the discovery cohort (Figure 6A; $P = .62$), but were statistically significant with respect to overall survival in the validation cohort (Figure 6B; $P = .0001$). In the discovery cohort, Kaplan–Meier 3-year survival rates were 45% and 55% in the < median versus ≥ median subgroups, respectively. In the validation cohort corresponding 3-year survival rates were 18% and 68%, respectively.

DISCUSSION

The overall goal of our work is to develop noninvasive imaging biomarkers that can be used to drive evolution-based adaptive therapeutic strategies for GBM. For any biomarker to be clinically useful, it must be computable reliably and reproducibly (24). MRI parameters such as ADC, T1, and T2, and with some limitations, also model-dependent parameters such as relative cerebral blood volume (rCBV), relative cerebral blood flow, and K^{trans}, are comparable between data sets when standardized

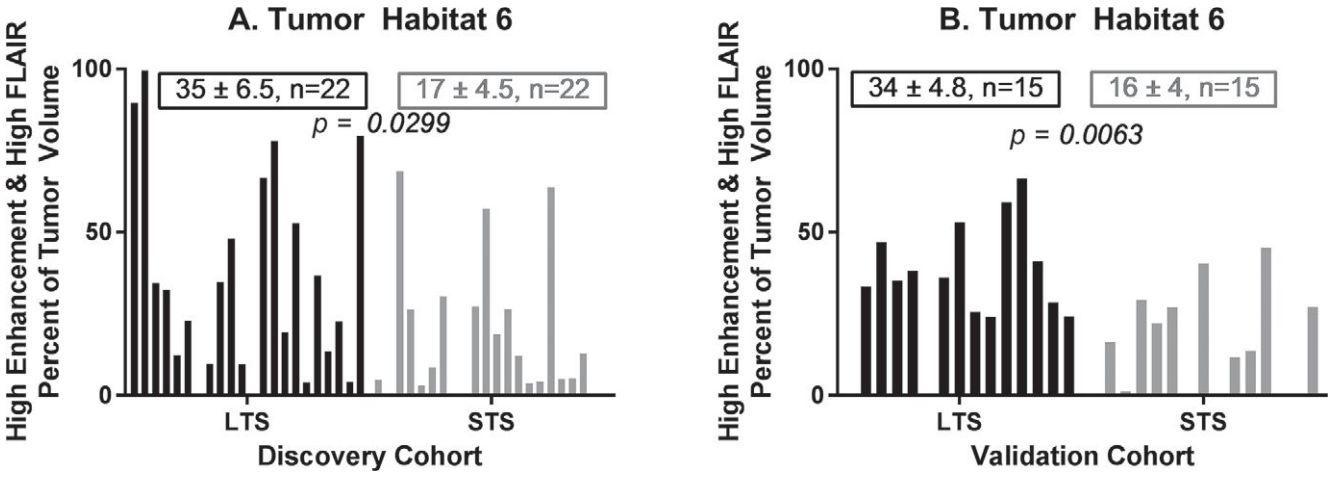


Figure 4. Habitat 6 (high enhancement and high FLAIR) was significantly higher in the LTS group relative to the STS group in both the (A) discovery and (B) validation cohorts.

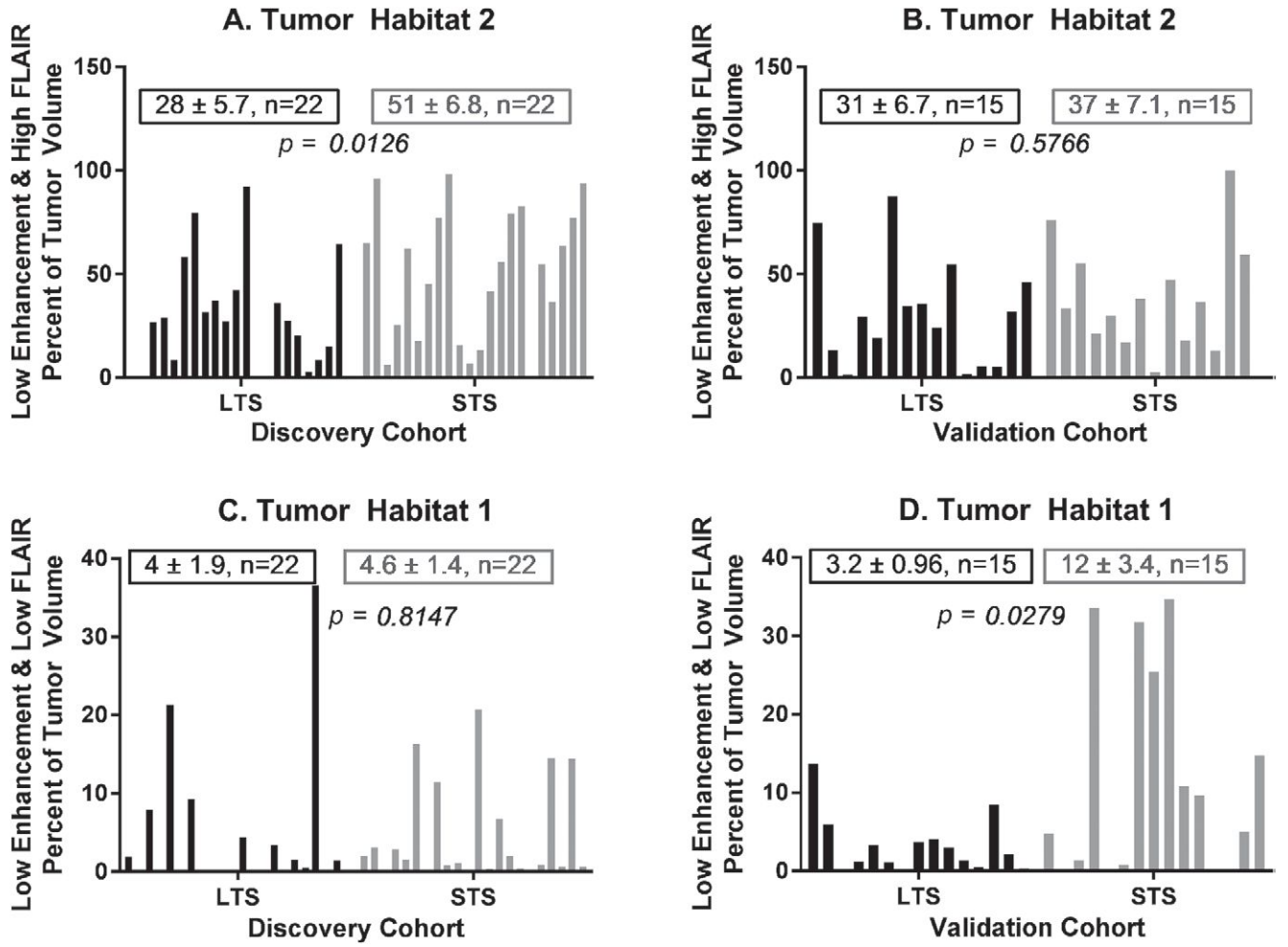


Figure 5. Habitat 2 (low enhancement and high FLAIR) was significantly lower in the LTS group relative to the STS group in the discovery cohort (panel A), but this difference was not recapitulated in the validation cohort (panel B). Habitat 1 (low enhancement and low FLAIR) was not significantly different between the LTS and STS groups in the discovery cohort (panel C), but was significantly lower in the LTS group in the validation cohort (panel D).

protocols are utilized (25-35). Parameter maps are therefore attractive for computing tumor habitats consistently across patients and scan dates, but these maps are not routinely collected as part of standard-of-care imaging. The subjects in our study received their initial diagnostic scans at a variety of institutions including at community radiology facilities, as a result of which there was great variability in the type and quality of scans that were available for retrospective analysis. In particular, we were unable to curate sufficient numbers of LTS subjects with available ADC maps at diagnosis. We therefore sought to compute intratumoral habitats using FLAIR, T1W, and T1W-CE scans after calibrating raw MRI pixel intensities against 2 reference tissues.

High signal on ΔT1W is indicative of either good perfusion or high microvascular leakiness. High intensity on FLAIR images in glioma represents a mixture of vasogenic edema, which arises from leakage of plasma into regions with low cell density, and tumor cell infiltration along long white matter tracts (36). Our retrospective study shows, in both a discovery cohort and a validation cohort, that tumors in LTS subjects have a significantly higher fraction of habitat 6 (high contrast enhancement and high FLAIR signal intensity) than STS. Particularly striking is the similarity in habitat 6 content of LTS tumors between the discovery and validation cohorts (35% and 34%, respectively) and of STS tumors between the discovery and validation cohorts (17% and 16%, respectively). We divided tumor regions with high signal intensity on ΔT1W calibrated difference images into 2 distinct habitats with either high or low FLAIR signal. Low FLAIR signal would be expected in regions with high contrast enhancement stemming from good perfusion, which would be conducive to high cellular density, although not necessarily where the enhancement arises from microvascular leakiness. Our results demonstrate the high contrast enhancement and high FLAIR signal habitat is strongly associated with patient survival.

In a preliminary study of pretreatment MRI examinations from 32 patients with GBM enrolled in the TCGA, Gatenby et al. showed that GBM tumor habitats defined on FLAIR and T1W-CE

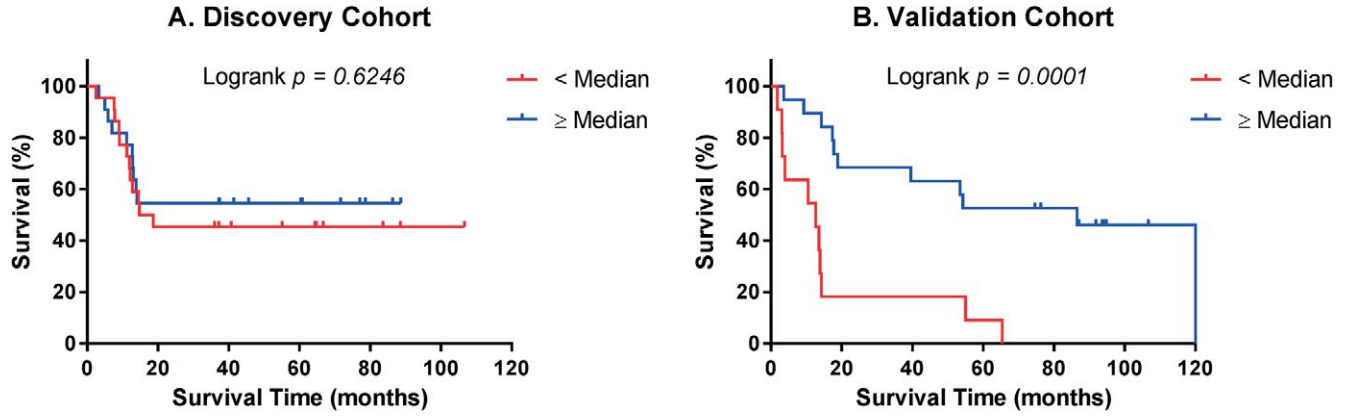


Figure 6. Kaplan–Meier plots of overall survival in the discovery cohort (A). Survival of patients with habitat 6 volume fraction ≥ median (5.77%, n = 22) and < median (n = 22). Kaplan–Meier plots of overall survival in the validation cohort (B). Survival of patients with habitat 6 volume fraction ≥ median from the discovery cohort (5.77%, n = 19) and < median (n = 11).

images could be used to differentiate patients who survived <400 days from patients who survived ≥400 days postdiagnosis (37). A follow-up study indicated that incorporating information from 3 MRI sequences, namely, T2W, FLAIR, and T1W-CE, improved prediction of survival time in patients with GBM (38). LaViolette et al. similarly clustered voxels into low, medium, and high classes on T1W, T1W-CE, FLAIR and apparent diffusion coefficient of water (ADC) maps to divide GBM tumors into 81 habitats, and identified 5 specific habitats that when present at higher volumes correlated with poorer prognosis (39). Recently, Juan-Albarracín et al. analyzed preoperative DSC-MRI and FLAIR scans of 50 patients with GBM to compute tumor habitats on the basis of rCBV, relative cerebral blood flow, and edema, and they found a surprising correlation between longer survival times and lower indices of perfusion (40). Boonzaier et al. report that tumor habitats reflecting low ADC values intersecting with high rCBV values demonstrate a significantly elevated choline-to-N-acetylaspartate ratio on 1H magnetic resonance spectroscopy, and that a higher proportion of this habitat within the NE region of GBM is associated with poor overall survival (41). Interpatient diversity in overall imaging patterns of growth and invasion has been associated with tumor aggressiveness and clinical outcomes across patients (42-45). Our investigation leverages unique resources of data including patients with exceptionally long follow-up for prognosis in glioblastoma.

Standard-of-care therapy in newly diagnosed GBM is maximal safe surgical resection followed by concomitant radiation therapy and temozolomide for 6 weeks, followed by adjuvant temozolomide for 6 monthly cycles (46). Thereafter, subjects in our retrospective study would each also have received a variety of investigational and/or palliative treatments, including extended cycles of temozolomide. Our findings suggest that one or more characteristics of the radiologically visible initial tumor mass define an intrinsic prognostically relevant tumor feature that continues to influence patient outcome months, and even years, after diagnosis. It is possible that the radiologic appearance of habitat 6 is a shared feature of disparate favorable markers in GBM, such as Isocitrate Dehydrogenase (IDH) mutation status (47), mesenchymal subtype (48) or lymphocyte cytokines such as CXCR4 (49). Alternately, one can hypothesize that components of the immune system in the LTS subjects retain the ability to recognize tumor antigens present in the original mass that are retained in the recurrent mass. Immune infiltrates in the tumor would be consistent with the MRI characteristics of habitat 6, namely, high contrast-enhancement and high tumor-associated edema. Pathological studies have shown that increased CD8+ T cell infiltrates in newly diagnosed GBM is associated with long-term survival (50), and we hypothesize that increased FLAIR signal in well-perfused—and presumably cellular—regions may be indicative of interstitial edema related to inflammatory changes caused by an immune response. A definitive biological interpretation of our finding requires further investigation.

Known weaknesses in our study include that the numbers in each survival group stratum were small and statistical power correspondingly limited to detect all but strong associations in the data. Specifically, while our analysis detected a significant difference between the LTS and STS groups in both the discovery and validation cohorts (Figure 4), on an individual patient basis, we could observe survival differences by a binary analysis around the median habitat 6 content in only the validation cohort (Figure 6). The need to improve calibration of raw MRI image intensities is revealed in the inconsistent significances of Habitats 1 and 2 in the discovery and validation cohorts (Figure 5). Additional covariates may also impinge upon our analysis. For example, in the discovery cohort, LTS and STS subjects were matched for parameters such as patient age and year of diagnosis, but LTS patients were nonetheless more educated and more likely to survive the completion of standard treatment. In the validation cohort, the LTS and STS groups were not matched for patient age and treatment regimens. It is unclear how these group differences might explain the present findings.

Only about 5% of patients with GBM undergoing standard of care survive ≥5 years postdiagnosis (46). Investigation of a

cohort of rare long-term survivors identifies a "habitat" on initial multiparametric MRI scans that is significantly different than in a control cohort. Our working hypothesis is that habitat 6 corresponds to a microenvironment that selects for glioma cells that are either innately less aggressive or are more amenable to control by tumor-infiltrating leukocytes. Habitat imaging has the potential to provide noninvasive longitudinal biomarkers of intratumoral evolutionary and ecological dynamics for the informed application of adaptive therapy to manage GBM.

Supplemental Materials

Supplemental Figures 1-3: http://dx.doi.org/10.18383/j.tom.2018.00052.sup.01

ACKNOWLEDGMENTS

We wish to thank the following people for their contributions to this work: Joo Kim, MD, for guidance on whole brain segmentation; Lila Kis and Doniya Milani for assistance with local data curation. We also wish to acknowledge research support from the National Institutes of Health (U54 CA193489; P30 CA076292 (IRAT Core); R01 CA116174; R01 NS060752, R01 CA164371, U54 CA210180, U54 CA143970, U54 CA193489, U01 CA220378), the James S. McDonnell Foundation (grant no. 220020400), the Ivy Foundation, and a 2017 Moffitt Team Science Award (Drs. Egan and Raghunand).

Disclosures: No disclosures to report.

Conflict of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. Hu LS, Ning S, Eschbacher JM, Gaw N, Dueck AC, Smith KA, Nakaji P, Plasencia J, Ranjbar S, Price SJ, Tran N, Loftus J, Jenkins R, O'Neill BP, Elmquist W, Baxter LC, Gao F, Frakes D, Karis JP, Zwart C, Swanson KR, Sarkaria J, Wu T, Mitchell JR, Li J. Multi-parametric MRI and texture analysis to visualize spatial histologic heterogeneity and tumor extent in glioblastoma. PLoS One. 2015;10:e0141506.
2. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455:1061–1068.
3. Sottoriva A, Spiteri I, Piccirillo SG, Touloumis A, Collins VP, Marioni JC, Curtis C, Watts C, Tavare S. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. Proc Natl Acad Sci U S A. 2013;110:4009–4014.
4. Gill BJ, Pisapia DJ, Malone HR, Goldstein H, Lei L, Sonabend A, Yun J, Samanamud J, Sims JS, Banu M, Dovas A, Teich AF, Sheth SA, McKhann GM, Sisti MB, Bruce JN, Sims PA, Canoll P. MRI-localized biopsies reveal subtype-specific differences in molecular and cellular composition at the margins of glioblastoma. Proc Natl Acad Sci U S A. 2014;111:12550–12555.
5. Autry A, Phillips JJ, Maleschlijski S, Roy R, Molinaro AM, Chang SM, Cha S, Lupo JM, Nelson SJ. Characterization of metabolic, diffusion, and perfusion properties in GBM: contrast-enhancing versus non-enhancing tumor. Transl Oncol. 2017;10:895–903.
6. Macyszyn L, Akbari H, Pisapia JM, Da X, Attiah M, Pigrish V, Bi Y, Pal S, Davuluri RV, Roccograndi L, Dahmane N, Martinez-Lage M, Biros G, Wolf RL, Bilello M, O'Rourke DM, Davatzikos C. Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques. Neuro Oncol. 2016;18:417–25.
7. Hu LS, Ning S, Eschbacher JM, Baxter LC, Gaw N, Ranjbar S, Plasencia J, Dueck AC, Peng S, Smith KA, Nakaji P, Karis JP, Quarles CC, Wu T, Loftus JC, Jenkins RB, Sicotte H, Kollmeyer TM, O'Neill BP, Elmquist W, Hoxworth JM, Frakes D, Sarkaria J, Swanson KR, Tran NL, Li J, Mitchell JR. Radiogenomics to characterize regional genetic heterogeneity in glioblastoma. Neuro Oncol. 2017;19:128–137.
8. Reardon DA, Wen PY. Glioma in 2014: unravelling tumour heterogeneity-implications for therapy. Nat Rev Clin Oncol. 2015;12:69–70.
9. Enriquez-Navas PM, Kam Y, Das T, Hassan S, Silva A, Foroutan P, Ruiz E, Martinez G, Minton S, Gillies RJ, Gatenby RA. Exploiting evolutionary principles to prolong tumor control in preclinical models of breast cancer. Sci Transl Med. 2016;8:327ra24.
10. Zhang J, Cunningham JJ, Brown JS, Gatenby RA. Integrating evolutionary dynamics into treatment of metastatic castrate-resistant prostate cancer. Nat Commun. 2017;8:1816.
11. Gatenby RA, Grove O, Gillies RJ. Quantitative imaging in cancer evolution and ecology. Radiology. 2013;269:8–15.
12. Manel S, Schwartz MK, Luikart G, Taberlet P. Landscape genetics: combining landscape ecology and population genetics. Trends Ecol Evol. 2003;18:189–197.
13. Turner MG. Landscape ecology: what is the state of the science? Annu Rev Ecol Evol Syst. 2005;36:319–344.
14. Vannier MW, Butterfield RL, Jordan D, Murphy WA, Levitt RG, Gado M. Multispectral analysis of magnetic resonance images. Radiology. 1985;154:221–224.
15. Vannier MW, Butterfield RL, Rickman DL, Jordan DM, Murphy WA, Biondetti PR. Multispectral magnetic resonance image analysis. Crit Rev Biomed Eng. 1987;15:117–144.
16. Gohagan JK, Spitznagel EL, Murphy WA, Vannier MW, Dixon WT, Gersell DJ, Rossnick SL, Totty WG, Destouet JM, Rickman DL, et al. Multispectral analysis of MR images of the breast. Radiology. 1987;163:703–707.
17. Carano RA, Ross AL, Ross J, Williams SP, Koeppen H, Schwall RH, Van Bruggen N. Quantification of tumor tissue populations by multispectral analysis. Mag Reson Med. 2004;51:542–551.
18. Barck KH, Willis B, Ross J, French DM, Filvaroff EH, Carano RA. Viable tumor tissue detection in murine metastatic breast cancer by whole-body MRI and multispectral analysis. Magn Reson Med. 2009;62:1423–1430.
19. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology. 2018;286:800–809.
20. Chinot OL, Wick W, Mason W, Henriksson R, Saran F, Nishikawa R, Carpentier AF, Hoang-Xuan K, Kavan P, Cernea D, Brandes AA, Hilton M, Abrey L, Cloughesy T. Bevacizumab plus radiotherapy-temozolomide for newly diagnosed glioblastoma. N Engl J Med. 2014;370:709–722.
21. Gilbert MR, Dignam JJ, Armstrong TS, Wefel JS, Blumenthal DT, Vogelbaum MA, Colman H, Chakravarti A, Pugh S, Won M, Jeraj R, Brown PD, Jaeckle KA, Schiff D, Stieber VW, Brachman DG, Werner-Wasik M, Tremont-Lukats IW, Sulman EP, Aldape KD, Curran WJ, Jr., Mehta MP. A randomized trial of bevacizumab for newly diagnosed glioblastoma. N Engl J Med. 2014;370:699–708.
22. Stupp R, Hegi ME, Mason WP, van den Bent MJ, Taphoorn MJ, Janzer RC, Ludwin SK, Allgeier A, Fisher B, Belanger K, Hau P, Brandes AA, Gijtenbeek J, Marosi C, Vecht CJ, Mokhtari K, Wesseling P, Villa S, Eisenhauer E, Gorlia T, Weller M, Lacombe D, Cairncross JG, Mirimanoff RO. Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. Lancet Oncol. 2009;10:459–466.
23. Otsu N. A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern. 1979;9:62–66.
24. Hayes DF. Biomarker validation and testing. Mol Oncol. 2015;9:960–966.
25. Leach MO, Morgan B, Tofts PS, Buckley DL, Huang W, Horsfield MA, Chenevert TL, Collins DJ, Jackson A, Lomas D, Whitcher B, Clarke L, Plummer R, Judson I, Jones R, Alonzi R, Brunner T, Koh DM, Murphy P, Waterton JC, Parker G, Graves MJ, Scheenen TW, Redpath TW, Orton M, Karczmar G, Huisman H, Barentsz J, Padhani A. Imaging vascular function for early stage clinical trials using dynamic contrast-enhanced magnetic resonance imaging. Eur Radiol. 2012;22:1451–1464.
26. Malyarenko D, Galban CJ, Londy FJ, Meyer CR, Johnson TD, Rehemtulla A, Ross BD, Chenevert TL. Multi-system repeatability and reproducibility of apparent diffusion coefficient measurement using an ice-water phantom. J Magn Reson Imaging. 2013;37:1238–1246.
27. Ellingson BM, Bendszus M, Boxerman J, Barboriak D, Erickson BJ, Smits M, Nelson SJ, Gerstner E, Alexander B, Goldmacher G, Wick W, Vogelbaum M, Weller M, Galanis E, Kalpathy-Cramer J, Shankar L, Jacobs P, Pope WB, Yang D, Chung C, Knopp MV, Cha S, van den Bent MJ, Chang S, Yung WK, Cloughesy TF, Wen PY, Gilbert MR. Consensus recommendations for a standardized Brain Tumor Imaging Protocol in clinical trials. Neuro Oncol. 2015;17:1188–1198.

28. Taouli B, Beer AJ, Chenevert T, Collins D, Lehman C, Matos C, Padhani AR, Rosenkrantz AB, Shukla-Dave A, Sigmund E, Tanenbaum L, Thoeny H, Thomassin-Naggara I, Barbieri S, Corcuera-Solano I, Orton M, Partridge SC, Koh DM. Diffusion-weighted imaging outside the brain: consensus statement from an ISMRM-sponsored workshop. J Magn Reson Imaging. 2016;44:521–540.
29. Huang W, Chen Y, Fedorov A, Li X, Jajamovich GH, Malyarenko DI, Aryal MP, LaViolette PS, Oborski MJ, O'Sullivan F, Abramson RG, Jafari-Khouzani K, Afzal A, Tudorica A, Moloney B, Gupta SN, Besa C, Kalpathy-Cramer J, Mountz JM, Laymon CM, Muzi M, Schmainda K, Cao Y, Chenevert TL, Taouli B, Yankeelov TE, Fennessy F, Li X. The impact of arterial input function determination variations on prostate dynamic contrast-enhanced magnetic resonance imaging pharmacokinetic modeling: a multicenter data analysis challenge. Tomography. 2016;2:56–66.
30. O'Connor JP, Aboagye EO, Adams JE, Aerts HJ, Barrington SF, Beer AJ, Boellaard R, Bohndiek SE, Brady M, Brown G, Buckley DL, Chenevert TL, Clarke LP, Collette S, Cook GJ, deSouza NM, Dickson JC, Dive C, Evelhoch JL, Faivre-Finn C, Gallagher FA, Gilbert FJ, Gillies RJ, Goh V, Griffiths JR, Groves AM, Halligan S, Harris AL, Hawkes DJ, Hoekstra OS, Huang EP, Hutton BF, Jackson EF, Jayson GC, Jones A, Koh DM, Lacombe D, Lambin P, Lassau N, Leach MO, Lee TY, Leen EL, Lewis JS, Liu Y, Lythgoe MF, Manoharan P, Maxwell RJ, Miles KA, Morgan B, Morris S, Ng T, Padhani AR, Parker GJ, Partridge M, Pathak AP, Peet AC, Punwani S, Reynolds AR, Robinson SP, Shankar LK, Sharma RA, Soloviev D, Stroobants S, Sullivan DC, Taylor SA, Tofts PS, Tozer GM, van Herk M, Walker-Samuel S, Wason J, Williams KJ, Workman P, Yankeelov TE, Brindle KM, McShane LM, Jackson A, Waterton JC. Imaging biomarker roadmap for cancer studies. Nat Rev Clin Oncol. 2017;14:169–186.
31. Klaassen R, Gurney-Champion OJ, Wilmink JW, Besselink MG, Engelbrecht MRW, Stoker J, Nederveen AJ, van Laarhoven HWM. Repeatability and correlations of dynamic contrast enhanced and T2* MRI in patients with advanced pancreatic ductal adenocarcinoma. Magn Reson Imaging. 2018;50:1–9.
32. Sorace AG, Wu C, Barnes SL, Jarrett AM, Avery S, Patt D, Goodgame B, Luci JJ, Kang H, Abramson RG, Yankeelov TE, Virostko J. Repeatability, reproducibility, and accuracy of quantitative MRI of the breast in the community radiology setting. J Magn Reson Imaging. 2018.
33. Malyarenko D, Fedorov A, Bell L, Prah M, Hectors S, Arlinghaus L, Muzi M, Solaiyappan M, Jacobs M, Fung M, Shukla-Dave A, McManus K, Boss M, Taouli B, Yankeelov TE, Quarles CC, Schmainda K, Chenevert TL, Newitt DC. Toward uniform implementation of parametric map Digital Imaging and Communication in Medicine standard in multisite quantitative diffusion imaging studies. J Med Imaging (Bellingham). 2018;5:011006.
34. Newitt DC, Malyarenko D, Chenevert TL, Quarles CC, Bell L, Fedorov A, Fennessy F, Jacobs MA, Solaiyappan M, Hectors S, Taouli B, Muzi M, Kinahan PE, Schmainda KM, Prah MA, Taber EN, Kroenke C, Huang W, Arlinghaus LR, Yankeelov TE, Cao Y, Aryal M, Yen YF, Kalpathy-Cramer J, Shukla-Dave A, Fung M, Liang J, Boss M, Hylton N. Multisite concordance of apparent diffusion coefficient measurements across the NCI Quantitative Imaging Network. J Med Imaging (Bellingham). 2018;5:011003.
35. Bane O, Hectors SJ, Wagner M, Arlinghaus LL, Aryal MP, Cao Y, Chenevert TL, Fennessy F, Huang W, Hylton NM, Kalpathy-Cramer J, Keenan KE, Malyarenko DI, Mulkern RV, Newitt DC, Russek SE, Stupic KF, Tudorica A, Wilmes LJ, Yankeelov TE, Yen YF, Boss MA, Taouli B. Accuracy, repeatability, and interplatform reproducibility of T1 quantification methods used for DCE-MRI: results from a multicenter phantom study. Magn Reson Med. 2018;79:2564–2575.
36. Villanueva-Meyer JE, Mabray MC, Cha S. Current clinical brain tumor imaging. Neurosurgery. 2017;81:397–415.
37. Zhou M, Hall L, Goldgof D, Russo R, Balagurunathan Y, Gillies R, Gatenby R. Radiologically defined ecological dynamics and clinical outcomes in glioblastoma multiforme: preliminary results. Transl Oncol. 2014;7:5–13.
38. Zhou M, Chaudhury B, Hall LO, Goldgof DB, Gillies RJ, Gatenby RA. Identifying spatial imaging biomarkers of glioblastoma multiforme for survival group prediction. J Magn Reson Imaging. 2017;46:115–123.
39. McGarry SD, Hurrell SL, Kaczmarowski AL, Cochran EJ, Connelly J, Rand SD, Schmainda KM, LaViolette PS. Magnetic resonance imaging-based radiomic profiles predict patient prognosis in newly diagnosed glioblastoma before therapy. Tomography. 2016;2:223–228.
40. Juan-Albarracín J, Fuster-Garcia E, Pérez-Girbés A, Aparici-Robles F, Alberich-Bayarri Á, Revert-Ventura A, Martí-Bonmatí L, García-Gómez JM. Glioblastoma: vascular habitats detected at preoperative dynamic susceptibility-weighted contrast-enhanced perfusion MR imaging predict survival. Radiology. 2018;287:944–954.
41. Boonzaier NR, Larkin TJ, Matys T, van der Hoorn A, Yan JL, Price SJ. Multiparametric MR imaging of diffusion and perfusion in contrast-enhancing and nonenhancing components in patients with glioblastoma. Radiology. 2017;284:180–190.
42. Baldock AL, Ahn S, Rockne R, Johnston S, Neal M, Corwin D, Clark-Swanson K, Sterin G, Trister AD, Malone H, Ebiana V, Sonabend AM, Mrugala M, Rockhill JK, Silbergeld DL, Lai A, Cloughesy T, McKhann GM, 2nd, Bruce JN, Rostomily RC, Canoll P, Swanson KR. Patient-specific metrics of invasiveness reveal significant prognostic benefit of resection in a predictable subset of gliomas. PLoS One. 2014;9:e99057.
43. Swanson KR, Rostomily RC, Alvord EC. A mathematical modelling tool for predicting survival of individual patients following resection of glioblastoma: a proof of principle. Br J Cancer. 2008;98:113–119.
44. Szeto MD, Chakraborty G, Hadley J, Rockne R, Muzi M, Alvord EC, Jr., Krohn KA, Spence AM, Swanson KR. Quantitative metrics of net proliferation and invasion link biological aggressiveness assessed by MRI with hypoxia assessed by FMISO-PET in newly diagnosed glioblastomas. Cancer Res. 2009;69:4502–4509.
45. Baldock AL, Yagle K, Born DE, Ahn S, Trister AD, Neal M, Johnston SK, Bridge CA, Basanta D, Scott J, Malone H, Sonabend AM, Canoll P, Mrugala MM, Rockhill JK, Rockne RC, Swanson KR. Invasion and proliferation kinetics in enhancing gliomas predict IDH1 mutation status. Neuro Oncol. 2014;16:779–786.
46. Stupp R, Mason WP, van den Bent MJ, Weller M, Fisher B, Taphoorn MJ, Belanger K, Brandes AA, Marosi C, Bogdahn U, Curschmann J, Janzer RC, Ludwin SK, Gorlia T, Allgeier A, Lacombe D, Cairncross JG, Eisenhauer E, Mirimanoff RO. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. N Engl J Med. 2005;352:987–996.
47. Price SJ, Allinson K, Liu H, Boonzaier NR, Yan JL, Lupson VC, Larkin TJ. Less invasive phenotype found in isocitrate dehydrogenase-mutated glioblastomas than in isocitrate dehydrogenase wild-type glioblastomas: a diffusion-tensor imaging study. Radiology. 2017;283:215–221.
48. Naeini KM, Pope WB, Cloughesy TF, Harris RJ, Lai A, Eskin A, Chowdhury R, Phillips HS, Nghiemphu PL, Behbahanian Y, Ellingson BM. Identifying the mesenchymal molecular subtype of glioblastoma using quantitative volumetric analysis of anatomic magnetic resonance images. Neuro Oncol. 2013;15:626–634.
49. Ma X, Shang F, Zhu W, Lin Q. CXCR4 expression varies significantly among different subtypes of glioblastoma multiforme (GBM) and its low expression or hypermethylation might predict favorable overall survival. Expert Rev Neurother. 2017;17:941–946.
50. Yang I, Tihan T, Han SJ, Wrensch MR, Wiencke J, Sughrue ME, Parsa AT. CD8+ T-cell infiltrate in newly diagnosed glioblastoma is associated with long-term survival. J Clin Neurosci. 2010;17:1381–1385.

[18F] FDG Positron Emission Tomography (PET) Tumor and Penumbra Imaging Features Predict Recurrence in Non–Small Cell Lung Cancer

Sarah A. Mattonen[1], Guido A. Davidzon[2], Shaimaa Bakr[3], Sebastian Echegaray[1], Ann N.C. Leung[1], Minal Vasanawala[4], George Horng[5], Sandy Napel[1], and Viswam S. Nair[1,6,7]

Departments of [1]Radiology, [2]Radiology, Division of Nuclear Medicine, and [3]Electrical Engineering, Stanford University, Stanford, CA; [4]Palo Alto VA Health Care System, Palo Alto, CA; [5]California Pacific Medical Center, San Francisco, CA; [6]Pulmonary & Critical Care Medicine, Moffitt Cancer Center & Research Institute, Tampa, FL; and [7]Morsani College of Medicine, University of South Florida, Tampa, FL

Corresponding Author:
Sarah A. Mattonen, PhD
James H. Clark Center, 318 Campus Drive, Room S355,
Stanford, CA 94305;
E-mail: smattonen@stanford.edu

Key Words: radiomics, lung cancer, risk stratification, recurrence, PET

Abbreviations: Non–small cell lung cancer (NSCLC), metabolic tumor volume (MTV), fluoro-2-deoxy-D-glucose (FDG), positron emission tomography (PET), computed tomography (CT), Dice similarity coefficient (DSC), mean absolute distance (MAD), gray-level co-occurrence matrix (GLCM), least absolute shrinkage and selection operator (LASSO), Akaike information criterion (AIC), hazard ratio (HR), confidence interval (CI), intraclass correlation coefficients (ICCs)

ABSTRACT

We identified computational imaging features on 18F-fluorodeoxyglucose positron emission tomography (PET) that predict recurrence/progression in non–small cell lung cancer (NSCLC). We retrospectively identified 291 patients with NSCLC from 2 prospectively acquired cohorts (training, n = 145; validation, n = 146). We contoured the metabolic tumor volume (MTV) on all pretreatment PET images and added a 3-dimensional penumbra region that extended outward 1 cm from the tumor surface. We generated 512 radiomics features, selected 435 features based on robustness to contour variations, and then applied randomized sparse regression (LASSO) to identify features that predicted time to recurrence in the training cohort. We built Cox proportional hazards models in the training cohort and independently evaluated the models in the validation cohort. Two features including stage and a MTV plus penumbra texture feature were selected by LASSO. Both features were significant univariate predictors, with stage being the best predictor (hazard ratio [HR] = 2.15 [95% confidence interval (CI): 1.56–2.95], $P < .001$). However, adding the MTV plus penumbra texture feature to stage significantly improved prediction ($P = .006$). This multivariate model was a significant predictor of time to recurrence in the training cohort (concordance = 0.74 [95% CI: 0.66–0.81], $P < .001$) that was validated in a separate validation cohort (concordance = 0.74 [95% CI: 0.67–0.81], $P < .001$). A combined radiomics and clinical model improved NSCLC recurrence prediction. FDG PET radiomic features may be useful biomarkers for lung cancer prognosis and add clinical utility for risk stratification.

INTRODUCTION

Lung cancer remains the most common cause of cancer death worldwide, and the 5-year survival rates of non–small cell lung cancer (NSCLC) remain quite poor despite advances in diagnosis and treatment (1, 2). Further, many patients will develop recurrence or progression following primary treatment. The absolute risk of any recurrence at 5 years post-treatment ranges from 33% to 52%, with the majority occurring at a distant site (3, 4). Among prognostic factors for predicting outcomes in NSCLC, tumor stage based on the American Joint Committee on Cancer (AJCC) staging system is currently considered the best for predicting outcomes (5). More accurate clinical, imaging, and molecular biomarkers will be extremely useful for stratifying patients who are at a higher risk of recurrence and who might benefit from adjuvant or more aggressive treatment options (6).

Maximum standardized uptake value (SUV_{max}) on fluorine-18F fluoro-2-deoxy-D-glucose (FDG) positron emission tomography (PET) imaging has also been shown to predict recurrence or death in NSCLC (7). However, this is a single-voxel metric; we hypothesized that applying a radiomics approach to extract more complex information (eg, texture) from standard medical images could provide additional prognostic information (8, 9).

While recent work has evaluated the potential for radiomics features to augment traditional metrics of response (10-12), the majority of studies to date have focused on only the metabolic

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2018.00026

tumor volume (MTV) on PET and, to the best of our knowledge, no study has investigated the peritumoral region. Tumor invasion from the main mass can be defined by infiltration of stroma, blood vessels, or visceral pleura (13). Recent studies have also shown the potential for tumor cells to spread into air spaces in the lung tissue adjacent to the tumor volume (14). It is well known that these features may present as border spiculation, vascular convergence, or pleural attachment surrounding the tumor on anatomical imaging, and that they may result in subtle heterogeneous uptake on PET imaging (15).

We investigated the potential of FDG-PET radiomics to predict recurrence in NSCLC by (1) assessing the variability in radiomic feature extraction from PET images and (2) building and validating a radiomics model to predict time to recurrence. We hypothesize that computational imaging features in the tumor and surrounding area on FDG-PET can augment clinical features to improve recurrence prediction.

METHODOLOGY

Patient Selection

We retrospectively analyzed a total of 291 patients with NSCLC from 2 distinct cohorts of prospectively acquired patients (n = 145 and n = 146). The study was approved by our Institutional Review Board, and all subjects signed informed consent before participation. Our study was also compliant with the Health Insurance Portability and Accountability Act.

The training cohort consisted of subjects from a pool of patients with early-stage NSCLC referred for surgical treatment at 2 local medical centers between 2008 and 2012 with preoperative PET/computed tomography (CT) performed before surgery (n = 145). This data set is publicly available on The Cancer Imaging Archive (16, 17). We used a second cohort (n = 146) for model validation. This was a cohort from 3 local medical centers between 2010 and 2016. Subjects were selected from patients undergoing evaluation for lung cancer by PET/CT imaging before definitive treatment as part of an observational biomarker study. In both the training and validation cohorts, there were no patients that received neoadjuvant therapy.

The AJCC seventh edition system was used for staging. Pathological staging was used in the training cohort and a combination of clinical and pathological staging in the validation cohort. Demographic differences between the training and validation cohorts were assessed using the Wilcoxon rank-sum test for continuous variables and the χ^2 test for categorical variables. All patients were followed per standard clinical protocol with clinical examination and imaging. We analyzed the combined endpoint of disease recurrence or progression. For stage I–IIIA subjects, we defined recurrence as either local, regional, or distant. For patients with stage IIIB–IV disease, we defined an event as any progression of disease. Time to event or last known follow-up was recorded from the date of pretreatment PET imaging.

Image Acquisition

Pretreatment FDG-PET/CT scans were acquired using a standard clinical protocol at 1 of 3 local medical centers. Images were acquired using either a GE Discovery VCT (GE Health care, Waukesha, WI), a GE Discovery LS PET/CT (GE Healthcare, Waukesha, WI), a Siemens Biograph mCT (Siemens Healthcare, Erlangen, Germany), or a Phillips Allegro/Gemini TF PET/CT (Phillips Healthcare, Cleveland, OH). Patients underwent scanning following fasting for a minimum of 6–8 h. A dose of 12–17 mCi of FDG was administered and patients underwent scanning from the skull base to mid-thigh using bed positions acquired every 2–5 minutes ~45–60 minutes after injection. Manufacturer-specific CT-based attenuated correction was performed using ordered subset expectation maximization reconstruction.

Region of Interest Delineations

Pretreatment PET images were converted to SUV units normalized by body weight. Two research assistants (S.M. and S.B.) were trained by a board-certified physician in Nuclear Medicine (G.D.) in using MIM Version 6.6 (MIM Software Inc., Cleveland, OH) to contour tumor MTVs using the semiautomatic PET-edge gradient-based segmentation tool. Both observers contoured all images independently in the training cohort. A subset of 21 images considered difficult to contour were reviewed by the same physician and re-delineated if necessary. To assess intraobserver variability, observer 1 (S.M.) contoured all images a second time after a delay of 3 months. We calculated the Dice similarity coefficient (DSC), mean absolute distance (MAD) of the boundary, and absolute volume difference between each set of contours to assess inter- and intraobserver variability of the MTV regions in the training cohort. Observer 1 alone contoured all images in the validation cohort.

We then generated a 3-dimensional penumbra region extending outward 1 cm from the surface of the MTV to sample surrounding uptake by using a 3D distance transform with a threshold of 1 cm. This distance was intuitively chosen to sample enough surrounding tissue given the voxel sizes of the PET images, while avoiding oversampling normal tissue. In addition to the MTV alone, we also evaluated the following 2 additional regions: the MTV plus penumbra and the penumbra only (excluding the MTV).

Feature Extraction

We extracted radiomics features in the MTV, penumbra, and MTV plus penumbra regions in both cohorts using The Quantitative Image Feature Engine (18) implemented in MATLAB R2016B (The MathWorks, Natick, MA). In the MTV, features included size (n = 4), sphericity (n = 1), local volume-invariant integral (LVII) shape (n = 39), histogram intensity (n = 12), and gray-level co-occurrence matrix (GLCM) texture (n = 144) (19, 20), for a total of 200 features. Because the penumbra region was generated from the MTV, 44 size and shape measures were not calculated in the penumbra and MTV plus penumbra regions (because they would not be independent measurements), for a total of 156 features in each. This resulted in a total of 512 features for analysis as summarized in Table 1. We set a fixed intensity bin size of 0.2 SUV for texture feature calculation to allow a meaningful comparison between images on the same SUV scale. This discretization may also reduce the differences between multiple scanners used in this study (21).

We then calculated intraclass correlation coefficients (ICCs) across the 3 sets of outlines for each radiomic feature to assess inter- and intraobserver variability. Robust features, defined as

Table 1. Number of Extracted Features

Region of Interest	Feature Type	Number of Features	Total Number of Features in ROI
Metabolic Tumor Volume (MTV)	Size	4	200
	Sphericity	1	
	LVII shape	39	
	Intensity	12	
	GLCM texture	144	
Penumbra	Intensity	12	156
	GLCM texture	144	
MTV + Penumbra	Intensity	12	156
	GLCM texture	144	
Total Number of Features			512

those with ICCs >0.8 in the training cohort, were selected for further analysis (22, 23).

Model Building and Validation

All radiomic features were normalized (Z-score transformation) before feature selection and model building. We further optimized the features through a generalized linear model via the least absolute shrinkage and selection operator (LASSO) (24) Cox regression using the *glmnet* package in R software version 3.4.3 (25). LASSO is a shrinkage and variable selection method for high-dimensional data, which was used to select top features to predict time to recurrence in the training cohort. The robust radiomic features and the 2 known clinical predictors (stage and SUV_{max}) were provided to LASSO. Alpha, the regularization parameter, was set to 1 (LASSO penalty) to minimize the number of selected features by shrinking most of the coefficients to zero and to minimize potential overfitting in the training cohort. In total, 100 randomizations of 4-fold cross-validation was used to reduce the effect of randomness in fold selection. The mean cross-validated error curves were averaged for each tuning parameter lambda value across all randomizations. The lambda and corresponding radiomic features associated with the minimum error were selected.

We built univariate and multivariate Cox proportional hazards models in the training cohort using the most frequently selected radiomic and/or clinical features. We evaluated the Akaike information criterion (AIC) to compare the quality of the different models, with lower AICs representing a higher quality model. We assessed the likelihood ratio *P*-value for the derived models to show recurrence prediction significance. HRs and 95% CIs were reported for individual variables. To evaluate nested models combining the clinical and/or radiomic features, the likelihood ratio test was used to compare the goodness of fit.

To verify prediction validity, we locked the coefficients of the variables in the top model generated from the training cohort and evaluated it in the validation cohort. The prognostic value was assessed using the concordance index with Noether's test to determine significance from random (0.5). We performed Kaplan–Meier analysis to separate high- and low-risk groups based on the median risk score in the training cohort. We performed a Student's *t* test for dependent samples to compare concordance indices between the models. All statistical analyses and model building were performed using R. Statistical significance was assessed at the $P < .05$ level.

RESULTS

Patient Demographics

The training and validation cohorts were similarly matched with regard to median age ($P = .057$) and tumor location ($P = .571$) (Table 2). The training cohort had a higher proportion of males ($P = .005$) and adenocarcinoma histology ($P = .035$). There was a slightly higher proportion of stage IV patients in the validation cohort ($P < .001$), resulting in a larger percentage of patients who recurred/progressed ($P = .038$). The median time to recurrence was 14 months (range, 2–97) in the training cohort and 15 months (range, 1–59) in the validation cohort. The median follow-up time for censored patients without an event was 50 months (range, 1–115) in the training cohort and 32 months (range, 1–76) in the validation cohort.

Segmentation Variability

Table 3 shows the Dice Similarity Coefficient (DSC), Mean Absolute Boundary Distance (MAD), and absolute volume difference between observers in the training cohort. Overall, semiautomatic segmentations were highly reproducible with an average DSC >0.9, MAD <1 mm, and volume differences <1 mL. When we inspected images with low DSC, high MAD, and/or high volume differences, we found that lesions that had the largest degree of variability tended to have a low uptake (eg, SUV_{max} <2), heterogeneous uptake, and/or were adjacent to structures with a similar metabolic uptake as the tumor (eg, the heart or mediastinum), making the precise boundary of the tumor difficult to determine. These features were evident in ~20% of the cases.

Feature Variability

Table 4 shows the ICCs of the 4 different classes of radiomic features in each of the 3 regions of interest. We found that a total

Table 2. Baseline Patient and Lesion Characteristics

		Training (n=145)	Validation (n=146)	*P*-value
Age, years		69 (42–87)	71 (41–96)	.057
Gender	Male	109 (75%)	87 (60%)	.005
Tumor Location	Right upper lobe	52 (36%)	50 (34%)	.571
	Right middle lobe	14 (10%)	9 (6%)	
	Right lower lobe	21 (14%)	26 (18%)	
	Left upper lobe	38 (26%)	34 (23%)	
	Left lower lobe	20 (14%)	27 (19%)	
Tumor Histology	Adenocarcinoma	113 (78%)	103 (71%)	.035
	Squamous cell	29 (20%)	30 (21%)	
	Non–small cell cancer not otherwise specified	3 (2%)	13 (9%)	
Tumor Stage	0[a]	4 (3%)	0 (0%)	<.001
	I	89 (61%)	100 (68%)	
	II	28 (19%)	13 (9%)	
	III	21 (14%)	17 (12%)	
	IV	3 (2%)	16 (11%)	
Recurrence/Progression	Yes	40 (28%)	57 (39%)	.038
	No	105 (72%)	89 (61%)	

Variables shown as median (range) or number (%).
[a] Pathological stage 0 disease is defined as a carcinoma in situ (TisN0M0) as per the American Joint Committee on Cancer (AJCC) 7th edition staging system.

Table 3. Inter- and Intraobserver Variability in Metabolic Tumor Volume (MTV) PET-edge Segmentations

Observer[a]	Dice Similarity Coefficient (DSC)	Mean Absolute Boundary Distance (MAD, mm)	Absolute Volume Difference (mL)[b]
A vs a (Intra)	0.916 (0.090)	0.548 (0.544)	0.71 (1.66)
A vs B (Inter)	0.917 (0.087)	0.559 (0.507)	0.58 (0.92)
a vs B (Inter)	0.904 (0.105)	0.628 (0.631)	0.79 (1.46)

All values are the mean (standard deviation).
[a] Observer 1 contoured each tumor twice (A and a) and observer 2 contoured each lesion once (B).
[b] For reference, the average [range] volumes of all MTV contours by the three observers were 15.4 [0.4–297.8], 15.3 [0.4–296.9], and 15.3 [0.3–296.0] mL.

Table 4. Intraclass Correlation Coefficients for All FDG-PET Radiomic Features

	MTV		Penumbra		MTV + Penumbra	
Feature Type	Inter-	Intra-	Inter-	Intra-	Inter-	Intra-
Size	0.996 (0.99–1.00)	0.994 (0.99–1.00)	–	–	–	–
Intensity	0.977 (0.89–1.00)	0.972 (0.84–1.00)	0.931 (0.48–0.99)	0.916 (0.36–0.99)	0.995 (0.98–1.00)	0.995 (0.98–1.00)
Shape	0.867 (0.37–0.98)	0.847 (0.39–0.98)	–	–	–	–
Texture	0.898 (0.50–0.99)	0.893 (0.48–0.99)	0.892 (0.14–0.99)	0.925 (0.50–0.99)	0.981 (0.28–1.00)	0.977 (0.66–1.00)

All values are shown as the mean (range).

Table 5. Number (percent) of Robust FDG-PET Radiomic Features Selected in Each Category by Virtue of an ICC > 0.8

Feature Type	MTV		Penumbra		MTV + Penumbra	
	Inter-	Intra-	Inter-	Intra-	Inter-	Intra-
Size	4 (100%)	4 (100%)	–	–	–	–
Intensity	12 (100%)	12 (100%)	11 (92%)	11 (92%)	12 (100%)	12 (100%)
Shape	27 (68%)	30 (75%)	–	–	–	–
Texture	115 (80%)	115 (80%)	118 (82%)	131 (91%)	144 (100%)	142 (99%)

of 435 of the 512 features (85%) had an ICC >0.8 (Table 5) and were considered robust to differences in the segmentations (22, 23).

Feature Selection and Model Training

Across the 100 randomizations, the average minimum cross-validation error was 10.5% at a lambda value of 0.1296 in the training cohort. This lambda generated 2 features with nonzero coefficients, stage, and 1 MTV plus penumbra GLCM texture feature (maximum probability). Although SUV_{max} has previously been shown to be associated with recurrence in NSCLC, it was not selected by LASSO as a top feature. However, it was found to be a significant univariate predictor in our cohort (Table 6), consistent with previous studies (7).

Figure 1 visualizes the Pearson correlation coefficients of the top features. For reference, correlation of the top features with MTV volume and SUV_{max} is also shown. All correlations were low and the radiomic feature showed no correlation with stage, volume, or SUV_{max}.

Univariate Cox regression model statistics, including the AIC, likelihood ratios, *P*-values, and HRs, are shown for the top features in Table 6. Both features were significant univariate predictors of time to recurrence. Overall, stage was the best univariate predictor.

Because stage was the best univariate predictor, the likelihood ratio test was performed to assess significant improvements to this well-established clinical model for recurrence prediction. Additional features were added to determine significant improvements to the model. Adding the MTV plus penumbra texture feature to stage significantly improved the model ($P = .006$). This multivariate model was a significant predictor of time to recurrence in the training cohort (likelihood ratio = 27.59, $P < .001$, concordance = 0.74 [95% CI: 0.66-0.81]). Both stage (HR = 1.92 [95% CI: 1.37–2.67], $P < .001$) and the radiomic texture feature (HR = 0.52 [95% CI: 0.30–0.91], $P = .02$) were significant covariates in the multivariate model. Adding SUV_{max} to stage did not significantly improve the clinical model performance ($P = .22$). It also did not significantly improve performance in the combined stage and radiomic model ($P = .73$).

Model Validation

Univariate results were confirmed in the validation cohort (Table 7), with all features being significant predictors of time to recurrence. The locked multivariate model from the training cohort, which included stage and the radiomic texture feature, was a significant predictor in the validation cohort (concordance = 0.74 [95% CI: 0.67–0.81], Noether's $P < .001$). We separated the patients into high- and low-risk groups on the basis of the median risk score in the training cohort. Kaplan–Meier time-to-recurrence curves for the multivariate model in both cohorts are shown in Figure 2. Recurrence was lower in the group below the median model risk score.

The multivariate model including stage and the radiomic feature significantly outperformed the best performing clinical model of stage in the training ($P = .036$) and validation ($P = .033$) cohorts. The combined model also outperformed the radiomic feature alone in both the training cohort ($P = .019$) and the validation cohort ($P < .001$).

Figure 3 exemplifies 2 patients with similar SUV_{max} that would typically be considered to be at a high risk of recurrence.

Table 6. Cox Proportional Hazards Model Statistics for Univariate Features in the Training Cohort

Feature	Akaike Information Criterion	Likelihood Ratio	*P*-value	HR [95% CI]	Concordance [95% CI]
Stage	341.7	19.98	<.001	2.15 [1.56–2.95]	0.68 [0.60–0.76]
Gray-level Cooccurrence Matrix Maximum Probability (MTV + Penumbra)	347.5	14.18	<.001	0.41 [0.23–0.74]	0.66 [0.57–0.74]
SUV_{max}	353.7	7.99	.005	1.06 [1.02–1.10]	0.67 [0.58–0.75]

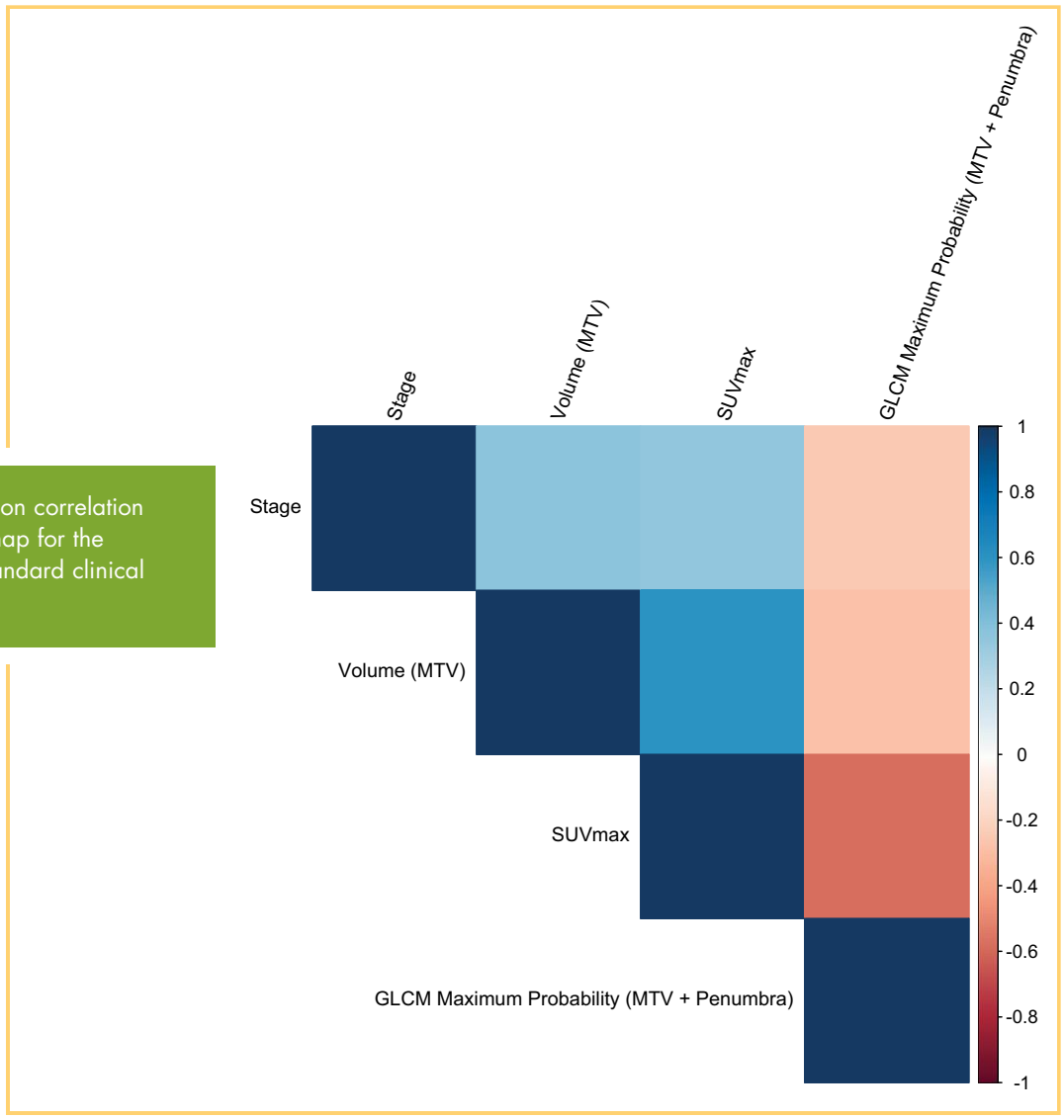


Figure 1. Pearson correlation coefficient heatmap for the radiomic and standard clinical variables.

Yet, the combined model including radiomics correctly predicted the recurrence status of each patient on the basis of the median risk value. Based on qualitative inspection, the high-risk patient had more heterogeneous uptake in the penumbra region compared with the low-risk patient.

DISCUSSION

We show here evidence that texture in the MTV and nearby surrounding region can predict recurrence in NSCLC. Furthermore, augmenting this radiomic feature with stage significantly improved performance over stage alone, which was validated in

Table 7. Cox Proportional Hazards Model Statistics for Univariate Features in the Validation Cohort

Feature	Akaike Information Criterion	Likelihood Ratio	*P*-value	HR [95% CI]	Concordance [95% CI]
Stage	475.6	35.7	<.001	2.13 [1.69–2.68]	0.69 [0.63–0.76]
Gray-level Cooccurrence Matrix Maximum Probability (MTV + Penumbra)	497.2	14.14	<.001	0.50 [0.33–0.76]	0.66 [0.60–0.72]
SUV_{max}	506.1	5.24	.02	1.03 [1.01–1.05]	0.67 [0.61–0.73]

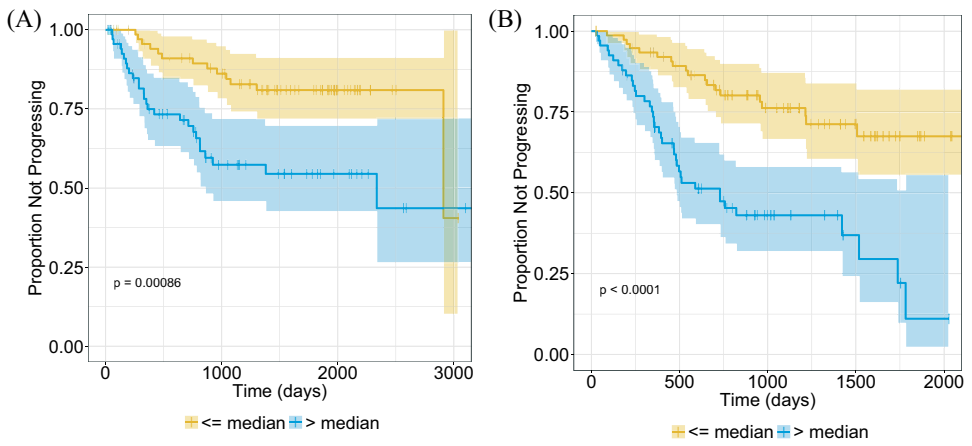


Figure 2. Kaplan–Meier curves for the multivariate stage and radiomic texture model risk scores in the training cohort (n = 145, $P < .001$) (A) and the validation cohort (n = 146, $P < .001$) (B). Patients have been stratified on the basis of median risk value in the training cohort. The shaded regions represent the 95% confidence intervals (CI) and "+" indicates censored data.

an independent data set. This model also showed potential value in risk-stratifying patients with NSCLC who are at high versus low risk of recurrence or progression. A general rule in modeling studies is that 10 patients are needed for every feature selected in the model (8). To minimize overfitting, our final model consisted of only 2 features. However further studies on larger sample sizes with additional features may improve prognostic performance and applicability to other cohorts.

The radiomic feature selected was a GLCM texture feature in the combined MTV plus penumbra volume. This feature, which describes local texture variations, suggests that patients whose PET images show a more heterogeneous texture, specifically in the penumbra region surrounding the MTV, are more likely to recur. This suggests the importance of image data in the surrounding region for recurrence prediction. This region may contain uptake not measured in the MTV (and not by the SUV_{max}) and could indicate areas of disease adjacent to the primary mass. The texture being detected in this region may be indicative of an invasive component of the tumor, for example, spiculations or tumor spread through blood vessels, but this requires further investigation (15).

Notably, size or shape features, including the commonly used metrics of maximum axial diameter and 3D volume, were not selected as predictive features. SUV_{max} was also not selected, and adding it to clinical or combined models did not significantly improve performance. This suggests that texture features may provide more useful information than traditional metrics for predicting recurrence/progression.

Previous work in the field of radiomics has evaluated FDG-PET features for outcome prediction in lung cancer. Jansen et al. found the GLCM energy texture feature was a significant predictor of overall survival in oligometastatic NSCLC (26). Others have shown that texture features may be beneficial for predicting local control, distant metastasis, and disease-free survival in lung cancer (10-12). However, the majority of studies to date have focused on only the MTV. To the best of our knowledge,

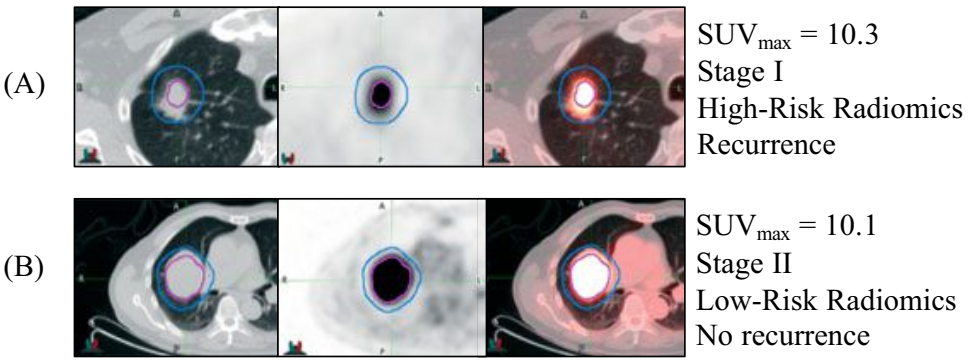


Figure 3. Example computed tomography (CT) image (left), corresponding positron emission tomography (PET) image (middle), and fused PET/CT images (right) for 2 patients, where the metabolic tumor volume (MTV) is encircled in magenta and the penumbra in between the magenta and blue outlines. Patients (A) and (B) had relatively high SUV_{max} values, but the radiomics model distinguished the high-risk patient (A) who recurred at 16-month follow-up and the low-risk patient (B) who had not recurred at just under 5 years of follow-up.

ours is the first study that evaluates the lung tumor penumbral region of PET images for recurrence prediction. Future work integrating CT imaging features or molecular data may improve prognostic performance.

Our study investigated PET/CT images from multiple scanners and institutions, potentially introducing variability in image data and quality and therefore the construction of a predictive model. We used a standard acquisition protocol across all institutions to minimize this variability (27, 28). This may still result in signal variations in the tumor and penumbra regions; therefore, further studies investigating single scanners are warranted and may improve model performance.

Previous work has also shown that PET radiomic features are dependent more on delineation variability than on reconstruction algorithm (29) and that texture features are less affected by difference in scanners (30). Many radiomic features also show high test–retest stability with repeat PET imaging (31). The PET-edge segmentation tool we used for tumor segmentation showed high reproducibility with associated radiomic feature robustness. Segmentations were performed with commercially available software (MIM Software, Inc.), making it an easily deployed and integrated system.

Our work is also applicable in a "real world," nonresearch setting, where different scanners and images of variable quality are routinely used for clinical assessment. However, additional external validation of this radiomics model is warranted to determine the impact of different scanners and acquisition protocols on model predictions.

Our study has several limitations. The primary limitation is that the penumbra region was not restricted to the lung volume, that is, it may at times have included the adjacent chest wall, major blood vessels, and/or mediastinum. However, as features were selected from within this region, it is providing relevant information for the prediction of recurrence. The effect of this and the efforts to minimize it remain the subject of further investigation. Owing to differences in breathing between the PET and CT images, accurate registration of the lung boundary is challenging. We also investigated only a single distance of 1 cm for the penumbra region; it is possible that larger or smaller distances could improve or degrade performance. Another limitation is the inherent low resolution of the PET images, limiting the amount of information we can analyze for each tumor owing to lower voxel quantities for smaller tumors. Finally, the sample sizes analyzed were relatively small, and validation of this model in larger data sets is warranted.

In conclusion, a PET texture feature in the metabolic tumor volume and surrounding region augmented staging for NSCLC recurrence prediction. This model may be useful in identifying patients who are at a higher risk of recurrence or progression and may assist physicians in determining what patients may benefit from adjuvant or personalized treatment options at the time of diagnosis.

ACKNOWLEDGMENTS

Equal contribution: "S.N and V.S.N contributed equally to this work."

The authors would like to acknowledge MIM Software Inc. for their assistance with segmentation software, Jalen Benson and Weiruo Zhang for their assistance with clinical data curation, and the Stanford Data Studio for statistical consulting. The authors would like to acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (NSERC) Postdoctoral Fellowship and the National Cancer Institute (NCI) R01 CA160251, U01 CA187947, and U01 CA196405.

Disclosures: Dr. Sandy Napel is a Consultant for Carestream Health Inc., on the Medical Advisory Board for Fovia Inc., a Scientific Advisor for EchoPixel Inc., and a Scientific Advisor for RADLogics Inc. However, he is not an employee of any of these companies and none of these conflicts are related to the data used and research completed in this manuscript.

Conflict of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. Lang-Lazdunski L. Surgery for nonsmall cell lung cancer. Eur Respir Rev. 2013; 22:382–404.
2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J Clin. 2018; 68:7–30.
3. Uramoto H, Tanaka F. Recurrence after surgery in patients with NSCLC. Transl Lung Cancer Res. 2014;3:242.
4. Consonni D, Pierobon M, Gail MH, Rubagotti M, Rotunno M, Goldstein A, Goldin L, Lubin J, Wacholder S, Caporaso NE, Bertazzi PA, Tucker MA, Pesatori AC, Landi MT. Lung cancer prognosis before and after recurrence in a population-based setting. J Natl Cancer Inst. 2015;107:djv059.
5. Ettinger DS, Wood DE, Akerley W, Bazhenova LA, Borghaei H, Camidge DR, et al. Non–small cell lung cancer, Version 6.2015. J Natl Compr Canc Netw. 2015;13:515–524.
6. Pignon J-P, Tribodet H, Scagliotti GV, Douillard J-Y, Shepherd FA, Stephens RJ, Le Chevalier T. Lung adjuvant cisplatin evaluation: a pooled analysis by the LACE Collaborative Group. J Clin Oncol. 2008;26:3552–3559.
7. Liu J, Dong M, Sun X, Li W, Xing L, Yu J. Prognostic value of 18F-FDG PET/CT in surgical non-small cell lung cancer: a meta-analysis. PLoS One. 2016;11: e0146195.
8. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology. 2015;278:563–577.
9. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, Zegers CM, Gillies R, Boellard R, Dekker A, Aerts HJ. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48:441–446.
10. Takeda K, Takanami K, Shirata Y, Yamamoto T, Takahashi N, Ito K, Takase K2, Jingu K. Clinical utility of texture analysis of 18F-FDG PET/CT in patients with Stage I lung cancer treated with stereotactic body radiotherapy. J Radiat Res. 2017;58:862–869.
11. Kirienko M, Cozzi L, Antunovic L, Lozza L, Fogliata A, Voulaz E, Rossi A, Chiti A, Sollini M. Prediction of disease-free survival by the PET/CT radiomic signature in non-small cell lung cancer patients undergoing surgery. Eur J Nucl Med Mol Imaging. 2018:45:207–217.
12. Wu J, Aguilera T, Shultz D, Gudur M, Rubin DL, Loo Jr BW, et al. Early-stage non–small cell lung cancer: quantitative imaging characteristics of 18F fluorodeoxyglucose PET/CT allow prediction of distant metastasis. Radiology. 2016;281:270–278.
13. Travis WD, Brambilla E, Noguchi M, Nicholson AG, Geisinger KR, Yatabe Y, Beer DG, Powell CA, Riely GJ, Van Schil PE, Garg K, Austin JH, Asamura H, Rusch VW, Hirsch FR, Scagliotti G, Mitsudomi T, Huber RM, Ishikawa Y, Jett J, Sanchez-Cespedes M, Sculier JP, Takahashi T, Tsuboi M, Vansteenkiste J, Wistuba I, Yang PC, Aberle D, Brambilla C, Flieder D, Franklin W, Gazdar A, Gould M, Hasleton P, Henderson D, Johnson B, Johnson D, Kerr K, Kuriyama K, Lee JS, Miller VA, Petersen I, Roggli V, Rosell R, Saijo N, Thunnissen E, Tsao M, Yankelewitz D. International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society International Multidisciplinary Classification of Lung Adenocarcinoma. J Thorac Oncol. 2011;6:244–285.
14. Kadota K, Nitadori J-i, Sima CS, Ujiie H, Rizk NP, Jones DR, Adusumilli PS, Travis WD. Tumor spread through air spaces is an important pattern of invasion and impacts the frequency and location of recurrences after limited resection for small stage I lung adenocarcinomas. J Thorac Oncol. 2015;10:806–814.

15. Ren J, Zhou J, Ding W, Zhong B. Clinicopathological characteristics and imaging features of pulmonary adenocarcinoma with micropapillary pattern. Zhonghua Zhong Liu Za Zhi. 2014;36:282–286. [Article in Chinese]
16. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, Tarbox L, Prior F. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging. 2013;26:1045–1057.
17. Bakr S, Gevaert O, Echegaray S, Ayers K, Zhou M, Shafiq M, Zheng H, Benson JA, Zhang W, Leung ANC, Kadoch M, D Hoang C, Shrager J, Quon A, Rubin DL, Plevritis SK, Napel S. A radiogenomic dataset of non-small cell lung cancer. Sci Data. 2018;5:180202.
18. Echegaray S, Bakr S, Rubin DL, Napel S. Quantitative Image Feature Engine (QIFE): an open-source, modular engine for 3D quantitative feature extraction from volumetric medical images. J Digit Imaging. 2018;31:403–414.
19. Haralick RM. Statistical and structural approaches to texture. Proceedings of the IEEE. 1979;67:786–804.
20. Haralick RM, Shanmugam K. Textural features for image classification. IEEE Trans Cybern. 1973:610–621.
21. Leijenaar RT, Nalbantov G, Carvalho S, Van Elmpt WJ, Troost EG, Boellaard R, Aerts HJ, Gillies RJ, Lambin P. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. Sci Rep. 2015;5:11075.
22. Parmar C, Velazquez ER, Leijenaar R, Jermoumi M, Carvalho S, Mak RH. Robust radiomics feature quantification using semiautomatic volumetric segmentation. PLoS One. 2014;9:e102107.
23. Pavic M, Bogowicz M, Würms X, Glatz S, Finazzi T, Riesterer O, Roesch J, Rudofsky L, Friess M, Veit-Haibach P, Huellner M, Opitz I, Weder W, Frauenfelder T, Guckenberger M, Tanadini-Lang S. Influence of inter-observer delineation variability on radiomics stability in different tumor sites. Acta Oncol. 2018;57:1070–1074.
24. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996:267–88.
25. Team RCR. A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013.
26. Jensen GL, Yost CM, Mackin DS, Fried DV, Zhou S, Gomez DR. Prognostic value of combining a quantitative image feature from positron emission tomography with clinical factors in oligometastatic non-small cell lung cancer. Radiother Oncol. 2018;126:362–367.
27. Nyflot MJ, Yang F, Byrd D, Bowen SR, Sandison GA, Kinahan PE. Quantitative radiomics: impact of stochastic effects on textural feature analysis implies the need for standards. J Med Imaging (Bellingham). 2015;2:041002.
28. Beichel RR, Smith BJ, Bauer C, Ulrich EJ, Ahmadvand P, Budzevich MM, Gillies RJ, Goldgof D, Grkovski M, Hamarneh G, Huang Q, Kinahan PE, Laymon CM, Mountz JM, Muzi JP, Muzi M, Nehmeh S, Oborski MJ, Tan Y, Zhao B, Sunderland JJ, Buatti JM. Multi-site quality and variability analysis of 3D FDG PET segmentations based on phantom and clinical image data. Med Phys. 2017;44:479–496.
29. van Velden FH, Kramer GM, Frings V, Nissen IA, Mulder ER, de Langen AJ, Hoekstra OS, Smit EF, Boellaard R. Repeatability of radiomic features in non-small-cell lung cancer [18F] FDG-PET/CT studies: impact of reconstruction and delineation. Mol Imaging Biol. 2016;18:788–795.
30. Tsujikawa T, Tsuyoshi H, Kanno M, Yamada S, Kobayashi M, Narita N, Kimura H, Fujieda S, Yoshida Y, Okazawa H. Selected PET radiomic features remain the same. Oncotarget. 2018;9:20734.
31. Leijenaar RT, Carvalho S, Velazquez ER, Van Elmpt WJ, Parmar C, Hoekstra OS, Boellaard R, Dekker AL, Gillies RJ, Aerts HJ, Lambin P. Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta Oncol. 2013;52:1391–1397.

TOMOGRAPHY® RESEARCH ARTICLE

Bias in PET Images of Solid Phantoms Due to CT-Based Attenuation Correction

Darrin W. Byrd[1], John J. Sunderland[2], Tzu-Cheng Lee[1], and Paul E. Kinahan[1]

[1]Department of Radiology, University of Washington, Seattle, WA; and [2]Department of Radiology, University of Iowa, Iowa City, IA

Corresponding Author:
Paul E. Kinahan, PhD
206-543-0236, University of Washington,
Box 357987, Seattle, WA 98195;
E-mail: kinahan@uw.edu

Key Words: PET, calibration, phantoms
Abbreviations: Computed tomography (CT), positron emission tomography (PET), Germanium isotope 68 (^{68}Ge), Computerized Imaging Reference Systems, Incorporated (CIRS), Image quality phantom (IQ), kilovolt potential (kVp), National Electrical Manufacturers Association (NEMA)

ABSTRACT

The use of computed tomography (CT) images to correct for photon attenuation in positron emission tomography (PET) produces unbiased patient images, but it is not optimal for synthetic materials. For test objects made from epoxy, image bias and artifacts have been observed in well-calibrated PET/CT scanners. An epoxy used in commercially available sources was infused with long-lived ^{68}Ge/^{68}Ga nuclide and measured on several PET/CT scanners as well as on older PET scanners that measured attenuation with 511-keV photons. Bias in attenuation maps and PET images of phantoms was measured as imaging parameters and methods varied. Changes were made to the PET reconstruction to show the influence of CT-based attenuation correction. Additional attenuation measurements were made with a new epoxy intended for use in radiology and radiation treatment whose photonic properties mimic water. PET images of solid phantoms were biased by between 3% and 24% across variations in CT X-ray energy and scanner manufacturer. Modification of the reconstruction software reduced bias, but object-dependent changes were required to generate accurate attenuation maps. The water-mimicking epoxy formulation showed behavior similar to water in limited testing. For some solid phantoms, transformation of CT data to attenuation maps is a major source of PET image bias. The transformation can be modified to accommodate synthetic materials, but our data suggest that the problem may also be addressed by using epoxy formulations that are more compatible with PET/CT imaging.

INTRODUCTION

With proper calibration, positron emission tomography (PET) accurately quantifies the concentration of radiolabeled molecules in patients noninvasively and with excellent sensitivity. Biomarkers computed from these measured concentrations have proven utility in managing the treatment of certain cancers (1-4). However, the acquisition of PET data is a physically complicated process, and the software required to convert the raw data to form an image relies on numerous approximations and empirical corrections. Poor calibration or nonoptimal processing of the data leads to biased images (5-7). This bias may reduce PET's prognostic value for patients and researchers (8).

One of the most important effects that the reconstruction must model is the interaction of 511-keV annihilation photons with tissues (in patients) or other materials (in calibration objects, which are commonly called "phantoms"). Without mathematical corrections, absorption of photons leads to reduced signal from central regions of PET images as well as edge artifacts. Scattered photons also affect raw PET data because PET's coincidence detection, which does not use physical collimation, cannot distinguish between scattered and unscattered photons for small deflections and therefore misplaces them in the raw projection data.

For modern PET scanners, the corrections for scattered and absorbed photons are calculated from computed tomography (CT) images that are acquired just before or after the PET scan (9, 10). CT volume images are mapped to attenuation images, commonly via a piecewise-linear transformation (11), whose final units are "attenuation coefficients" that represent the probability of an annihilation photon being "attenuated" (absorbed or scattered) per unit length. With these attenuation images, the scanner is able to estimate the required data corrections for scatter and absorption that are applied during the reconstruction.

However, CT-based attenuation correction suffers from a known limitation in that there is no unique relationship between CT pixelwise image values (Hounsfield Units) and attenuation coefficients at the energy of PET photons. Figure 1 shows this problem. The disparity in the absorption properties of bone and soft tissue varies with photon energy, and it is much greater at lower CT photon energies than at PET energy. The piecewise-linear transformation succeeds in producing sufficiently accu-

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2018.00043

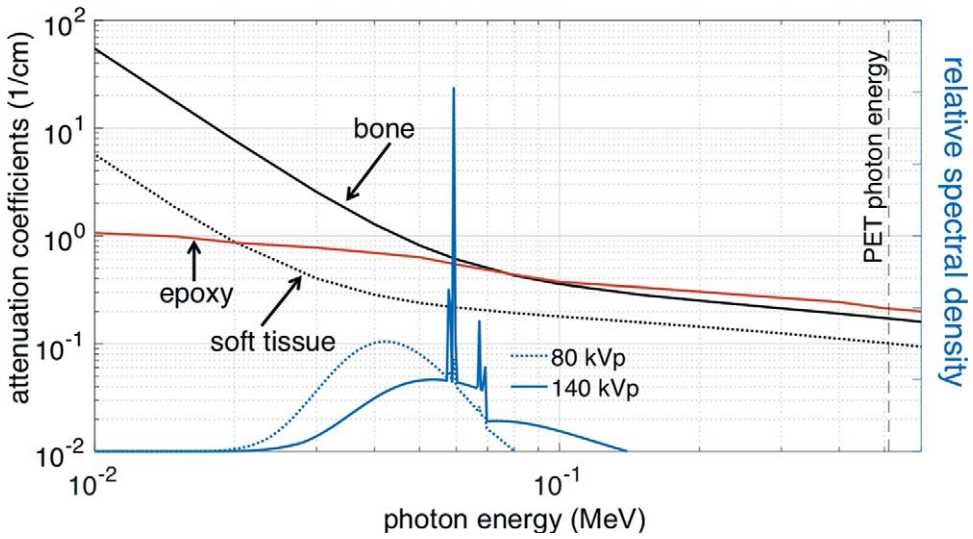


Figure 1. Attenuation coefficients for bone and soft tissues (black lines) and filtered bremsstrahlung spectra for 80- and 140-kVp computed tomography (CT) scans (blue lines). Tissue attenuation values are from the National Institute of Standards and Technology (12). X-ray energies are from the Catsim software package (13). Epoxy composition is from the PubChem database (14).

rate attenuation images because human tissues have predictable chemical compositions and their CT image values can be coarsely grouped into soft tissue and bone. This allows bone to be scaled separately, thus avoiding overestimation of Compton scatter at PET's 511-keV energy. Figure 2 shows a continuous, piecewise-linear transformation used in a modern PET/CT scanner.

Tests of quantitative accuracy for PET scanners often involve phantoms that are carefully designed to validate a PET/CT scanner's correction of all physical effects, including detection sensitivity and random coincidences (15, 16). Among current published standards and accreditation organizations, water-filled phantoms containing short-lived radionuclides predominate (17, 18). These phantoms' physical properties are well-matched to patient scans, to the extent that water-filled phantom scans are routinely used to measure the calibration factors used to convert clinical scans from a scanner's arbitrary units to true nuclide concentration. However, recent reports suggest that this calibration process may result in increased variability in PET signal, likely owing to difficulties in repeatedly refilling short-lived phantoms each time the phantom is used (19).

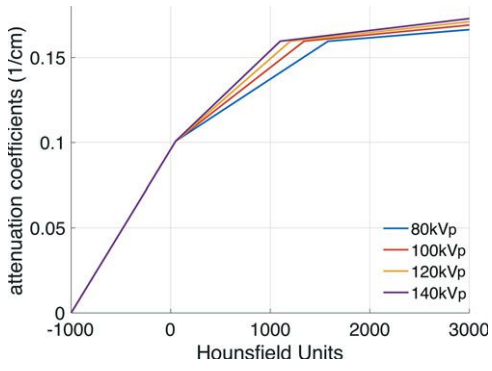


Figure 2. Estimated attenuation coefficients at 511 keV versus Hounsfield Units from CT images with varying characteristic voltage. Coefficients were copied from a modern clinical positron emission tomography (PET)/CT scanner.

This problem can potentially be solved by using phantoms infused with long-lived radionuclides, which can be measured repeatedly without refilling. These phantoms follow highly predictable decay curves, allowing bias to be computed at multiple time points with fewer confounding factors. In this work, we investigate an important drawback of long-lived phantoms: they are usually constructed from solid materials to mitigate the risk of spilling, and these solid materials have attenuation properties that are not accurately estimated by the CT-based attenuation estimation used for human tissues (Figure 2). This can lead to image bias. Below, we examine the bias in attenuation images and reconstructed PET images of several solid long-lived phantoms, and we show that CT-based attenuation correction underestimates photon absorption by the epoxy used in their construction.

METHODOLOGY

Phantoms were constructed using epoxy with and without the admixture of long-lived positron-emitting ^{68}Ge/^{68}Ga (^{68}Ge). Phantoms were imaged by PET/CT and by 511-keV transmission scans. Modifications to the X-ray tube voltage and reconstruction software were used to investigate the dependence of PET image bias on CT-based attenuation correction.

Image Quality Phantom

A National Electrical Manufacturers Association (NEMA) Image Quality (IQ) phantom (15) (Data Spectrum Corporation, Durham, NC) was filled with solid epoxy and ^{68}Ge at an initial background concentration of 7.19 kBq/mL. The phantom, shown in Figure 3A, contained spherical inserts at the sizes specified in the NEMA test standard, but with the modification that all spheres were filled to the same concentration of radionuclide (ie, the phantom contained no nonradioactive spheres). Sphere contrast was 7.7:1 relative to background, and the same epoxy formula was used to fill both background and spheres. The phantom was scanned on 3 commercial PET/CT scanners: a Discovery STE (General Electric Healthcare, Waukesha, WI), a

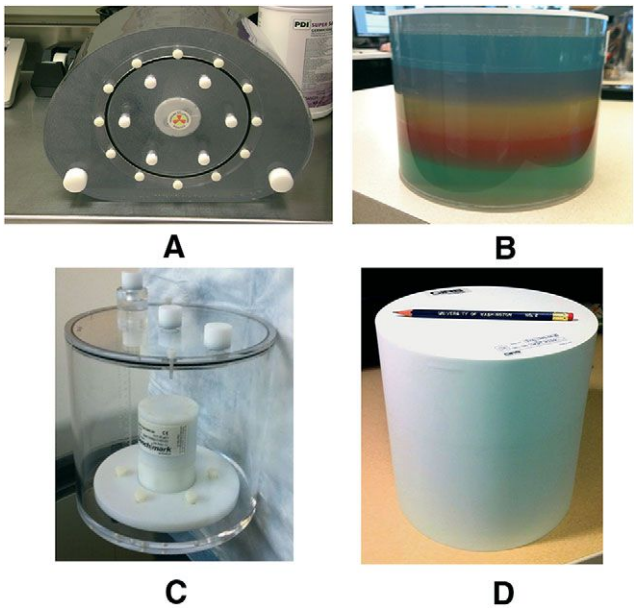


Figure 3. Image Quality (IQ) phantom filled with epoxy infused with ^{68}Ge (A). Nonradioactive epoxy phantom made from the same epoxy as the IQ phantom and the X-Cal phantom (B). X-Cal phantom mounted for scanning in an American College of Radiology flood phantom (C). CIRS 20cm *Plastic Water LR* phantom (D).

Biograph (Siemens Healthcare, Knoxville, TN), and a Philips Gemini TF Big Bore (Philips Healthcare, Best, The Netherlands.). The dependence of reconstructed signal on CT X-ray energy (commonly characterized by the potential in kilovolts [kVp] applied to the X-ray tube) and variations in the reconstruction software were investigated.

Nonradioactive Epoxy Phantom

A 20-cm cylinder (Figure 3B) was filled with the same epoxy used in the IQ phantom, but without the addition of any radionuclide. This phantom was scanned on a General Electric Discovery LS and a Siemens HR+. Both scanners used positron sources to measure photon absorption at the same energy measured in clinical PET scans, 511 keV. Filtered backprojection was used to generate attenuation images. The phantom was also CT-scanned on the General Electric Discovery STE scanner, and attenuation images resulting from 80-, 100-, 120-, and 140-kVp CT scans were copied from the scanner console and read in MATLAB (MathWorks, Natick, MA).

X-Cal Phantom

A comercially-available 45-mm cylinder phantom was scanned inside a 20-cm water-filled American College of Radiology flood phantom (Figure 3C) on the same Discovery STE as the IQ phantom. The 45-mm phantom is sold as part of a "cross-calibration" kit sold by RadQual, LLC (Weare, NH) and we consequently refer it to it as the X-Cal phantom. It was made from the same epoxy as the above sources. We have previously reported on its signal properties and the bias between measured values and known tracer concentration (19, 20).

Nonradioactive PlasticWater Phantom

A nonradioactive 20-cm-diameter cylinder (Figure 3D) was constructed from a different epoxy that was formulated to better match the attenuation properties of human tissues. The cylinder

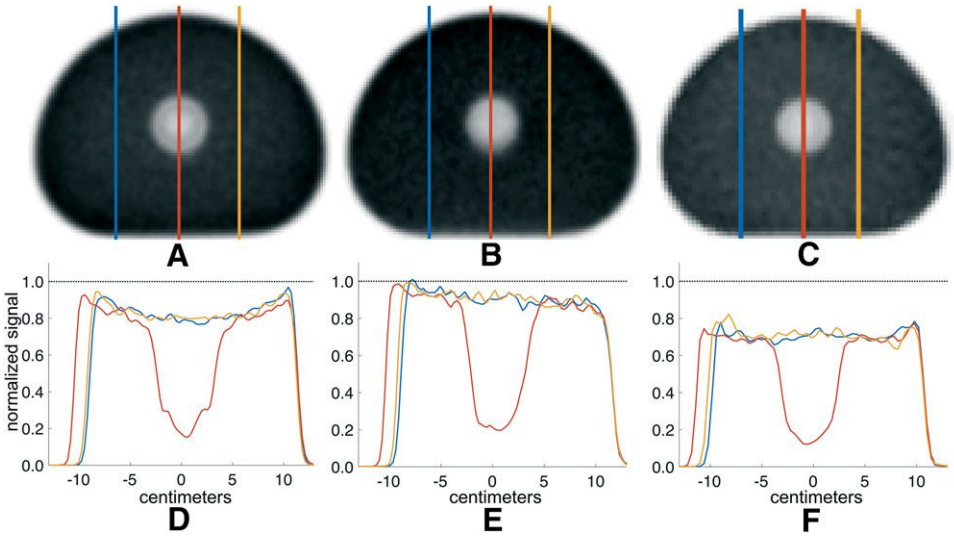


Figure 4. Images and profiles of PET signal from the IQ phantom in three scanners. Signal has been normalized by the known nuclide concentration, with truth represented by the horizontal black dotted line. The 3 columns show scanner models from 3 manufacturers: a General Electric Discovery STE (A and D), a Siemens Biograph (B and E), and a Philips Gemini TF Big Bore (C and F). Data were averaged over 3 cm axially and acquired over a decay-compensated duration of 60 minutes. Colored lines in the images correspond to the locus of points shown in the profiles. Color windows are matched between images.

Table 1. PET Signal (measured/known) in the Background Region of the IQ Phantom with Varying X-ray Tube Voltage and Modified CT Rescaling for 2 of the Scanners in this Study

kVp	Siemens	General Electric
80	0.92	0.75
100	0.97	0.78
120	0.95	0.80
140	0.95	0.82
120 (mod'd AC)		1.04

was constructed by Computerized Imaging Reference Systems, Incorporated, or CIRS (Norfolk, VA), and was filled with their "LR" epoxy. The phantom was scanned on a General Electric Discovery LS to generate 511-keV transmission images as well as CT-based attenuation images.

Variations in Imaging Parameters

CT photon energies were varied by changing the X-ray voltage in CT scans of the IQ phantom and the nonradioactive phantoms. CT voltage modifications were done using the scanners' user-facing interfaces and spanned the available settings of 80, 100, 120 and 140 kVp. Where possible, the impact of CT voltage on PET measured radioactivity concentration was assessed.

Modification of CT-Based Attenuation Correction

An additional variation for the data measured on the General Electric Discovery STE scanner was the modification of the rescaling functions shown in Figure 2 to provide more accurate conversion of the CT images to attenuation images for the IQ and X-Cal phantoms. In particular, for the domain that contained the phantoms' image values of ~80–90 Hounsfield Units, the rescaling coefficient (slope) for 120-kVp CT images was increased. The new coefficients for the 2 phantoms were chosen to make the resulting attenuation images agree with values obtained from the 511-keV transmission scans of the phantoms. Modifications to the attenuation conversion were made in MATLAB, and reconstructions were performed using code from General Electric.

RESULTS

The IQ phantom PET images demonstrated quantitative bias. For all 3 PET/CT scanners, the bias was spatially variable. Figure 4 shows data from axially averaged (ie, thick-slice) images from the 3 scanners. For each scanner, the figure depicts data from ~60-minutes' worth of scanning (scan durations were corrected to a common time point to compensate for phantom decay).

Table 1 shows the PET background signal divided by the known nuclide concentration for the images in Figure 4. Background signal was computed as in the NEMA standard using 28-mm regions (15).

Figure 5 shows the attenuation image from the transmission scan of the nonradioactive epoxy phantom (Figure 3B) on the GE Discovery LS. The values obtained on the Siemens HR+ were similar. Figure 5 also shows profiles through this transmission image as well as the attenuation values that were estimated from CT data on the General Electric Discovery STE. The profiles show that the CT-based attenuation images do not agree with the values obtained using positron annihilation photons. While varying X-ray tube voltage does lead to varying CT signal, no user-selectable tube voltage led to agreement between the CT-based attenuation images and the transmission image. Region of interest means in the CT-based attenuation images were 0.095, 0.096, 0.097, and 0.098 cm^{-1} as the CT voltage varied. In the transmission scan, the value was 0.105 cm^{-1}.

Figure 6A shows the attenuation images generated in the reconstruction before and after our modification of the algorithm. Figure 6B, shows PET data reconstructed with each attenuation image. It can be seen that the accuracy of the signal is improved. The bottom row of Table 1 shows that the modifications to the attenuation correction lead to more accurate PET signal.

Figure 7A shows a transaxial slice containing the spherical inserts in an image made with the modified attenuation correction algorithm. Figure 7B shows mean signal from regions of interest drawn on the spheres. Averaged over sphere sizes, the signal was 1.20 times larger in the images with modified attenuation correction, indicating that if solid phantoms are used for

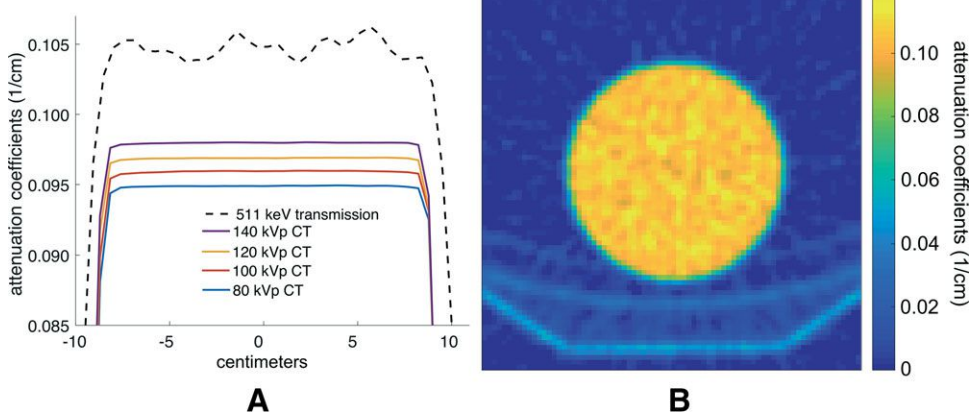


Figure 5. Profiles showing CT-based attenuation estimates in the solid epoxy phantom as well as 511-keV transmission measurement (dashed line) (A).The attenuation image produced using 511-keV photons (B).

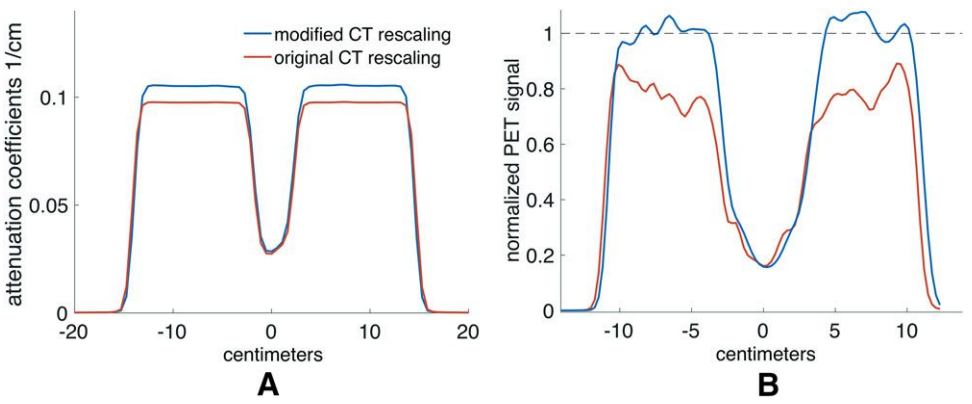


Figure 6. Profiles through CT-based attenuation images before and after our modifications to the CT attenuation correction algorithm (A). Profiles through the reconstructed PET images made with the pre- and postmodification attenuation data, showing improved signal accuracy (B). PET signal has been divided by known phantom background activity concentration.

resolution measurements, some compensation for attenuation bias must be applied.

Figure 8 shows the signal from PET images of the X-Cal source before and after modification to the attenuation correction. With the modified algorithm, the signal is visibly more accurate. Using XCaliper (20), a previously reported method for drawing regions of interest in the X-Cal phantom, the bias was measured as −4.7% using the standard attenuation correction algorithm and 0.5% with our modifications. Different scale factors were used in the respective modifications to the CT rescaling for the X-Cal and IQ phantoms.

Figure 9 shows profiles through transmission scans of the PlasticWater epoxy phantom and a similarly sized water-filled cylinder. Also shown are CT-based attenuation estimates. It can be seen that the PlasticWater epoxy better matches the water values in the transmission scans. In addition, the bias between the transmission scan and the CT-based attenuation images is similar to that seen in an actual water phantom, as Table 2 also shows.

DISCUSSION

Inaccuracy of PET activity concentration measurements in solid epoxy phantoms has been previously observed and is a challenge to their use in determining scanner calibration accuracy. We have investigated signal bias in solid phantoms made from an epoxy that is used in commercially available sources. While several factors may affect long-lived phantom bias, such as scatter correction and prompt gamma emission by ^{68}Ge, our results show that bias is greatly reduced by modification to the CT-based attenuation correction algorithm.

Although we did not evaluate attenuation images for all scanners used, Figure 4 shows that each scanner exhibited signal bias over all or part of the phantom images. It is expected that scanners from different manufacturers would behave similarly, as the generation of X-rays, and therefore the transformation needed to generate attenuation images, is substantially similar across scanner types (9).

The transmission measurements made with 511-keV photons provide a more accurate estimate of attenuation because, in contrast to CT photons, their probability of various modes of scatter interactions (ie, Compton, photoelectric, Rayleigh) is precisely the same as for the photons emitted during a PET scan. Figure 5A shows that regardless of X-ray energy, the CT-based attenuation values have a negative bias versus the transmission scan. Because the PET reconstruction uses these attenuation values to compensate for lost photons, we would expect PET images of the phantom to inherit this negative bias, as was observed. While the PET image bias does change with CT values, as shown in Table 1, X-ray energy cannot be varied arbitrarily and no user-selectable setting led to unbiased PET images.

As Figures 6B and 8 and Table 1 show, modifying the reconstruction to improve the accuracy of attenuation images leads to more accurate PET measurements and reduced bias. Figure 7B shows that recovery curves, which are used to char-

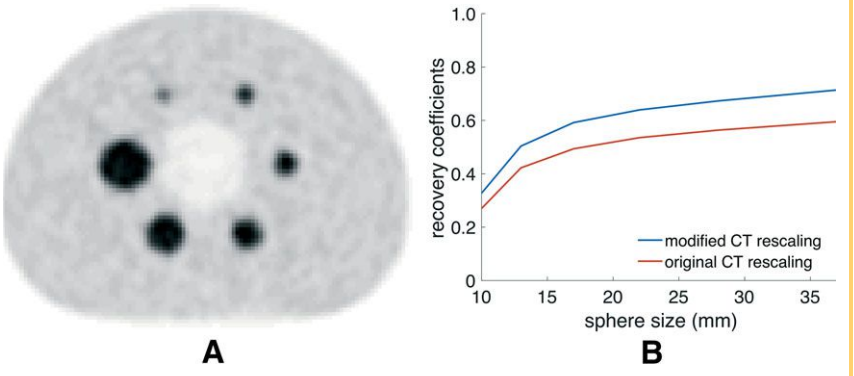


Figure 7. IQ phantom reconstructed with modified CT-based attenuation correction, showing uniformity in the background region (A). Mean region of interest divided by known concentration (signal recovery) for the depicted spheres (B).

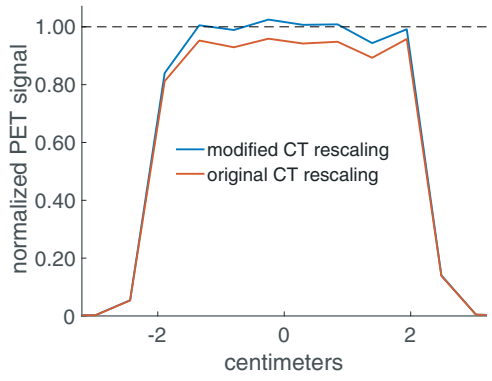


Figure 8. PET signal from the 45-mm-diameter X-Cal phantom divided by known phantom activity concentration. Here, the modified attenuation correction used slightly different scaling than was used for the IQ phantom reconstructions.

acterize resolution, also exhibit reduced bias when attenuation correction is accurate.

The precise modifications required to generate accurate attenuation images were object-dependent. For the IQ phantom and the X-Cal phantom, the correct coefficient was determined by measuring the Hounsfield Units of the epoxy in each scan and choosing the new slope of the rescaling formula that led to 0.105 1/cm. The IQ and X-Cal phantoms had Hounsfield Units of 92.1 and 82.2, respectively, although they were made of the same material. We did not attempt to correct the CT transformation for multiple X-ray tube voltages, but we note that because CT values themselves depend on voltage, the optimal modifications for one voltage will not work for others. We further expect that they would change if the experiment were repeated with a different epoxy. In all, this indicates that correcting bias in epoxy with this method would require premeasured look-up tables so that the appropriate rescaling factors are available for a range of scan scenarios.

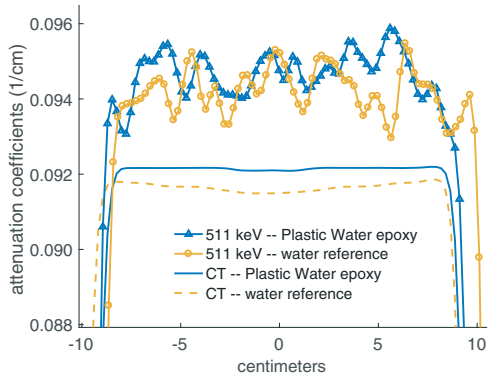


Figure 9. Profiles through 511-keV transmission scans and standard CT-based attenuation images for the PlasticWater LR phantom and a similarly sized water phantom.

Table 2. Attenuation Values (1/cm) Measured by Large Regions of Interest in the PlasticWater LR phantom and Similarly Sized Aqueous Phantom from CT Scans (First 4 Rows) and Transmission Measurements (Bottom Row)

	LR	Water
80 kVp	0.0922	0.0917
100 kVp	0.0922	0.0916
120 kVp	0.0921	0.0916
140 kVp	0.0922	0.0916
511 keV trans	0.0949	0.0944

An alternative approach would be the use of an epoxy whose attenuation properties are better suited to the transformation used by the scanner. The PlasticWater phantom was investigated with this in mind. As Table 2 shows, the attenuation estimates from a GE Discovery LS do not show a dependence on X-ray tube voltage, and the bias of attenuation estimates versus the transmission scan is reduced. Further, the bias between CT-based and 511-keV transmission measurements closely resembles that of an actual water phantom (Figure 9), for which the scanner's reconstruction algorithm is presumably well-calibrated. That is, the slight bias of CT-based water attenuation coefficients may be intentionally introduced to compensate for other approximations in the algorithm, such as imperfect scatter correction. It is therefore plausible, although not confirmed here, that a radioactive PET phantom constructed from the PlasticWater epoxy would exhibit bias similar to a water-filled phantom.

We emphasize that bias in attenuation estimates of our solid phantoms made from CT images (Figure 5A) does not imply that clinical patient images are similarly affected. Rather, it is a property of an empirical optimization in the reconstruction that favors clinical patient images over other materials.

While our study was limited in scope, using a small number of scanners, we expect that the bias seen in our PET images could be replicated on most scanners using CT-based attenuation correction, owing to the similarity in the way their X-rays are generated and detected.

Future work should better characterize the robustness and trade-offs of applying software modifications versus using new materials for phantom construction. In particular, the fabrication of long-lived radioactive phantoms can present unique manufacturing challenges, and the authors make no claims about the fitness of the specific materials used in the present study for this purpose. Software modifications would require vendor participation, but have the advantage of being compatible, in principle, with any existing phantom whose attenuation properties are known. Standardization of phantom size and composition may lead to object-dependence being a smaller hurdle for software-based bias reduction.

CONCLUSIONS

Solid, long-lived PET phantoms can suffer signal bias owing to physical factors. We have shown that corrections for photon attenuation computed from CT images can be a significant source of bias. Modifications to the reconstruction algorithm can reduce the errors in CT-based attenuation estimates, although the required parameters are likely to depend on X-ray tube voltage, the type of epoxy used, and the geometry of the phantom. The use of epoxy that better matches the photon-scattering properties of water appears to be a promising alternative to algorithmic corrections if they can be manufactured reliably, which is a nontrivial task.

ACKNOWLEDGMENTS

This work was supported by NIH-SAIC Contract 24XS036-004, NIH Grant R01CA169072, and NIH Grant U01 CA148131. The authors would like to thank Adam Strohn, Levent Sensoy, Joshua Scheuermann, and Joel Karp for their help in coordinating phantom transfers and acquiring data, and Steve Kohlmeyer and Keith Allberg for their assistance with phantom construction and testing.

Disclosure: Dr. Kinahan reports grants from NIH, during the conduct of the study; other from PET/X LLC, outside the submitted work; and Research Grant from GE Healthcare.

REFERENCES

1. Avril N, Sassen S, Schmalfeldt B, Naehrig J, Rutke S, Weber WA, Werner M, Graeff H, Schwaiger M, Kuhn W. Prediction of response to neoadjuvant chemotherapy by sequential F-18-fluorodeoxyglucose positron emission tomography in patients with advanced-stage ovarian cancer. J Clin Oncol. 2005;23:7445–7453.
2. Weber WA. Assessing tumor response to therapy. J Nucl Med. 2009;50 (*Suppl* 1):1S–10S.
3. Fletcher JW, Djulbegovic B, Soares HP, Siegel BA, Lowe VJ, Lyman GH, Coleman RE, Wahl R, Paschold JC, Avril N, Einhorn LH, Suh WW, Samson D, Delbeke D, Gorman M, Shields AF. Recommendations on the use of 18F-FDG PET in oncology. J Nucl Med. 2008;49:480–508.
4. Shankar LK, Hoffman JM, Bacharach S, Graham MM, Karp J, Lammertsma AA, Larson S, Mankoff DA, Siegel BA, Van DA, Annick VdA, Yap J, Sullivan D. Consensus recommendations for the use of 18F-FDG PET as an indicator of therapeutic response in patients in national cancer institute trials. J Nucl Med. 2006;47: 1059–1066.
5. Kinahan PE, Fletcher JW. PET/CT standardized uptake values (SUVs) in clinical practice and assessing response to therapy. Semin Ultrasound CT MR. 2010;31: 496–505.
6. Boellaard R. Standards for PET image acquisition and quantitative data analysis. J Nucl Med. 2009;50:11S–20S.
7. Kumar V, Nath K, Berman CG, Kim J, Tanvetyanon T, Chiappori AA, Gatenby RA, Gillies RJ, Eikman EA. Variance of standardized uptake values for FDG-PET/CT greater in clinical practice than under ideal study settings. Clin Nucl Med. 2013;38:175–182.
8. Kurland BF, Doot RK, Linden HM, Mankoff DA, Kinahan PE. Multicenter trials using 18F-fluorodeoxyglucose (FDG) PET to predict chemotherapy response: effects of differential measurement error and bias on power calculations for unselected and enrichment designs. Clin Trials. 2013;10:886–895.
9. Carney JPJ, Townsend DW, Rappoport V, Bendriem B. Method for transforming CT images for attenuation correction in PET/CT imaging. Med phys. 2006;33: 976–983.
10. Kinahan PE, Townsend DW, Beyer T, Sashin D. Attenuation correction for a combined 3D PET/CT scanner. Med Phys. 1998;25:2046–2053.
11. Abella M, Alessio AM, Mankoff DA, MacDonald LR, Vaquero JJ, Desco M, Kinahan PE. Accuracy of CT-based attenuation correction in PET/CT bone imaging. Phys Med Biol. 2012;57:2477–2490.
12. Hubbell JH, Seltzer M. Tables of X-Ray Mass Attenuation Coefficients and Mass Energy-Absorption Coefficients (version 1.4). National Institute of Standards and Technology: Gaithersburg, MD.
13. De Man B, Basu S, Chandra N, Dunham B, Edic P, Iatrou M, McOlash S, Sainath P, Shaughnessy C, Tower B. CatSim: a new computer assisted tomography simulation environment. *Proceedings Volume 6510, Medical Imaging 2007: Physics of Medical Imaging , 65102G*; 2007.
14. NCFBI. PubChem Compound Database; CID=169944.
15. Daube-Witherspoon ME, Karp JS, Casey ME, DiFilippo FP, Hines H, Muehllehner G, Simcic V, Stearns CW, Adam L-E, Kohlmyer S. PET performance measurements using the NEMA NU 2-2001 standard. J Nucl Med. 2002;43:1398–1409.
16. Scheuermann JS, Reddin JS, Opanowski A, Kinahan PE, Siegel BA, Shankar LK, Karp JS. Qualification of national cancer institute-designated Cancer Centers for Quantitative PET/CT Imaging in Clinical Trials. J Nucl Med. 2017;58:1065–1071.
17. Boellaard R, Delgado-Bolton R, Oyen WJG, Giammarile F, Tatsch K, Eschner W, Verzijlbergen FJ, Barrington SF, Pike LC, Weber WA. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. Eur J Nucl Med Mol Imaging. 2015;42:328–354.
18. Doot RK, Scheuermann JS, Christian PE, Karp JS, Kinahan PE. Instrumentation factors affecting variance and bias of quantifying tracer uptake with PET,âïCT. Med Phys. 2010;37:6035–6046.
19. Byrd D, Christopfel R, Arabasz G, Catana C, Karp J, Lodge MA, Laymon C, Moros EG, Budzevich M, Nehmeh S, Scheuermann J, Sunderland J, Zhang J, Kinahan P. Measuring temporal stability of positron emission tomography standardized uptake value bias using long-lived sources in a multicenter network. J Med Imaging. 2018;5:011016.
20. Byrd DW, Doot RK, Allberg KC, MacDonald LR, McDougald WA, Elston BF, Linden HM, Kinahan PE. Evaluation of cross-calibrated (68)Ge/(68)Ga phantoms for assessing PET/CT measurement bias in oncology imaging for single- and multicenter trials. Tomography. 2016;2:353–360.

FLT PET Radiomics for Response Prediction to Chemoradiation Therapy in Head and Neck Squamous Cell Cancer

Ethan J. Ulrich[1,2], Yusuf Menda[3], Laura L. Boles Ponto[3], Carryn M. Anderson[4], Brian J. Smith[5], John J. Sunderland[3], Michael M. Graham[3], John M. Buatti[4], and Reinhard R. Beichel[1,6]

Departments of [1]Electrical and Computer Engineering, [2]Biomedical Engineering, [3]Radiology, [4]Radiation Oncology, [5]Biostatistics, and [6]Internal Medicine, University of Iowa, Iowa City, IA

Corresponding Author:
Reinhard R. Beichel, PhD
Department of Electrical and Computer Engineering, University of Iowa, 4016 Seamans Center for the Engineering Arts and Sciences, Iowa City, IA 52242;
E-mail: reinhard-beichel@uiowa.edu

Key Words: PET, FLT, radiomics, prediction, head and neck cancer

Abbreviations: ^{18}F-fluorothymidine (FLT), positron emission tomography (PET), chemoradiation therapy (CRT), radiation therapy (RT), head and neck squamous cell cancer (HNSCC), distant metastasis (DM), local recurrence (LR), volumes of interest (VOIs), gray-level size zone matrix (GLSZM), standardized uptake value (SUV), metabolic tumor volume (MTV), total lesion glycolysis (TLG)

ABSTRACT

Radiomics is an image analysis approach for extracting large amounts of quantitative information from medical images using a variety of computational methods. Our goal was to evaluate the utility of radiomic feature analysis from ^{18}F-fluorothymidine positron emission tomography (FLT PET) obtained at baseline in prediction of treatment response in patients with head and neck cancer. Thirty patients with advanced-stage oropharyngeal or laryngeal cancer, treated with definitive chemoradiation therapy, underwent FLT PET imaging before treatment. In total, 377 radiomic features of FLT uptake and feature variants were extracted from volumes of interest; these features variants were defined by either the primary tumor or the total lesion burden, which consisted of the primary tumor and all FLT-avid nodes. Feature variants included normalized measurements of uptake, which were calculated by dividing lesion uptake values by the mean uptake value in the bone marrow. Feature reduction was performed using clustering to remove redundancy, leaving 172 representative features. Effects of these features on progression-free survival were modeled with Cox regression and *P*-values corrected for multiple comparisons. In total, 9 features were considered significant. Our results suggest that smaller, more homogenous lesions at baseline were associated with better prognosis. In addition, features extracted from total lesion burden had a higher concordance index than primary tumor features for 8 of the 9 significant features. Furthermore, total lesion burden features showed lower interobserver variability.

INTRODUCTION

Concomitant chemoradiation is used as an organ-sparing treatment strategy for advanced oropharyngeal and larynx cancers. Although outcomes vary based on stage, site, and other factors including human papilloma virus status, the 3-year progression-free survival of patients with advanced-stage head and neck cancer after chemoradiation therapy (CRT) is ~60% (1). Patients in whom cancer recurs after initial CRT are considered for salvage surgery; but, patients with presalvage Stage IV disease and those with presalvage Stage III disease at recurrence have poor prognosis with a median survival of <6 month and 14 months, respectively (2). As both new targeted therapies and radiation therapy (RT) delivery methods are developed, there is a need to develop biomarkers that may help stratify patients *a priori* for different treatment modalities or that can predict the likelihood of durable response versus ultimate failure earlier during therapy to allow for adaptive treatment approaches.

Positron emission tomography (PET) with ^{18}F-fluorodeoxyglucose (FDG PET) is widely used in pretreatment staging and post-therapy evaluation of head and neck cancers after RT or CRT. Because of its high negative predictive value in detection of recurrent disease, the National Comprehensive Cancer Network Guidelines now recommend omitting consolidative surgery (neck dissection) if the post-therapy FDG PET obtained at least 12 weeks after initial therapy is negative for residual tumor (3). However, the role of FDG PET in predicting failure of CRT or monitoring treatment response to (chemo)radiation during or early after treatment is not well established (12 weeks after initial therapy is typically required).

Radiation therapy and chemotherapy affect proliferation rates in treated tumors. In addition, pretreatment proliferation

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2018.00038

rates may be a determinant of sensitivity to chemotherapy and RT. Assessment of cancer proliferation rates and changes in cell proliferation rate may therefore accurately predict ultimate therapeutic response. 3′-deoxy-3′-^{18}F-fluorothymidine (FLT), a thymidine analogue that is not incorporated into DNA, is the most widely studied PET agent for imaging cell proliferation. The intracellular trapping of FLT is regulated by thymidine kinase 1, a key enzyme in DNA synthesis, with high activity during the proliferative phase of the cell cycle and low activity in the quiescent phase (4). Several studies have shown that untreated head and neck cancers can be imaged with FLT PET with a high tumor-to-background contrast (5–9).

Radiomics is an image analysis approach with the goal of extracting large amounts of quantitative information from medical images using a variety of computational methods. Extracted features include measurements of intensity (uptake), shape, and texture. The objective of this study was to evaluate the utility of FLT PET radiomic features obtained at baseline in the prediction of treatment response in patients with head and neck squamous cell cancer (HNSCC). The present work provides a basis for further optimization of predictive FLT PET features, which can then be further evaluated in future clinical trials.

METHODOLOGY

Patients

A single-center prospective study was performed in patients who had histologically confirmed HNSCC and were scheduled to receive definitive concurrent CRT per standard cancer care. Other eligibility criteria included a Karnofsky score of ≥60, acceptable bone marrow reserve (absolute neutrophil count, ≥1.5 K/mL; platelet count, ≥100 K/mL) and kidney (serum creatinine, ≤2.1 mg/dL), and liver function (bilirubin, ≤1.0 mg/dL; ALT/AST, ≤2.5 times upper limits of normal for the institution). These criteria generally excluded patients who were not robust enough to receive combined modality therapy. Patients were excluded if they had chemotherapy or radiotherapy within 4 weeks before the study (no induction chemotherapy) or were receiving investigational drugs or nucleoside analogues (such as 5-Fluorouracil that could interfere with FLT uptake). All patients were scheduled to undergo a baseline FLT PET scan within 30 days of the initiation of CRT. This was generally done the week before starting treatment. Platinum-based chemotherapy was started the first day of radiotherapy, either with high-dose cisplatinum or a combination of cisplatinum or carboplatinum combined with a taxane. Patients were followed every 3 months with clinical exams for the first year per our clinical routine and 2–4 times per year subsequently. Surveillance FDG PET scans were obtained at 3–4 months after treatment. Subsequent follow-up imaging was individualized on the basis of symptoms and clinical findings. This research was approved by the University of Iowa Institutional Review Board, and all subjects signed an informed consent. The research was conducted according to the principles of the Declaration of Helsinki and Good Clinical Practice.

In total, 30 patients with squamous cell head and neck cancer, including 27 oropharyngeal cancers, 1 unknown primary, and 2 laryngeal cancers, were available for analysis. There were 26 male and 4 female patients with an age range of 36–76 years (median, 57 years). The demographics of the patients including distribution of tumor stages are summarized in Table 1. After a median follow-up of 26 months (range, 7–36 months), 8 patients died of disease, 1 patient was alive with distant metastasis (DM), and 21 patients had no evidence of disease. Among the 8 patients who died from the disease, 4 patients had local recurrence (LR), 1 patient had local recurrence and distant metastasis (LR + DM), and 3 patients had DM alone at the time of initial recurrence or progression. Three patients underwent salvage surgery after completion of radiotherapy because of local recurrence and had no evidence of disease at last follow-up. The median follow-up in patients with no evidence of disease was 25 months.

Table 1. Overview of Patients in the FLT PET Study (n = 30)

Patient Characteristics	Categories	Total [%]	Median [Range]
Age at diagnosis (years)			57 [36–76]
Sex	Male	26 [86.7]	
	Female	4 [13.3]	
Site	Oropharynx	27 [90.0]	
	Larynx	2 [6.7]	
	Unknown primary	1 [3.3]	
T-Stage	Tx	1 [3.3]	
	T1	1 [3.3]	
	T2	15 [50.0]	
	T3	7 [23.3]	
	T4	6 [20.0]	
N-Stage	N0	5 [16.7]	
	N1	5 [16.7]	
	N2	16 [53.3]	
	N3	4 [13.3]	
Overall Stage	II	2 [6.7]	
	III	9 [30.0]	
	IVA	13 [43.3]	
	IVB	6 [20.0]	
Follow-Up (Months)			22.0 [4.6–36.0]
Survival Status	Progression-free survival	21 [70]	
	Progression or death	9[a][30]	

[a] Consists of 4 patients with LR, 4 patients with DM, and 1 patient with LR + DM.

FLT PET Imaging

For the synthesis of FLT, fluorine-18 fluoride was reacted with 3′-anhydrothymidine-5′-benzoate following the procedure of

Machulla et al. (10). The benzoate protecting group was removed with base hydrolysis and the product purified by semiprep HPLC with 10% ethanol/90% isotonic saline as the mobile phase with typical yields of 5%–8%. FLT was infused via a syringe pump over 2 minutes followed by 10-mL saline flush administered manually. The administered activity of FLT was 2.6 MBq/kg (0.07 mCi/kg) with a maximum dose of 185 MBq (5 mCi). Imaging was performed on a Siemens ECAT EXACT HR + PET scanner (Siemens Medical Solutions USA, Inc., Knoxville, TN) for 40 minutes, starting 60 minutes after injection. Transmission imaging was performed before the injection of FLT. Whole-body scans were obtained for 28 patients, and scans of the head and neck region were obtained for only 2 patients. Images were iteratively reconstructed (2 iterations = 8 subsets, Gaussian 8.0 mm, zoom = 1.2) with a resulting voxel size of 4.29 × 4.29 × 4.29 mm.

Image Analysis

For primary tumors and FLT-avid lymph nodes, volumes of interest (VOIs) defined by high FLT uptake above background were generated by a nuclear medicine physician using a semiautomated segmentation software developed for head and neck tumors in PET (11). Primary tumors were segmented on FLT PET in all patients except for 1 patient who had an unknown primary tumor site. FLT-avid nodal metastases in the neck were identified in 23 patients. In total, 83 lesions/VOIs were identified using the semiautomated PET segmentation tool. Each VOI received an individual label. Subsequently, these labels were used to define 2 different measurement region categories (ie, VOIs) from which radiomic features were extracted. The first measurement region category PT consisted of VOIs representing primary tumor only. The second category LB was the total lesion burden, which corresponds to the primary tumor and all FLT-avid nodes combined. To calculate quantitative features for LB, all lesion previously segmented in a FLT scan were combined into 1 image mask, forming a single VOI. For each measurement region, radiomic features describing intensity, shape, and texture properties were calculated by using the open-source packages PET-IndiC (12) and pyradiomics (13). All features were derived from standardized uptake value (SUV) normalized PET images. A total of 104 quantitative baseline PT features and an additional 99 baseline LB features were extracted from each patient. Note that 5 shape features (ie, slice maximum 2D diameter, column maximum 2D diameter, row maximum 2D diameter, maximum 3D diameter, and sphericity) are meant for single, connected VOIs, so these were excluded from the LB features.

For texture features, the histogram bin size was fixed at 0.25 SUV. The selected bin size follows van Velden et al. (14), where the total number of bins will be ~64 bins, depending on the lesion SUV range. A fixed bin size is used rather than a fixed number of bins because lesion SUV ranges vary among patients and fixing the number of bins is less appropriate for the clinical setting (15).

In addition to SUV-based measurements, normalized measurements of uptake were calculated by dividing lesion SUVs by the mean SUV in the bone marrow. The goal of normalization is to compare the cell proliferation in cancerous tissue to that of a normal structure. Normalization of SUVs was accomplished by generating a VOI around the largest vertebra completely visible in the field of view using the same segmentation software described above. In total, 30 vertebral VOIs were created using the semiautomated segmentation tool. For most patients, the L5 vertebra was segmented. The L4 vertebra was segmented for 1 patient owing to the L5 vertebra not being completely within the field of view. Because 2 patients had PET scans that did not include lumbar vertebrae, the T4 or T6 vertebra were segmented instead. Patient SUVs were then normalized by dividing by the mean vertebral SUV, and radiomic features were again calculated from the lesion VOIs. In total, 87 vertebra-normalized PT features and 87 vertebra-normalized LB features were generated from each patient. Note that normalization is not applicable for 11 features (ie, shape features) and has no effect on Q1–Q4 distributions, skewness, and kurtosis. For texture features based on normalized uptake, the histogram bin size was fixed at 0.125 (unitless). Note that the bin size is reduced compared with unnormalized texture features (bin size, 0.25), because normalization reduces the lesion intensity ranges compared with unnormalized lesions. In total, 377 baseline radiomic features were extracted from each patient.

Feature Reduction

Redundancy of quantitative features was reduced by using a clustering algorithm. The goal of feature reduction was to replace highly correlated features with a single representative feature. Such a step could be achieved by utilizing a PCA-based feature selection step [eg, FactoMineR (16)]. However, due to the sparseness of our feature space, a more appropriate feature selection method was utilized. First, the similarities of features were calculated by determining the Pearson correlation (r) for all pairs of features. Next, features were clustered according to similarity using an affinity propagation (AP) clustering algorithm (17), an unsupervised dimension reduction technique that others have utilized in the analysis of quantitative imaging features (18–20). An advantage of AP clustering over k-means clustering is that the total number of clusters at the output is automatically determined. Moreover, the algorithm is able to handle infinite dissimilarities, meaning 2 features that are highly dissimilar will not be placed in the same cluster. Therefore, to allow features with only strong correlations defined by $r \geq 0.90$ to be clustered together, all features with pairwise similarity values less than 0.90 were artificially set to have infinite dissimilarity before application of the AP clustering algorithm. As output, the algorithm produces a reduced set of representative exemplar features. An exemplar feature can be either a single feature with no strong correlations with other features or a representative of a cluster containing ≥2 features. The feature reduction step was performed using the *apcluster* package (21) in version 3.2.3 of the R statistical software (22).

Statistical Evaluation

Survival analysis was conducted to estimate and test the effects of quantitative features in the reduced set on progression-free survival (PFS). Time to event for PFS was defined as time from start of treatment to recurrence or death. Effects on survival during the 36-mo, post-treatment period were of primary interest. Hence, subjects who did not experience an event by month

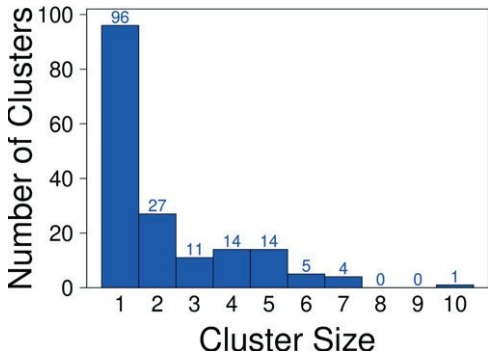


Figure 1. Cluster size distribution for the 172 clusters identified in the feature reduction step.

36 were censored at that point in time for the analysis. Cox regression was used to model the effects of individual quantitative features on survival. Using multiple predictors in a Cox regression model on a small cohort has the potential to overfit the patient data. Therefore, a Cox model with a single predictor was chosen to avoid overfitting. Estimated effects are summarized with hazard ratios (HRs) and the concordance (*c*)-index. The *c*-index is an estimate of the probability that, out of 2 randomly selected patients, the model can discriminate which patient will survive longer (23). Values can range from 0.0 to 1.0, with 0.5 indicating absence of discriminant value for the model, 0.7 indicating reasonable discriminant value, and 1.0 indicating perfect discriminant value. Two-sided *P*-values for tests of significance of features in the models are reported. To account for multiple statistical tests, the false-discovery rate (FDR) was computed using the Benjamini–Hochberg method (24). Features with a FDR of 10% were identified as significant. All statistical tests were performed using the *survival* package (25) for R.

Interobserver Variability Analysis

To study the variabiltity of feature measurement, a second observer independently generated segmentations (VOIs) for the same 83 lesions and features were calculated as described above. The features extracted from the second observer's VOIs were then compared to the features extracted from the VOIs of the first observer. Agreement in feature measurement was compared using the intraclass correlation coefficient (ICC). To investigate the impact of interobserver segmentations on model performance, a separate model for each predictive feature was generated using segmentations by the second observer. The performance of these models were then compared to the initial models from the first observer. Differences of model performance were reported as changes in *c*-index values.

RESULTS

Feature Reduction

The feature reduction step took the 377 FLT features as input and clustered similar features together to produce 172 uncorrelated clusters. Figure 1 shows the distribution of cluster sizes. Ninety-six clusters had a size of one, meaning there were 96 features (25.5%) that were not highly correlated with any other feature ($r < 0.9$). The remaining 76 clusters had size of ≥ 2, with the maximum being a size of 10.

Correlation of Baseline Features With Treatment Outcome

Feature performance was estimated using each feature as a predictor in a univariate Cox regression model. A total of 37

Table 2. Comparison of Predictive FLT Features (Progression-Free Survival) With 3 Commonly Used Features, SUV_{max}, SUV_{peak}, and SUV_{mean}

Feature (VOI, normalization)	*P*-Value	HR [95% CI]	FDR	*c*-Index
Gray-Level Non-Uniformity[a] (LB, N)	0.0002	3.11 [1.70, 5.68]	0.043	0.86
Gray-Level Non-Uniformity[b] (LB, N)	0.0012	3.12 [1.56, 6.24]	0.058	0.72
Spherical Disproportion (LB, U)	0.0012	4.10 [1.56, 10.80]	0.058	0.74
Information Measure of Correlation 2[c] (LB, U)	0.0017	0.32 [0.16, 0.65]	0.058	0.79
Zone Percentage[b] (LB, N)	0.0020	0.18 [0.04, 0.78]	0.058	0.75
Gray-Level Non-Uniformity[a] (LB, U)	0.0020	2.21 [1.40, 3.47]	0.058	0.83
Q1 Distribution (LB, U)	0.0042	0.36 [0.17, 0.75]	0.088	0.78
Volume (LB, U)	0.0043	2.44 [1.38, 4.32]	0.088	0.74
Information Measure of Correlation 1[c] (LB, U)	0.0046	4.07 [1.23, 13.42]	0.088	0.78
SUV_{max} (LB, U)	0.1916	0.60 [0.27, 1.33]	0.395	0.66
SUV_{peak}[d] (LB, U)	0.3341	0.69 [0.32, 1.48]	—	0.63
SUV_{mean}[d] (LB, U)	0.5038	0.76 [0.34, 1.71]	—	0.62

Abbreviations: VOI, volume of interest; HR, hazard ratio; CI, confidence interval; FDR, false-discovery rate; PT, primary tumor; LB, lesion burden; U, unnormalized; N, normalized.
[a] Calculated from the gray-level run length matrix (GLRLM).
[b] Calculated from the gray-level size zone matrix (GLSZM).
[c] Calculated from the gray-level co-occurrence matrix (GLCM).
[d] Not selected in feature reduction step, so FDR was not calculated.

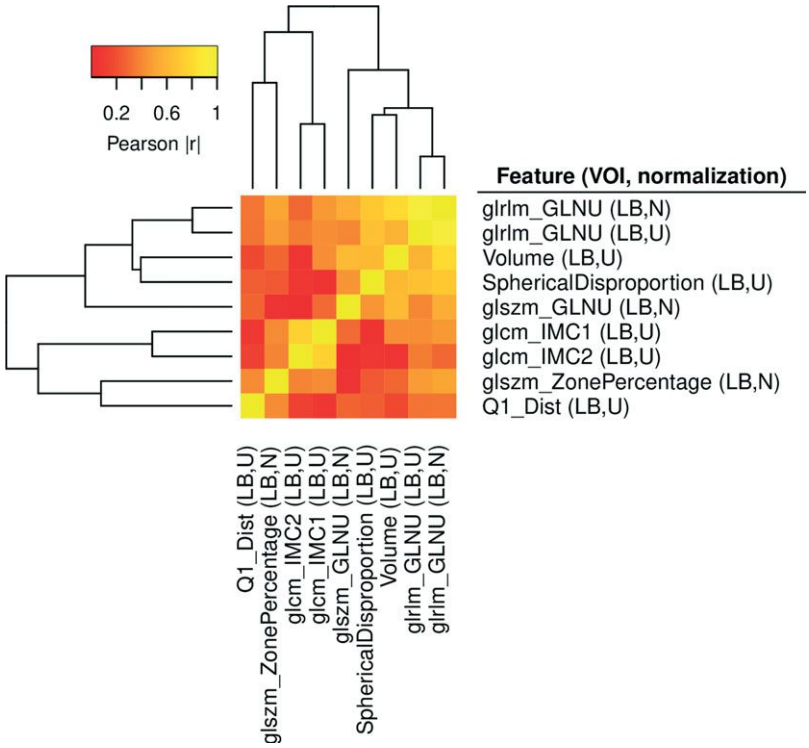


Figure 2. Heatmap of correlations among the 9 baseline ^{18}F-fluorothymidine (FLT) features with the best performance.

exemplar baseline features (21.5%) had *P*-values below the 5% level. After adjusting for multiple testing to control the false-discovery rate, a total of 9 baseline features were identified as significant at the set 10% FDR level. Table 2 summarizes the unadjusted Cox regression *P*-values, estimated hazard ratios with corresponding confidence intervals, FDRs, and *c*-index values for the 9 significant features as well as for 3 commonly used features (ie, SUV_{max}, SUV_{peak}, and SUV_{mean}). SUV_{peak} was defined as the highest average uptake within a 1 cm^3 sphere that is completely contained within the VOI. Note that SUV_{peak} and SUV_{mean} were not selected in the feature reduction step, so univariate analyses were done separately and no FDRs were calculated for SUV_{peak} and SUV_{mean}. The clinical parameter for primary tumor stage (T-stage) was not significantly associated with survival.

Figure 2 shows a heatmap of correlations among the 9 significant features. The feature reduction step used a high correlation threshold ($r \geq 0.90$), so moderate correlations among the best-performing features still exist. By showing the correlations of the features in a heatmap, good-performing lesion characteristics, rather than individual features, may be observed. For example, features that measure lesion size (eg, volume) and shape (eg, spherical disproportion) had good performance. Also, measures of lesion heterogeneity (eg, gray-level nonuniformity and zone percentage) had good performance.

Interobserver Variability Analysis

Table 3 shows the results of the variability analysis for the 9 significant features and the commonly used features SUV_{max}, SUV_{peak}, and SUV_{mean}. Gray-level nonuniformity from the gray-level size zone matrix (GLSZM) had moderate agreement between the 2 observers. The other 8 significant features had strong agreement between the 2 observers. Both SUV_{max} and SUV_{peak} had perfect agreement between the 2 observers and SUV_{mean} had strong agreement.

To assess model performance stability, the segmentations of the second observer were used to produce a second model for

Table 3. Interobserver Agreement for Predictive FLT Features and 3 Commonly Used Features, SUV_{max}, SUV_{peak}, and SUV_{mean}

Feature (VOI, normalization)	Measurement Agreement
Gray-level Non-Uniformity[a] (LB, N)	0.99
Gray-level Non-Uniformity[b] (LB, N)	0.75
Spherical Disproportion (LB, U)	0.96
Information Measure of Correlation 2[c] (LB, U)	0.98
Zone Percentage[b] (LB, N)	0.91
Gray-level Non-Uniformity[a] (LB, U)	0.99
Q1 Distribution (LB, U)	0.90
Volume (LB, U)	0.99
Information Measure of Correlation 1[c] (LB, U)	0.95
SUV_{max} (LB, U)	1.00
SUV_{peak} (LB, U)	1.00
SUV_{mean} (LB, U)	0.94

Measurement agreement was calculated as the Intraclass Correlation Coefficient (ICC) between the feature values of the first and second observer.

Abbreviations: VOI, volume of interest; LB, lesion burden; U, unnormalized; N, normalized.

[a] Calculated from the gray-level run length matrix (GLRLM).

[b] Calculated from the gray-level size zone matrix (GLSZM).

[c] Calculated from the gray-level co-occurrence matrix (GLCM).

Table 4. Differences of Model Performance Due to Interobserver Segmentation Variability

Feature (VOI, normalization)	Δc-index
Gray-Level Non-Uniformity[a] (LB, N)	0.00
Gray-Level Non-Uniformity[b] (LB, N)	−0.01
Spherical Disproportion (LB, U)	−0.03
Information Measure of Correlation 2[c] (LB, U)	0.03
Zone Percentage[b] (LB, N)	0.01
Gray-Level Non-Uniformity[a] (LB, U)	0.01
Q1 Distribution (LB, U)	−0.07
Volume (LB, U)	−0.01
Information Measure of Correlation 1[c] (LB, U)	0.03

Change Calculations are the Difference (Δ) of the *c*-Indices Between the Model of the First Observer and the Model of the Second Observer.
Abbreviations: VOI, volume of interest; LB, lesion burden; U, unnormalized; N, normalized.
[a] Calculated from the gray-level run length matrix (GLRLM).
[b] Calculated from the gray-level size zone matrix (GLSZM).
[c] Calculated from the gray-level co-occurrence matrix (GLCM).

each of the predictive features shown in Table 2. Table 4 shows the performance differences of the second model in reference to the first model for the univariate predictors with the best performance. For most features, only small changes in performance (*c*-index) were observed, indicating that model performance was stable. Only 1 feature (Q1 distribution) had a change in *c*-index >5 percentage points.

DISCUSSION

Performance

In this work, we investigated associations of patient outcomes with radiomic features derived from FLT PET lesion segmentations. Radiomics generates many features that can be highly correlated from each subject, so a feature reduction step was included to remove redundancies from the feature space. Despite this reduction, a large number of features were not highly correlated and tested in the performance analysis, so controlling the false-discovery rate was used to reduce false positives. A total of 9 FLT features were considered significant.

Our results suggest that a favorable prognosis is associated with a small lesion size, a more sphere-like lesion shape, and homogeneous intensity. Figure 3 shows the baseline scans of 2 patients with different outcomes and different FLT-avid lesion shapes. The surviving patient (Figure 3A) has a small, sphere-like lesion. The patient later classified with progressive disease (Figure 3B) has large lesions with a large, irregular surface area. Our results also suggest that lesion texture/homogeneity of intensity may be an indicator of outcome. Figure 4 shows the baseline scans of 2 patients with different outcomes and different lesion textures. The surviving patient (Figure 4A) has lesions with smaller regions of more uniform texture. The patient later classified with progressive disease (Figure 4B) has large regions and an overall nonuniform texture.

The authors are not aware of any publications that normalize lesion uptakes with the mean vertebral uptake before analysis of response prediction for HNSCC with FLT PET. Three out of the 5 best-performing intensity-based features were normalized with the mean vertebral uptake. Table 5 compares the *c*-indices of intensity-based FLT features with and without normalization. Texture features from the gray-level co-occurrence matrix have poorer performance after normalization. Texture features from the gray-level run length matrix and the GLSZM have a small increase in performance after normalization. Due to our small cohort of patients, more analysis is needed on a larger patient population to determine if these differences are significant.

All features identified as having an association with patient outcome were calculated from the total lesion burden (Table 2). This suggests that important information about the disease is found not only in the primary tumor, but also in the FLT-avid lymph nodes. Table 6 compares the *c*-indices of the 9 best-performing FLT features calculated from the primary tumor and the total lesion burden. All but 1 feature (ie, information measure of correlation 1) had higher performance when calculated from the total lesion burden. Furthermore, the interoperator agreement (ICC) average and standard deviation of the 9 best-performing FLT features for primary tumor and the total lesion burden was 0.88 ± 0.13 and 0.94 ± 0.08, respectively. Thus, FLT PET features derived from total lesion burden show higher agreement, and 8 out of the 9 best-performing features had strong agreement between different observers (Table 3). As

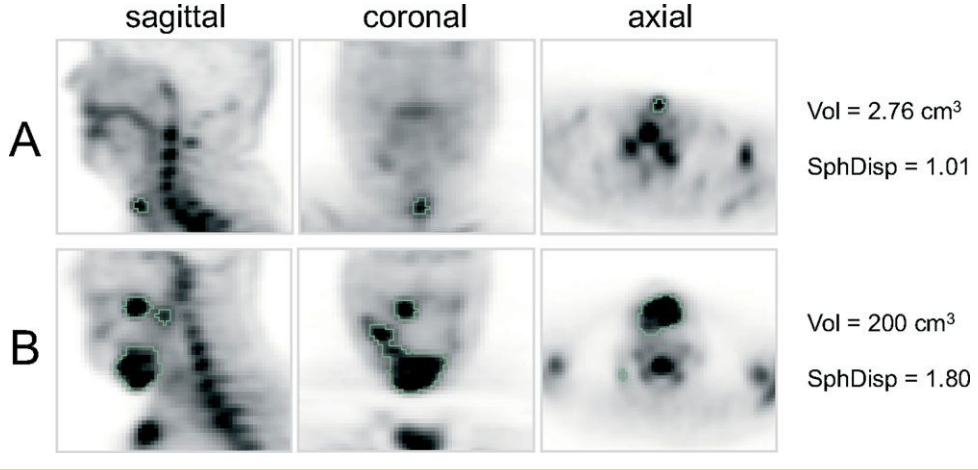


Figure 3. Baseline FLT scan slices showing differences in lesion size and shape. Patient later classified as progression-free survival at follow-up (A). Patient later classified as progression at follow-up (B). A favorable prognosis was associated with small tumor volume (Vol) and a lower spherical disproportion (SphDisp).

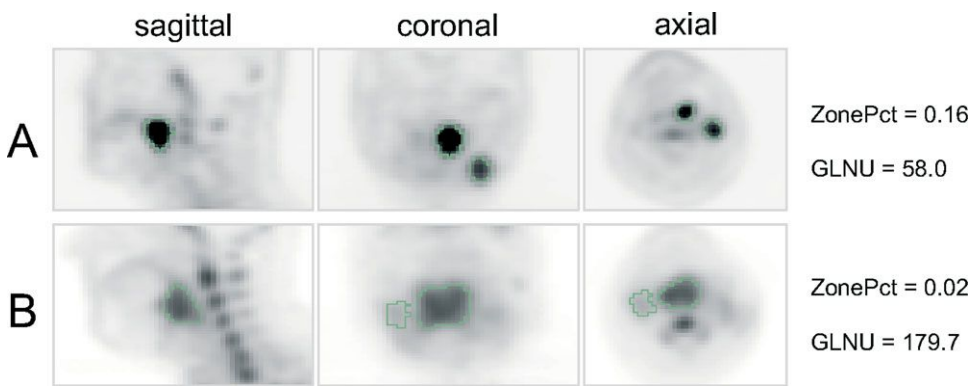


Figure 4. Baseline FLT scan slices showing differences in lesion texture. Patient later classified as progression-free survival at follow-up (A). Patient later classified as progression at follow-up (B). A favorable prognosis was associated with more homogeneous lesions and finer textures. Gray-level nonuniformity from the gray-level run length matrix (GLNU) has a lower value for more uniform regions. Zone percentage from the gray-level size zone matrix (ZonePct) has a higher value for regions with finer textures.

stated before, more analysis is needed on a larger patient population to determine if these differences are significant.

Related Work

The association of standard FLT features and outcome has been previously studied. For example, Hoshikawa et al. reported that baseline FLT tumor volume and total lesion proliferation (TLP) were predictive of locoregional tumor control in 32 patients with HNSCC treated with CRT and surgery (26). We found similar results to their findings for the total lesion burden volume ($P = .004$) and total lesion burden TLP ($P = .012$) for predicting 3-y progression-free survival. Note that total lesion burden TLP in our analysis was not selected during the feature reduction step. Hoshikawa et al. later reported that baseline FLT tumor volume, TLP, and SUV_{max} were predictive of locoregional tumor control in 53 patients with HNSCC treated with RT or CRT (27). Our results are not similar to their findings for unnormalized SUV_{max} ($P = .192$). This may be due to our smaller patient cohort (30 vs. 53). However, Linecker et al. reported earlier that high FLT uptake is associated with poor outcome in 20 patients treated with RT and CRT (8).

The authors are aware of 2 other publications that report correlations of FLT based radiomic features and patient outcomes. Willaime et al. reported that radiomic features were predictive of treatment response in 11 breast cancer patients treated with chemotherapy (28). However, the different cancer site and treatment type does not allow for a meaningful comparison with our results. Majdoub et al. (29) reported that tumor proliferative volume and textural features are predictive of disease-free survival in 45 patients with HNSCC treated with RT and CRT. They found that large, more heterogeneous lesions

Table 5. Comparison of c-Index Values for Unnormalized and Normalized Features

Feature	Unnormalized	Normalized
Gray-Level Non-Uniformity[a]	0.83	**0.86**
Gray-Level Non-Uniformity[b]	0.66	**0.72**
Information Measure of Correlation 2[c]	**0.79**	0.63
Zone Percentage[b]	0.73	**0.75**
Information Measure of Correlation 1[c]	**0.78**	0.56

Higher c-Index Values for Each Feature are Indicated in Bold.
[a] Calculated from the gray-level run length matrix (GLRLM).
[b] Calculated from the gray-level size zone matrix (GLSZM).
[c] Calculated from the gray-level co-occurrence matrix (GLCM).

Table 6. Comparison of c-Index Values for Features Calculated from the Primary Tumor and the Total Lesion Burden

Feature (Normalization)	Primary Tumor	Lesion Burden
Gray-Level Non-Uniformity[a] (N)	0.71	**0.86**
Gray-Level Non-Uniformity[b] (N)	0.50	**0.72**
Spherical Disproportion (U)	0.49	**0.74**
Information Measure of Correlation 2[c] (U)	0.75	**0.79**
Zone Percentage[b] (N)	0.68	**0.75**
Gray-Level Non-Uniformity[a] (U)	0.71	**0.83**
Q1 Distribution (U)	0.64	**0.78**
Volume (U)	0.59	**0.74**
Information Measure of Correlation 1[c] (U)	**0.79**	0.78

Higher c-Index Values for Each Feature are Indicated in Bold.
Abbreviations: U, unnormalized; N, normalized.
[a] Calculated from the gray-level run length matrix (GLRLM).
[b] Calculated from the gray-level size zone matrix (GLSZM).
[c] Calculated from the gray-level co-occurrence matrix (GLCM).

were associated with a less favorable prognosis, which is consistent with our findings.

In current clinical practice, FDG PET imaging is commonly utilized for assessment of response to treatment. For this purpose, simple quantitative image features like SUVmax, SUVpeak (30), metabolic tumor volume (MTV), or total lesion glycolysis (TLG) have been proposed, out of which SUVmax is most widely adopted. In a recent study, Castelli et al. (31) summarized the results of 45 studies regarding the predictive value of such FDG PET features with respect to clinical outcome in HNC treatment with chemoradiotherapy (CRT). The study concluded that MTV and TLG in pretreatment PET scans showed good correlation with disease free survival (DFS) or overall survival (OS). In this work, we have investigated FLT PET derived image features. At this stage, it is unclear which imaging approach (ie, tracer) results in better predictive performance. For example, the volume defined by above normal tracer uptake showed good performance on FLT data (Table 2) as well as in FDG PET studies (31). However, to decide which approach is preferable, a dedicated study is needed.

Limitations

This study has several limitations. The HPV (human papilloma virus) status, which is now a well-known prognostic factor in oropharyngeal cancers, was not available for this cohort as it was not routinely obtained when subjects were enrolled in this study. Furthermore, the effects of repeated scans and image reconstruction parameters on FLT-based radiomic features was not determined. Willaime et al. did investigate test–retest variability of texture features in breast cancer using FLT PET (28). They report similar results to a study by Tixier et al., which investigated the test-retest variability of FDG PET texture features using 16 patients with esophageal cancer (32). Both studies found that measures of tumor homogeneity and entropy had good repeatability. Leijenaar et al. investigated the repeatability of FDG PET texture features in non–small cell lung cancer (33). A majority of features (71%) were stable during test-retest analysis.

Yan et al. reported that zone percentage of the GLSZM was sensitive to image reconstruction parameters and should be used with caution (34). Their work used 20 patients with lung lesions imaged with FDG PET. Zone percentage was associated with patient outcome in our results, and it is a measure of fine textures. It is reasonable to expect that high variability of zone percentage calculations by different image reconstruction parameters would also occur in FLT PET. Reconstruction parameters were held constant for the images in our study.

CONCLUSION

In conclusion, radiomics is a useful approach for extracting large amounts of information from tumor images. We investigated the association of patient outcomes with radiomic features extracted from tumors imaged with FLT PET. Radiomics features performed favorably compared to standard clinical stage. We found that smaller, more homogenous lesions at baseline were associated with a better prognosis in 30 patients with head and neck cancer. Therefore, for future studies of FLT-based prediction of outcome, we recommend including radiomic features of lesion size, shape, and texture features that measure lesion homogeneity. We also recommend that radiomic features be calculated from the total lesion burden, rather than the primary tumor only, so that the largest amount of disease information is used for analysis. Our findings enable future optimization of FLT-based features which can then be assessed in validation studies.

ACKNOWLEDGMENTS

This research was funded in part by National Institutes of Health grants U01 CA140206, U24 CA180918, R21 CA130281, P30 CA086862, and U54 UL1TR002537. We thank Drs. G. Leonard Watkins and Kenneth Dornfeld for their contribution. We are indebted to Kellie Bodeker for providing regulatory oversight for our study.

Disclosures: No disclosures to report.

Conflict of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. Ang KK, Zhang Q, Rosenthal DI, Nguyen-Tan PF, Sherman EJ, Weber RS, Galvin JM, Bonner JA, Harris J, El-Naggar AK, Gillison ML, Jordan RC, Konski AA, Thorstad WL, Trotti A, Beitler JJ, Garden AS, Spanos WJ, Yom SS, Axelrod RS. Randomized phase III trial of concurrent accelerated radiation plus cisplatin with or without cetuximab for stage III to IV head and neck carcinoma: RTOG 0522. J Clin Oncol. 2014;32:2940–2950.
2. Goodwin JWJ. Salvage surgery for patients with recurrent squamous cell carcinoma of the upper aerodigestive tract: when do the ends justify the means? Laryngoscope. 2000;110(3 Pt 2 *Suppl* 93):1–18.
3. Sadowski SM, Neychev V, Millo C, Shih J, Nilubol N, Herscovitch P, Pacak K, Marx SJ, Kebebew E. Prospective study of 68Ga-DOTATATE positron emission tomography/computed tomography for detecting gastro-entero-pancreatic neuroendocrine tumors and unknown primary sites. J Clin Oncol. 2016;34:588–596.
4. Rasey JS, Grierson JR, Wiens LW, Kolb PD, Schwartz JL. Validation of FLT uptake as a measure of thymidine kinase-1 activity in A549 carcinoma cells. J Nucl Med. 2002;43:1210–1217.
5. Cobben DC, van der Laan BF, Maas B, Vaalburg W, Suurmeijer AJ, Hoekstra HJ, Jager PL, Elsinga PH. 18F-FLT PET for visualization of laryngeal cancer: comparison with 18F-FDG PET. J Nucl Med. 2004;45:226–231.
6. Hoshikawa H, Kishino T, Mori T, Nishiyama Y, Yamamoto Y, Inamoto R, Akiyama K, Mori N. Comparison of 18F-FLT PET and 18F-FDG PET for detection of cervical lymph node metastases in head and neck cancers. Acta Otolaryngol. 2012;132:1347–1354.
7. Hoshikawa H, Nishiyama Y, Kishino T, Yamamoto Y, Haba R, Mori N. Comparison of FLT-PET and FDG-PET for visualization of head and neck squamous cell cancers. Mol Imaging Biol. 2011;13:172–177.
8. Linecker A, Kermer C, Sulzbacher I, Angelberger P, Kletter K, Dudczak R, Ewers R, Becherer A. Uptake of 18F-FLT and 18F-FDG in primary head and neck cancer correlates with survival. Nuklearmedizin. 2008;47:80–85; quiz N12.
9. Menda Y, Boles Ponto LL, Dornfeld KJ, Tewson TJ, Watkins GL, Schultz MK, Sunderland JJ, Graham MM, Buatti JM. Kinetic analysis of 3′-deoxy-3′-18F-fluorothymidine (18F-FLT) in head and neck cancer patients before and early after initiation of chemoradiation therapy. J Nucl Med. 2009;50:1028–1035.
10. Machulla HJ, Blocher A, Kuntzsch M, Piert M, Wei R, Grierson JR. Simplified labeling approach for synthesizing 3′-deoxy-3′-[18F]Fluorothymidine ([18F]FLT. J Radioanal and Nucl Chem. 2000;243:843–846.
11. Beichel RR, van Tol M, Ulrich EJ, Bauer C, Chang T, Plichta KA, Smith BJ, Sunderland JJ, Graham MM, Sonka M, Buatti JM. Semiautomated segmentation of head

and neck cancers in 18F-FDG PET scans: a just-enough-interaction approach. Med Phys. 2016;43:2948–2964.
12. Ulrich EJ, van Tol M, Bauer C, Fedorov A, Beichel RR, Buatti JM. PET-IndiC extension documentation. In: 3D Slicer Wiki Internet. Available from: https://www.slicer.org/wiki/Documentation/Nightly/Extensions/PET-IndiC.
13. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan RGH, Fillion-Robin JC, Pieper S, Aerts HJWL. Computational radiomics system to decode the radiographic phenotype. Cancer Res. 2017;77:e104–e107.
14. van Velden FH, Kramer GM, Frings V, Nissen IA, Mulder ER, de Langen AJ, Hoekstra OS, Smit EF, Boellaard R. Repeatability of radiomic features in non-small-cell lung cancer 18F]FDG-PET/CT studies: impact of reconstruction and delineation. Mol Imaging Biol. 2016;18:788–795.
15. Leijenaar RT, Nalbantov G, Carvalho S, Van Elmpt WJ, Troost EG, Boellaard R, Aerts HJ, Gillies RJ, Lambin P. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. Sci Rep. 2015;5:11075.
16. Lê S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. J Stat Softw. 2008;25:1–18.
17. Frey BJ, Dueck D. Clustering by passing messages between data points. Science. 2007;315:972–976.
18. Bartholmai BJ, Raghunath S, Karwoski RA, Moua T, Rajagopalan S, Maldonado F, Decker PA, Robb RA. Quantitative CT imaging of interstitial lung diseases. J Thoracic Imaging. 2013;28.
19. Maldonado F, Boland JM, Raghunath S, Aubry MC, Bartholmai BJ, Hartman TE, Karwoski RA, Rajagopalan S, Sykes AM, Yang P. Noninvasive characterization of the histopathologic features of pulmonary nodules of the lung adenocarcinoma spectrum using computer-aided nodule assessment and risk yield (CANARY)—a pilot study. J Thoracic Oncol. 2013;8:452–460.
20. Zhu Y, Li H, Guo W, Drukker K, Lan L, Giger ML, Ji Y. Deciphering genomic underpinnings of quantitative MRI-based radiomic phenotypes of invasive breast carcinoma. Sci Rep. 2015;5:17787.
21. Bodenhofer U, Kothmeier A, Hochreiter S. APCluster: an R package for affinity propagation clustering. Bioinformatics. 2011;27:2463–2464.
22. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2015. Available from: https://www.R-project.org/.
23. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA. 1982;247:2543–2546.
24. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. 1995;57:289–300.
25. Therneau TM, Lumley T. Package 'survival' version 2.42-6 [Internet]; 2018 July 13. Available from: https://cran.r-project.org/package=survival.
26. Hoshikawa H, Yamamoto Y, Mori T, Kishino T, Fukumura T, Samukawa Y, Mori N, Nishiyama Y. Predictive value of SUV-based parameters derived from pretreatment 18F-FLT PET/CT for short-term outcome with head and neck cancers. Ann Nucl Med. 2014;28:1020–1026.
27. Hoshikawa H, Mori T, Yamamoto Y, Kishino T, Fukumura T, Samukawa Y, Mori N, Nishiyama Y. Prognostic value comparison between (18)F-FLT PET/CT and (18)F-FDG PET/CT volume-based metabolic parameters in patients with head and neck cancer. Clin Nucl Med. 2015;40:464–468.
28. Willaime JMY, Turkheimer FE, Kenny LM, Aboagye EO. Quantification of intratumour cell proliferation heterogeneity using imaging descriptors of 18F fluorothymidine-positron emission tomography. Phys Med Biol. 2012;58:187.
29. Majdoub M, Visvikis D, Texier F, Hoeben B, Visser E, Cheze-Le Rest C, Hatt M. Proliferative 18F-FLT PET tumor volumes characterization for prediction of locoregional recurrence and disease-free survival in head and neck cancer. In: *SNMMI 2013: Society of Nuclear Medicine and Molecular Imaging Annual Meeting*. Vancouver, Canada; 2013. Available from: https://hal.archives-ouvertes.fr/hal-00936229.
30. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: Evolving considerations for PET response criteria in solid tumors. J Nucl Med. 2009;50 *Suppl* 1:122S–50S.
31. Castelli J, De Bari B, Depeursinge A, Simon A, Devillers A, Roman Jimenez G, Prior J, Ozsahin M, de Crevoisier R, Bourhis J. Overview of the predictive value of quantitative 18 FDG PET in head and neck cancer treated with chemoradiotherapy. Crit Rev Oncol Hematol. 2016;108:40–51.
32. Tixier F, Hatt M, Le Rest CC, Le Pogam A, Corcos L, Visvikis D. Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET. J Nucl Med. 2012;53:693.
33. Leijenaar RT, Carvalho S, Velazquez ER, Van Elmpt WJ, Parmar C, Hoekstra OS, Hoekstra CJ, Boellaard R, Dekker ALAJ, Gillies RJ, Aerts HJWL, Lambin P. Stability of FDG-PET radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta Oncol. 2013;52:1391–1397.
34. Yan J, Chu-Shern JL, Loi HY, Khor LK, Sinha AK, Quek ST, Tham IW, Townsend D. Impact of image reconstruction settings on texture features in 18F-FDG PET. J Nucl Med. 2015;56:1667–1673.

TOMOGRAPHY® RESEARCH ARTICLE

ePAD: An Image Annotation and Analysis Platform for Quantitative Imaging

Daniel L. Rubin, Mete Ugur Akdogan, Cavit Altindag, and Emel Alkim

Department of Biomedical Data Science, Radiology, and Medicine (Biomedical Informatics Research), Stanford University, Stanford, CA

Corresponding Author:
Daniel Rubin, MD, MS
Medical School Office Building (MSOB) 1265 Welch Road,
X335 Stanford, CA 94305-5464;
E-mail: dlrubin@stanford.edu

Key Words: medical image annotation, biomarker evaluation, feature extraction, AIM (Annotation and Image Markup), DICOM SR (DICOM Structure Report)

Abbreviations: Electronic Physician Annotation Device (ePAD), Quantitative Imaging Network (QIN), positron emission tomography (PET), magnetic resonance imaging (MRI), regions of interest (ROIs), Picture Archiving and Communication System (PACS), Attenuation Distribution across the Long Axis (ADLA), Pixel Intensity Distributions (PID), gray-level co-occurrence matrices (GLCMs)

ABSTRACT

Medical imaging is critical for assessing the response of patients to new cancer therapies. Quantitative lesion assessment on images is time-consuming, and adopting new promising quantitative imaging biomarkers of response in clinical trials is challenging. The electronic Physician Annotation Device (ePAD) is a freely available web-based zero-footprint software application for viewing, annotation, and quantitative analysis of radiology images designed to meet the challenges of quantitative evaluation of cancer lesions. For imaging researchers, ePAD calculates a variety of quantitative imaging biomarkers that they can analyze and compare in ePAD to identify potential candidates as surrogate endpoints in clinical trials. For clinicians, ePAD provides clinical decision support tools for evaluating cancer response through reports summarizing changes in tumor burden based on different imaging biomarkers. As a workflow management and study oversight tool, ePAD lets clinical trial project administrators create worklists for users and oversee the progress of annotations created by research groups. To support interoperability of image annotations, ePAD writes all image annotations and results of quantitative imaging analyses in standardized file formats, and it supports migration of annotations from various propriety formats. ePAD also provides a plugin architecture supporting MATLAB server-side modules in addition to client-side plugins, permitting the community to extend the ePAD platform in various ways for new cancer use cases. We present an overview of ePAD as a platform for medical image annotation and quantitative analysis. We also discuss use cases and collaborations with different groups in the Quantitative Imaging Network and future directions.

INTRODUCTION

Advances in molecular medicine are providing many new treatments that promise to be safer and more effective than traditional cytotoxic treatments by targeting the molecular characteristics of each patient's tumor (1-3). As these new targeted treatments enter clinical trials, there is a growing need to derive quantitative characteristics from images of cancer lesions ("quantitative imaging biomarkers") that accurately assess the clinical benefit of these treatments (surrogate endpoints in clinical trials). Tumor shrinkage is the hallmark of response to traditional cytotoxic cancer therapies (4), and thus linear measurement of target lesions is the imaging biomarker used in most clinical trials using criteria such as Response Evaluation Criteria in Solid Tumors (RECIST) (5-7), Response Assessment in Neuro-Oncology (RANO) (8, 9), and International Harmonization Criteria (10). However, targeted, noncytotoxic therapies may arrest cancer growth and improve progression-free survival without necessarily shrinking tumors (11-14). Simple linear measurement may underestimate treatment response (15-18), in addition to having other limitations (7, 19). Alternative imaging biomarkers may be more promising than linear measurement for assessing response, especially with targeted therapeutic agents, as they can capture specific imaging features related to biological alterations in tumors during treatment (eg, heterogeneity, hypoxia, or changes in tumor microenvironment) (20-24), unlike tumor shrinkage (15, 25-27). Indeed, quantitative imaging biomarkers that reliably detect the results of anticancer agents (as opposed to detecting only change in tumor size) are desirable for all classes of therapeutic agents (28). Such new imaging biomarkers could become surrogate endpoints in clinical trials, as regulatory approval can be based on surrogate endpoints that document clinical benefit (29).

Development of imaging biomarkers follows a life cycle, starting with discovery and validation ("emerging biomarkers"), then translation and incorporation into clinical trials, and eventually to qualification for clinical use as surrogate endpoints for evaluating treatments ("qualified biomarkers") (30). A number of research groups are working on the discovery/validation of

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2018.00055

the spectrum and developing new quantitative imaging biomarkers, including the Quantitative Imaging Network (QIN) (31) and the broader community (32-39). On the translation end of the spectrum, many of the new imaging biomarkers are ready to be translated for use in clinical trials, such as tumor volume (40), changes in contrast enhancement on computed tomography (41), radiotracer uptake on positron emission tomography (PET) (32, 42-46), kinetic parameters in dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) (47-49), and spatial maps of such parameters (50, 51); however, very few of these new imaging biomarkers have yet to be incorporated into clinical trials for assessing treatment response.

Current image viewing and annotation tools are limited in their ability to support incorporating new imaging biomarkers into clinical trials in 4 major ways. First, although there are several commercial and open-source tools available to assess cancer lesions (52-55), they generally support very few measures of cancer lesions, such as linear dimension of target lesions, and they cannot be readily extended to deploy novel imaging biomarkers. Newer algorithms for computing imaging biomarkers are generally written in a variety of languages such as Java, Python, and C/C++, or exist within single toolkits [eg, MATLAB and 3D Slicer (56, 57)], which may not be compatible with current image assessment tools. Second, current lesion assessment tools are designed for only tracking cancer lesions in clinical practice, and they generally do not provide workflow management and study oversight features needed for assessing new image biomarkers in clinical trials. Third, there are no decision tools that use new imaging biomarkers for assessing treatment response in patients and overall drug effectiveness in clinical trial cohorts. Such decision-making requires calculating a variety of response measures in patients and across cohorts—tasks generally done by hand, making it difficult to compare multiple alternative imaging biomarkers. Fourth, it is costly and difficult to accrue aggregate data needed to qualify new imaging biomarkers as surrogate endpoints for clinical trials (58). Qualification of new imaging biomarkers requires collecting context-specific assessments of the performance of the biomarker relative to clinical outcomes (59). It is challenging to acquire sufficient data that link imaging biomarker data with clinical outcomes, such as survival (60). Efforts such as the Quantitative Imaging Biomarker Alliance (QIBA) are creating consensus on processes to qualify new imaging biomarkers (61), but their ultimate success depends on expanding public data sets (62) and leveraging many studies from individual laboratories and cooperative groups, which currently cannot be repurposed for this task because image annotations (or biomarker values extracted from them) are not recorded in standardized formats.

We developed ePAD—one of the research projects of the QIN—to address all of these challenges by developing a modular software platform integrating image viewing with computation of emerging and validated quantitative imaging biomarkers, facilitating translation of novel biomarkers into clinical trials as surrogate endpoints. In this paper, we will present ePAD's core architecture and describe the ways in which it meets the foregoing challenges. We also describe active research projects that are leveraging ePAD.

THE ePAD PLATFORM

We describe the design of ePAD and its core architecture, presenting this information from 4 different perspectives that address 4 major challenges mentioned above: (1) as a platform enabling the computing of novel imaging biomarkers of cancer treatment response, (2) as a workflow management and study oversight tool enabling the oversight for assessing new image biomarkers in clinical trials, (3) as a clinical decision support tool for the treatment response assessment using current and new imaging biomarkers, and (4) as infrastructure to permit researchers to aggregate evidence needed to show that new imaging biomarkers predict survival, which can be useful in qualifying them as surrogate endpoints in clinical trials.

Image Annotations in ePAD

A key distinguishing feature of ePAD is its support for standardized formats for image annotations, specifically Annotation and Image Markup (AIM) (63) and DICOM segmentation objects (64). AIM is an information model developed by the National Cancer Imaging Program of NCI for storing and sharing image metadata (65-67), such as lesion identification, location, size measurements, regions of interest (ROIs), radiologist observations, anatomic locations of abnormalities, calculations, inferences, and other qualitative and quantitative image features. The image metadata also include information about the image, such as the name of imaging procedure and how or when the image was acquired. AIM supports controlled terminologies, enabling semantic interoperability. In the use case of lesion annotation in cancer, the value of AIM is recording lesion identifiers (enabling unambiguous tracking of lesions across longitudinal images), anatomic locations of lesions, lesion types (target, nontarget, new lesion, or resolved lesion), and study types (baseline or follow up). This semantic information is critical for automating the generation of tabular summaries of lesions, and it also enables automating comparing the response assessment in patients according to different imaging biomarkers (see Section "Clinical Decision Support Tool for Treatment Response Assessment"). AIM has recently been incorporated into DICOM Structured Report (DICOM SR) (68), with specifications for saving AIM in DICOM-SR (69).

Architecture of ePAD

ePAD Components. ePAD (70-72) is a freely available quantitative imaging informatics platform (http://epad.stanford.edu) distributed as a virtual machine or as Docker containers. Users can download virtual machine or Docker version of ePAD and host it in their own environment. This enables them to restrict the access to their private networks, typically to the hospital network. These machines generally do not have access to the internet. The core architecture of ePAD is shown in Figure 1. The ePAD platform comprises the following 5 main components: (1) the ePAD viewer, a zero-footprint web image viewer and image annotator, (2) ePAD web services, providing a programming interface to ePAD services, (3) an image database, (4) an image annotation database, and (5) plugin modules (server-side and client-side for extending the ePAD platform). The image database, image annotation database, and ePAD web services comprise the "back end" of ePAD. The ePAD plugin modules extend the functionality of ePAD, and while most of the plugins de-

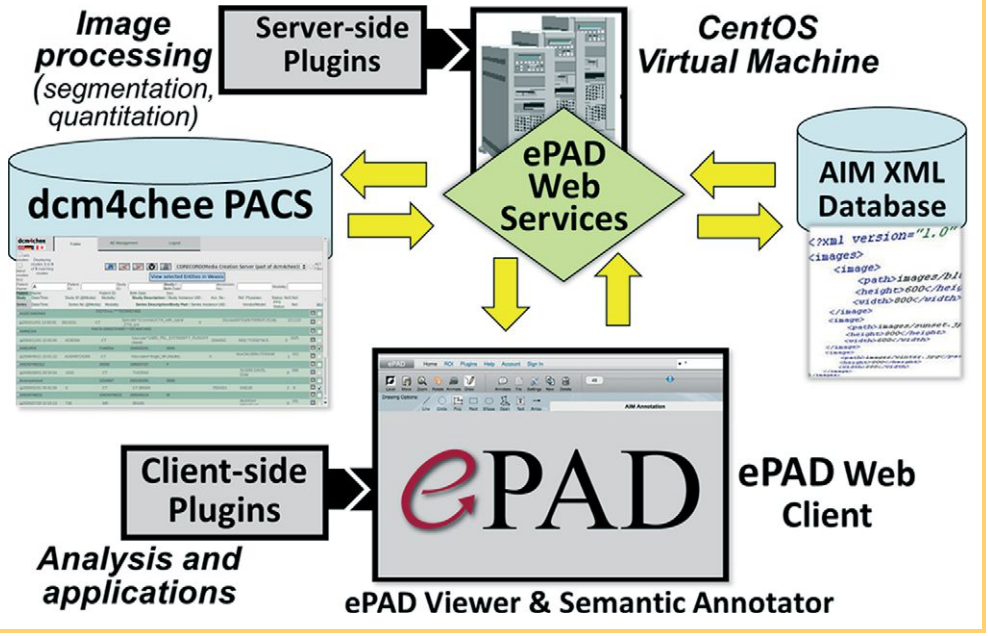


Figure 1. Architecture of the ePAD platform, which comprises an image database (dcm4chee PACS as a cache for images), a database of image annotations (AIM XML database), the ePAD Viewer (a web application), ePAD Web Services that communicate data between the image and annotation databases and the ePAD Viewer, and back-end (server-side) and front-end (client-side) plugins enabling the community to extend the ePAD platform.

scribed below were developed by us, we also describe several developed by the community, and such community engagement will enable ePAD to foster an ecosystem enabling continued evolution of the platform to meet the needs of researchers broadly.

1. ePAD Viewer. The ePAD viewer is a web application providing the look and feel of a Picture Archiving and Communication System (PACS) to the user, who can browse patient studies and open them to view images. To display images, the ePAD viewer queries an embedded PACS database [dcm4chee (73)] and stores image annotations in ePAD's annotation database. The ePAD viewer was written using HTML5 (74), Java, JavaScript, and the Google Web Toolkit (http://www.gwtproject.org), which supports image rendering with controls for image display (eg, zooming, panning, and window/level) within the Web browser. Drawing and editing image annotations are accomplished with HTML5 Scalable Vector Graphics (SVG).

An important component of the ePAD viewer is its image annotation window (Figure 2). The ePAD viewer ensures the minimum information necessary to create a meaningful image annotation is collected from the user: the lesion name, the lesion type (target, nontarget, new lesion, or resolved lesion) and the anatomic location of the lesion, and the study time point (baseline or follow-up). The ePAD viewer automatically labels each lesion with a name (eg, "Lesion1") to enable unambiguous determination of the same lesion on serial imaging studies (75). To specify the content of annotations, ePAD uses AIM templates (76) that are created by a separate freely available application. AIM templates specify the data elements to be provided by the user when making image annotations. All answer choices in ePAD templates are controlled terminology lists such as RadLex (77). The ePAD viewer prompts the user if certain values in the templates are inconsistent or incomplete (66). The ePAD viewer permits creating 2 types of ROI, coordinate based and pixel map based. The former is saved as coordinates in the AIM file (63), and the latter is saved as a DICOM segmentation object (64).

2. ePAD Web Services. The ePAD viewer uses a set of RESTful Web services (78) to communicate with the back end of ePAD to retrieve images and save image annotations, as well as authenticating user credentials and invoking image calculation methods that need to be executed on the server. The ePAD Web Services provides programmatic access to the image database and the annotation database that are components of the ePAD back end (Figure 1). The ePAD Web Services is typically hosted on a server that resides within an institution's firewall so that all traffic between the ePAD viewer and the ePAD Web Services resides within the institution's Intranet. Thus, users can use ePAD to evaluate image data containing protected health information, provided the network on which ePAD is hosted is secure. Another model for hosting ePAD is a centralized, hosted version, which could provide publicly available images (which should be deidentified for public dissemination). The ePAD Web Services are used by plugin developers to extend ePAD's functionality, either as client-side or as server-side plugins (Figure 1). Plugin developers can use the ePAD Web services to access annotations and images in their own applications or to provide extensions to the ePAD platform.

3. Image Database. Medical images in DICOM format are managed by an open-source PACS called dcm4chee (73). This PACS contains a DICOM image receiver and a programming interface that permits the ePAD Web Services to manage imaging studies within ePAD. The DICOM image database provides a temporary storage depot for images for image display and annotation in ePAD. The AIM annotations and DICOM segmentation objects in ePAD are saved indefinitely, however, as these annotations comprise the user-generated data in ePAD. Because DICOM images are large, the ePAD back end converts them into a lossless compressed PNG image object ("packed PNG") that takes each 16-bit pixel in a DICOM image and packs it into a PNG color channel before returning it to the ePAD viewer, where it is unpacked. This approach significantly reduces the volume of data provided by the server and speeds performance of the ePAD viewer. To further speed image display performance, ePAD supports the Web-Accessible DICOM Objects [WADO (79)] protocol to retrieve lossy JPG images, while the lossless packed PNGs are initially loading.

4. Annotation Database. As the user makes annotations on images in ePAD viewer, it creates AIM files. All AIM annotations are stored in an XML database [eXist (80)]. The AIM annotation database is accessible via functions in the ePAD Web Services,

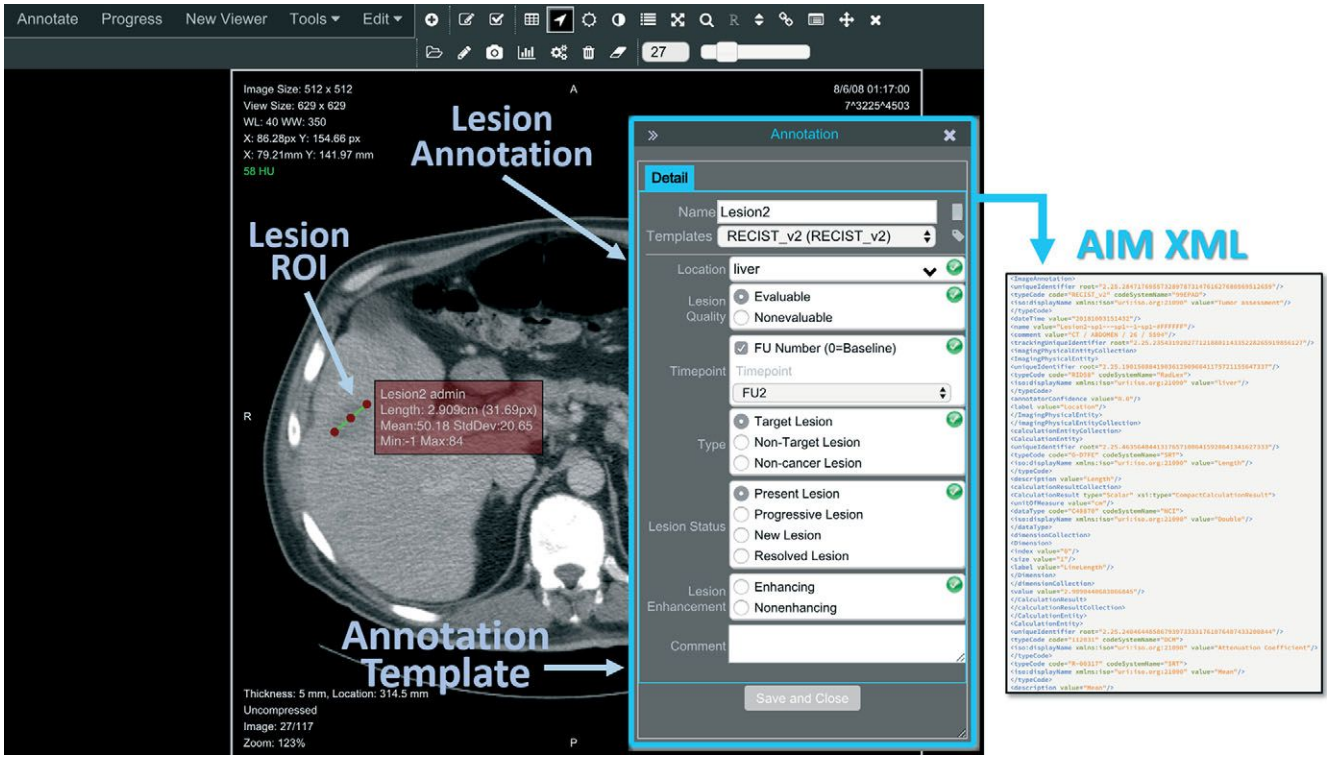


Figure 2. ePAD viewer and annotation window. Images are displayed in the ePAD web viewer, and the user records image annotations in using drawing tools (eg, to create an ROI, shown on the left) and an annotation window (to record qualitative image features, shown on right).

and it is the key resource that ePAD queries for lesion tracking and summarizing longitudinal changes in cancer treatment response, as described in Section "Clinical Decision Support Tool for Treatment Response Assessment."

5. Plugin Modules. Developers can create server-side and client-side plugins to access the data collected by ePAD to provide a new functionality. The server-side code can be written in a variety of languages, such as MATLAB, python, C/C++, or Java. We and other groups have created plugins to build a variety of features to address the challenges of (1) computing novel imaging biomarkers of cancer treatment response, (2) providing workflow management and study oversight features for assessing new image biomarkers, (3) creating clinical decision support tools for treatment response assessment using current and new imaging biomarkers, and (4) permitting researchers to aggregate evidence needed to show that new imaging biomarkers predict survival, which can be useful in qualifying them as surrogate endpoints in clinical trials.

Plugins currently available in ePAD are listed in the following sections.

JJVector Feature Extraction Plugin. JJVector is a 2D feature extraction plugin we developed that analyses closed-shape annotations and extracts 2D radiomics features based on the intensity values from the ROI and the surrounding tissue of its associated organ (81). The plugin saves the calculated feature values in an AIM file that can be downloaded in different formats from ePAD, such as an excel summary sheet to be used in other applications such as training machine learning models.

ADLA Biomarker Plugin. The Attenuation Distribution across the Long Axis (ADLA) plugin implements the ADLA semiquantitative imaging biomarker for assessing treatment response in solid malignancies and a measure of intralesional heterogeneity. We built this plugin in collaboration with prior works that created it (82, 83). ePAD calculates the standard deviation along the long axis to compute ADLA and saves it in the AIM file to be used for analyses such as response assessment as an alternative imaging biomarker. ePAD also generates an ADLA histogram of pixel values within the ROI when the long axis is selected (Figure 3).

Perfusion Analysis Plugin. A contributor developed an ePAD plugin deploying an algorithm for computing T1 perfusion maps on dynamic contrast-enhanced studies based on his prior work (84). The plugin analyses the multiframe MRI images having different phases of dynamic contrast enhancement and calculates a T1 map for the imaged volume. The plugin scales the T1 map to 8 bits to save as a standard DICOM object (a probability DICOM Segmentation object) and paints the mask on the image using a color lookup table (Figure 4).

Riesz Texture Feature Plugin. A contributor developed an ePAD plugin that computes image texture features based on Riesz wavelets (85). The latter are a subtype of convolutional approaches that can quantify image derivatives of any order and at multiple scales. The image derivatives are aligned along dominant local orientations, allowing characterization of the local organization of the image direction, with invariance to the local orientation of anatomical structures. These image derivatives have an intuitive interpretation, and the Riesz features have shown to provide valuable imaging measurements in various medical applications.

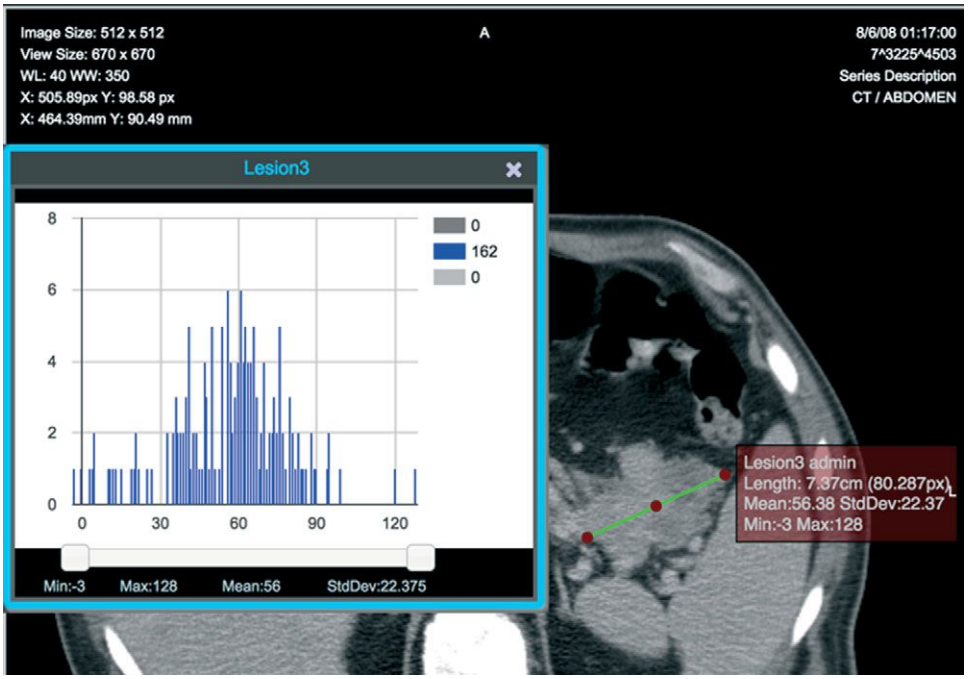


Figure 3. ADLA histogram on a line annotation on a cancer lesion created and visualized in ePAD.

Quantitative Image Feature Engine (QIFE). QIFE is an open-source feature-extraction framework we created that computes 3D radiomics features for ROIs that are created as DICOM segmentation objects (86). ePAD stores these image features in an AIM file for further analysis in radiomics studies or as alternative imaging biomarkers of response.

Quantitative Feature Explore (QFExplore) Plugin Suite. The Quantitative Feature Explore (QFExplore) is a suite of plugins we developed for the ePAD platform, enabling the exploration and validation of imaging biomarkers in a clinical environment (85). Imaging features that can be extracted using QFExplore include histogram bins of Pixel Intensity Distributions (PID), statistical moments of PIDs (ie, mean, standard deviation, skewness, kurtosis), gray-level co-occurrence matrices (GLCMs), and Riesz wavelets (87). Figure 5 illustrates QFExplore plugin suite's feature comparison functions in action. The ROIs are visualized on

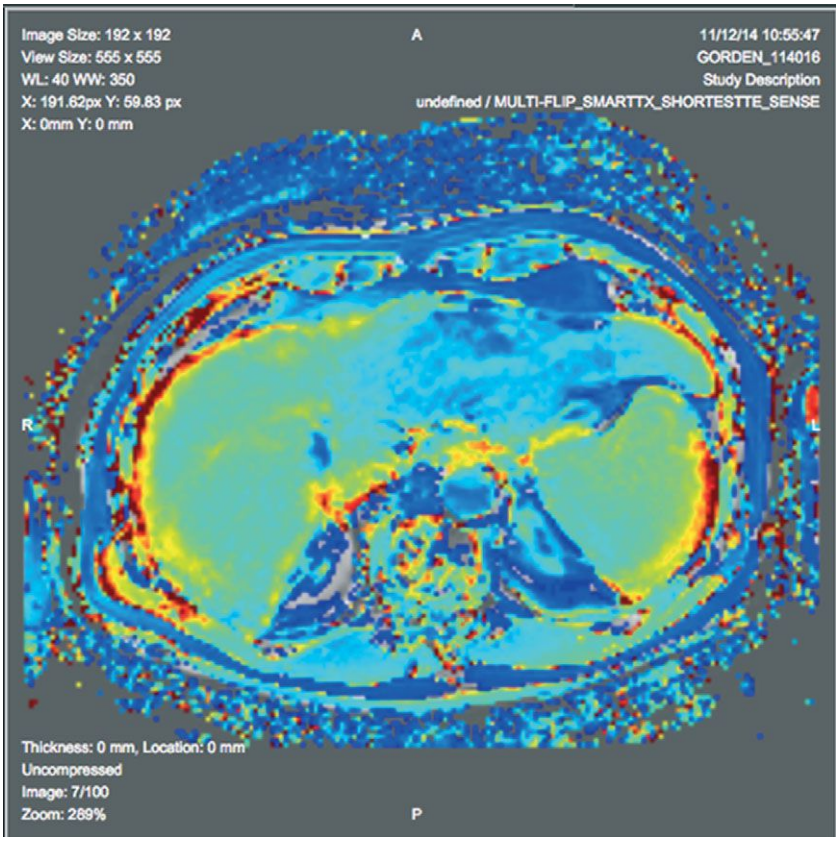


Figure 4. T1 perfusion map generated by ePAD plugin derived from the algorithm in Jarrett et al.'s study (112), with the map overlaid on magnetic resonance (MR) image using a color lookup table.

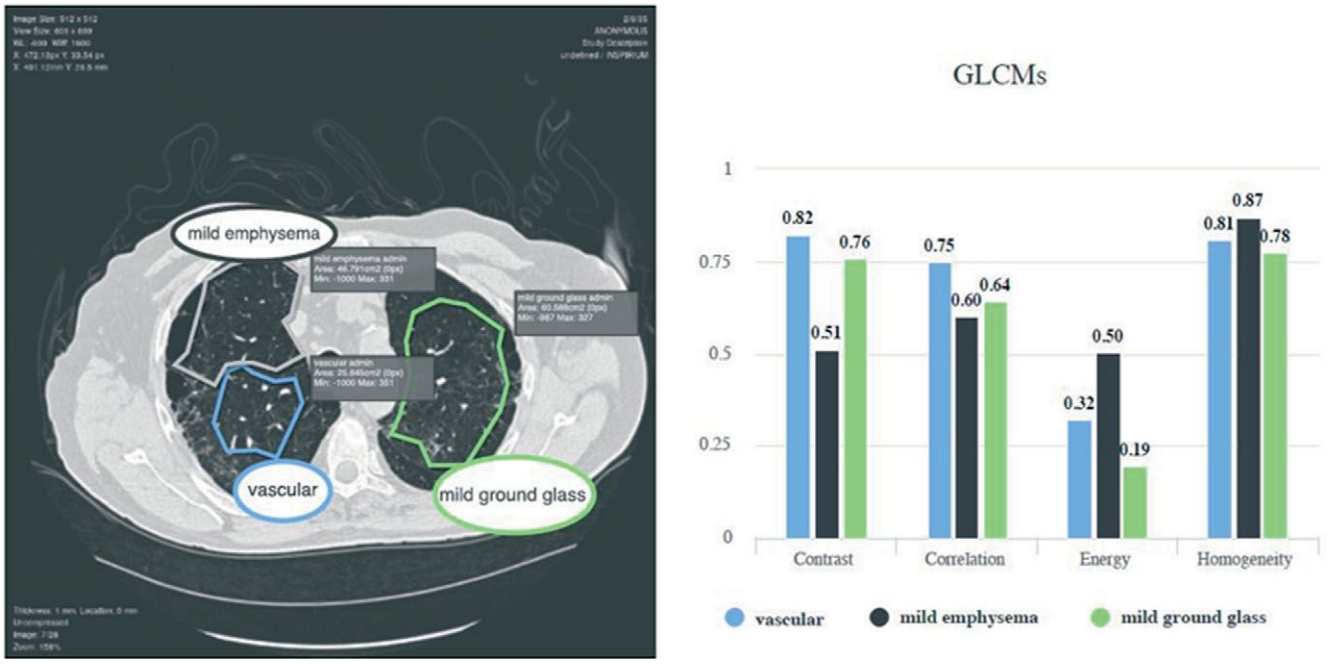


Figure 5. QF Explore Plugin Suite: gray-level co-occurrence matrix feature extraction and comparison chart. The user can compare the feature values for various regions of interest (ROIs). GLCM contrast and correlation is higher for vascular ROIs (85).

the left, while color-coded gray-level co-occurrence matrices values are displayed in a chart on the right.

Quantitative Feature Pipeline (QIFP). We created the QIFP, a cloud-based platform for building processing pipelines of image analysis algorithms (88). It provides a Docker library of image analysis algorithms for preprocessing, segmentation, and feature extraction that can be assembled into pipelines. The QIFP is integrated with ePAD so that any processing pipeline for generating quantitative imaging biomarkers can be executed in ePAD (or ePAD annotations can be consumed and used in QIFP processing pipelines).

ePAD APPLICATIONS

ePAD includes several applications that are part of the platform and accessible via menu tabs in the ePAD viewer.

Computing and Comparing Imaging Biomarkers

A need that is critical for research is its ability to compute a variety of alternative imaging biomarkers besides linear dimension (used in RECIST and similar criteria). In a given clinical trial, patient response to treatment can be computed using a variety of imaging biomarkers, and a sizeable collection of data can be amassed if this is done across clinical trials that could ultimately be useful in comparing and evaluating alternative imaging biomarkers as secondary endpoints of response. Different imaging biomarker algorithms are written in different languages, and ePAD enables incorporating them into its image analysis workflow through its plugin mechanism described above. These plugins can execute source code modules written in MATLAB, Java, C/C++, or other languages, letting biomarker algorithm developers add their existing code to ePAD easily.

When users make annotations on images, ePAD automatically analyzes each annotation to generate the image biomarkers that the user chooses, and it saves them in AIM format. It also computes the minimum, maximum, standard deviation and mean for all the pixels that are inside the ROI. If the ROI is a line, ePAD calculates the length. If the ROI comprises 2 perpendicular lines, ePAD will calculate the length of the long axis and short axis. Additional features and biomarker candidates can be calculated by various plugins.

Workflow Management and Study Oversight

The ePAD viewer includes an application that provides a summary panel of annotations designed to streamline the task of summarizing for the radiologist all prior measurements and images in prior studies of each patient to convey the list of lesions previously measured, and which need to be measured on the current study. To populate this summary display, the ePAD viewer queries ePAD's annotation database to find all the lesions from the prior exams and list them for the user. This provides the user with a worklist of lesion measurements that need to be made for each imaging study. It also links each measurement to the image from which it was obtained. When the user clicks on a measurement, the corresponding image is retrieved and the measurement is displayed.

ePAD also facilitates oversight and managing image readings for clinical trial researchers and study administrators via user roles, worklists, and study progress monitoring. Project owners and administrators can create users and assign them

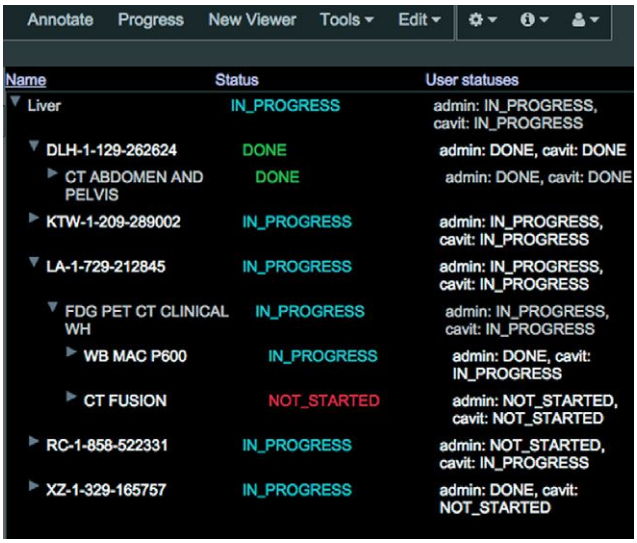


Figure 6. Progress view of ePAD visualizing a particular project ("Liver") that contains 5 patients. The status column shows the overall status for that series/study or patient, and the user statuses column shows the status of annotations that have been created by each ePAD user associated with that project.

specific roles to control their access to imaging data and annotations created by other users. Users or study supervisors can create worklists for people and assign to a reader. Using worklists allows the supervisors to divide the readings to multiple readers. A study progress monitoring application module in ePAD monitors the status of image annotations made in clinical trials and summarizes them in a table in the ePAD viewer. Study administrators can follow the image annotations made in multiple studies by group of users assigned to a particular study. The application can also track the progress of the annotation process by identifying which subjects/studies are fully annotated by all the annotators, which annotators have completed the annotation process for each subject and which subjects/studies have not yet been annotated yet (Figure 6). This functionality has been helpful the MGH/HST Martinos Center for Biomedical Imaging used this for MEDICI project (89), which used ePAD.

Clinical Decision Support Tool for Treatment Response Assessment

ePAD has applications to assist decision-making based on image biomarker assessments in the following 2 major cancer research tasks: determine treatment response in patients (ePAD longitudinal annotation report) and evaluate treatment effectiveness by determining the cohort-based treatment response (ePAD waterfall plot). We built these applications using ePAD Web services to retrieve AIM annotations and their associated images to track target lesions and compute cancer treatment response according to selected imaging biomarkers.

Longitudinal Annotation Reporting. ePAD supports longitudinal annotation tracking, which provides a summary of quantitative image features across time. This is the basis for RECIST and other reports of response assessment. However, ePAD can generate such reports bases on any quantitative imaging biomarker it can collect from image annotations. It analyzes all the annotations of a subject and populates 3 dropdown menus to facilitate selecting them by shape, template, and measurement type (Figure 7). Users can select the basis for the longitudinal annotation report based on the selected measurement types. If a measurement is not present for a particular time point of a lesion, the table display it as a missing value. The summary section of the report will be filled automatically for the measurement type.

ePAD can generate a RECIST report by querying the annotations that are of linear type (Figure 7) and calculating sum of lesion dimensions on the images of each time point. RECIST report generation supports line and perpendicular lines annotations, as well as an image-based response rate (the percentage change in the sum of lesion dimensions compared with baseline). ePAD applies the RECIST rules to classify the response rate to determine the response category (ie, stable disease, partial response, complete response, and progressive disease). This information is displayed with the lesion measurements in the ePAD viewer (Figure 7). ePAD also checks the consistency of the annotations to determine if the anatomic location of the lesion

Figure 7. A tumor burden report (using linear measurement as the imaging biomarker and RECIST response criteria) and a longitudinal annotation report of a patient having 4 time points and 3 lesions. This report is automatically generated from ePAD's image annotations and enables clinicians to determine image-based treatment response in the patient.

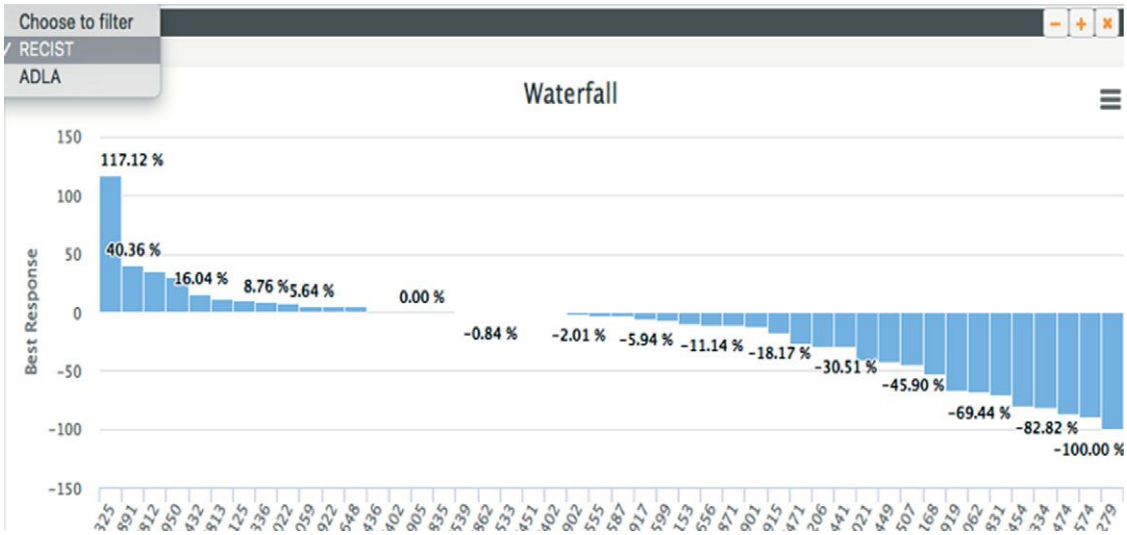


Figure 8. Waterfall report plot based on linear measurement as the imaging biomarker of response and RECIST as the response criteria, showing the best response score for each patient in the study cohort. This plot enables researchers to assess the effectiveness of cancer treatment in the cohort, and a variety of these plots can be generated using different imaging biomarkers of response (upper left corner).

is specified consistently on different time points of the patient; otherwise, the measurement will be marked as error to notify the user. The report also marks missing annotations for a lesion as error. The user can open the annotation in the ePAD viewer by clicking on the annotation measurement on a specific time point. The user can also open all annotations of the lesions on all time points by clicking the lesion name.

Besides using longitudinal measurement of lesions, ePAD can generate reports of lesion response based on other imaging biomarkers, such as the ADLA biomarker (82). The report evaluates the progress of the disease using the sum of ADLA scores for each timeline, similar to RECIST.

Waterfall Plots. Waterfall plots are bar graphs showing the response of a cohort of patients to the same cancer treatment. The height of the bars represents the best overall response the patient had during the course of treatment, and each bar (patient) is ordered from best to worst response, which resembles a waterfall. These plots are highly useful for seeing how well a patient cohort responded to treatment, with the percentage of patients with positive response indicating effective treatment. ePAD generates waterfall plots of user-specified patient cohorts by computing the longitudinal annotation-based response in each patient in the cohort and ordering the response from best to worst response (Figure 8). The plot can be based on longitudinal measurement of lesions as the basis of evaluating response (ie, RECIST), but importantly, it can also be based on using newer imaging biomarkers of response such as ADLA or other imaging biomarkers that have been recently introduced by researchers. If the user selects to use RECIST, the waterfall plot module analyzes every subject in the cohort, generates the RECIST tables, gets the best response for each subject, and plots it in a decreasing order forming a waterfall plot. If the user selects to use ADLA, a waterfall plot is generated based on an ADLA table that ePAD computes for each subject, using the standard deviation of the line annotations on lesions as the measurement type (82). Then, the best response from the ADLA table for each patient is used to create the waterfall plot. Users can drill down to more granular data within the waterfall plot; the user can access the table that is used to make the best response rate computation by clicking the specific bar in the waterfall plot.

Application for Aggregating Evidence for Evaluating New Imaging Biomarkers

As noted earlier, ePAD has plugins to compute a variety of image biomarkers. Some of these plugins assume that cancer lesions are circumscribed, and if the images input into these biomarker plugins were annotated using only line annotations (eg, as part of RECIST measurements), ePAD can generate ROIs that circumscribe lesions automatically by executing image segmentation plugins that use the line or point annotations as seeds. In addition to segmentation, there are quantitative image analysis plugins that may operate on the entire image, and ePAD supports those as well. Current automated segmentation plugins and other analysis plugins available in ePAD are listed in the following sections.

Automated Segmentation in PET Images. The plugin invokes automated segmentation of cancer lesions seen on PET images (90). It is triggered with a seed point ROI within the lesion. It analyzes the image volume to create a 3D ROI enclosing the lesion and creates a DICOM segmentation object marking the volume of the lesion. The DICOM segmentation object is added to ePAD with its associated AIM annotation file and displayed on the image series as a mask.

Automated 2D Lesion Segmentation. ePAD has a 2D lesion segmentation plugin, LesionSeg. The plugin is triggered with drawing a polygon or a long axis line within a lesion. It analyzes the image and creates a polygon ROI stored as an AIM file containing the contours of the lesion (91).

Automated Image Segmentation of QF Explore Plugin Suite. The QFExplore plugin suite has a plugin for automatically segmenting lungs in a DICOM image volume (85). The plugin

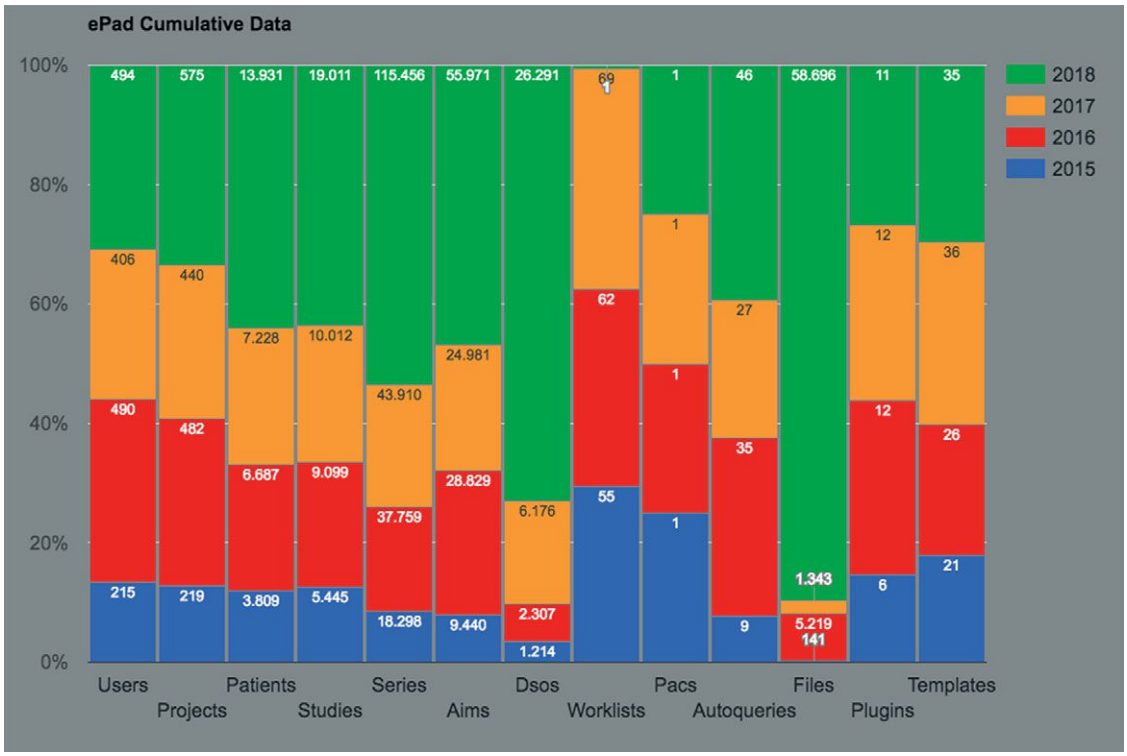


Figure 9. Cumulative ePAD statistics collected from ePAD instances between 2015 and 2018.

analyses the volume, segments the lung volume, and creates a DICOM segmentation object.

ePAD USAGE

To track usage statistics, ePAD collects anonymous data from all ePAD machines that are connected to internet (if the statistics are not disabled by the user). The statistics consist the number of users, projects, patients, studies, series, AIM annotations, DICOM segmentation objects, plugins, and templates that exist on the ePAD instances. Figure 9 shows the ePAD usage statistics collected from 2015 to 2018. For plugins and templates, the maximum number of entities is reported, as many are the same versions of the plugins and templates across ePAD instances. For all the other entities, the values reported by each ePAD instance are computed by getting the latest reported values for each year and summing them to obtain the total number. For example, in 2018, over 19,000 imaging studies were hosted in various ePAD instances worldwide, and over 55,000 annotations were created in ePAD on those imaging studies. As ePAD collects only the number of entities for privacy purposes, the numbers are cumulative; that is, this does not mean 55,000 annotations were created during 2018, but it means that at the end of 2018, 55,000 annotations existed on ePAD instances. In addition, currently there are a maximum of 11 plugins and 35 templates that are being used across all ePAD instances.

INTEROPERABILITY

One of the key aims of ePAD is to facilitate collaborations among research sites and repurposing of their existing data, which we achieve by supporting standards and interoperability for images and annotations.

ePAD saves all image annotations that it collects using existing standards, in particular AIM (63) and DICOM segmentation objects (64), for volumetric ROIs. ePAD also supports the DICOM-SR standard via the dcmqi library (92) for volumetric ROI annotations. Recently AIM was harmonized with the DICOM standard, which provided DICOM-SR support of AIM annotation types under Supplement 200 with specifications for saving AIM in DICOM-SR (68, 69). ePAD also supports DICOM radiation therapy (DICOM-RTs) and tiff image files. ePAD analyses the DICOM-RT objects and extracts its ROI contours using the DICOM file interface library developed by MAASTRO (93). It then creates a DICOM segmentation object for each contour and saves it and an AIM file. ePAD also supports uploading tiff files and creates a DICOM image series from them using the patient identification number, patient name, study description, and series description supplied by the user. The file list is analyzed, and a DICOM file is created for each tiff file. The instance numbers of the DICOM files are ordered in the alphabetical order of tiff filenames.

ePAD also has migration tools that were developed in collaboration with various laboratories that enable ePAD to leverage the existing annotations created by other software tools, including ROIs exported from Osirix (94) and Mint Lesion (53). Specifically, ePAD analyzes the exported proprietary file from Osirix via ExportROIs plugin and creates an AIM file for each ROI in the file. ePAD also creates AIM files from JavaScript Object Notation (JSON) objects that are created from the Mint Lesion commercial system.

USE CASES FOR ePAD IN QIN AND OTHER RESEARCH

Many research studies in that require viewing and annotating radiology images for making measurements of lesions or extracting radiomics features from them could benefit from using ePAD. Clinical trials of cancer treatments can be particularly helped given ePAD's workflow support and multireader support features, its support of interoperability standards, as well as its ability to compute many imaging biomarkers seamlessly within routine image annotation workflow. ePAD has been used by many researchers worldwide to support clinical research and clinical trials, and it has supported many published studies (75, 81, 85, 95-110), and it has been shown to improve the workflow of measuring target lesions (111). We briefly highlight support it has provided several projects in NCI's QIN.

Vanderbilt QIN. In collaboration with Vanderbilt QIN, "Quantitative MRI for Predicting Response of Breast Cancer to Neoadjuvant Therapy" in which this group developed algorithms for computing quantitative perfusion maps of MRI images to deduce biomarkers of treatment response (112), we deployed their biomarker algorithms as a plugin to ePAD. As these researchers incorporate these perfusion analyses into clinical trials, ePAD will be able to deploy them as part of the image interpretation workflow.

Dana Farber Cancer Institute QIN. The QIN project at Dana Farber, "Genotype and Imaging Phenotype Biomarkers in Lung Cancer (113)," developed pyRadiomics, a flexible platform that extracts a large panel of predefined features from medical images and is useful in characterizing cancer lesions. We incorporated pyRadiomics into ePAD as a Docker module that runs on the QIFP platform (88) (see above) so that users can invoke generation of these image features as part of image analysis workflows in clinical trials.

ECOG-ACRIN QIN. The QIN project within the ECOG-ACRIN cooperative group, "ECOG-ACRIN-Based QIN Resource for Advancing Quantitative Cancer Imaging in Clinical Trials," is leveraging ePAD as a testbed for evaluating the deployment of imaging biomarkers into clinical trials. Currently this project is comparing ability of ePAD to evaluate a variety of quantitative imaging biomarkers as part of the routine workflow of image viewing and annotation in clinical trials.

American College of Radiology (ACR) Core Laboratory. The ACR has a data archive and research toolkit called DART Portal (114) that operates as a gateway to browse and query data for research, quality improvement, and clinical study operational purposes. They are adding ePAD as an interface to DART to enable collecting image annotations as part of clinical trials in AIM format and storing that in DART.

DISCUSSION

Response assessment in patients with cancer in clinical trials is based on analysis of CT and magnetic resonance images (115). Objective criteria, such as RECIST, are critical to evaluation of response assessment in clinical trials, but lesion measurements vary with user experience, and they are often inconsistent or incomplete (105). There is a pressing need to recognize signals in radiology images that optimally assess and predict response to treatment. Tumor shrinkage is the hallmark of response to cytotoxic cancer therapies (4), and thus, linear measurement of target cancer lesions is the imaging biomarker used in current response criteria such as RECIST and International Harmonization Criteria (10). However, new targeted, noncytotoxic therapies arrest cancer growth and improve progression-free survival without necessarily shrinking tumors (11-14); thus, simple linear measurement may underestimate treatment response (15-18), and may not be the best proxy for tumor activity. To address these limitations, researchers are developing quantitative imaging biomarkers that may better assess the benefit of new treatments, but they have been challenging to introduce into clinical trial workflows.

In the paper, we have presented ePAD from 4 different viewpoints to highlight how it addresses key challenges for incorporating quantitative imaging biomarkers into clinical trials. First, it provides a platform for computing a variety of imaging biomarkers. Second, it provides a workflow management and study oversight tool enabling oversight for assessing new image biomarkers in clinical trials. Third, it provides clinical decision support tools to help clinical researchers assess treatment response using current and new imaging biomarkers. Fourth, ePAD provides infrastructure to permit researchers to aggregate evidence about how well imaging biomarkers predict response, which may help in qualifying them as surrogate endpoints in clinical trials.

There are many existing commercial and freely available tools available for medical image viewing and annotation, and although ePAD provides similar capability in terms of image viewing and drawing shapes on images, it provides many unique features that address many unmet needs in evaluating images clinical research. Osirix (94) and ClearCanvas (55) are 2 medical image annotation applications that provide similar image viewing capabilities, although they depend on a thick client, limiting collaboration as they are platform-dependent, while ePAD is a web-based viewer and requires no installation for users other than hosting a single instance of the ePAD machine for all users. In addition, Osirix saves its annotations in propriety format and supports exporting ROIs using ExportROIs plugin. On the other hand, ePAD and ClearCanvas supports AIM format. ePAD also supports the new DICOM-SR AIM object. Beyond the open-source tools, we recognize that several commercial tools are available for image viewing and analysis to enable response assessment. These tools were developed to enable evaluating established criteria such as RECIST in clinical trials; however, such tools are not optimal for research studies that wish to include novel imaging biomarkers of treatment response (eg, those being developed by NCI's QIN). This gap was a primary motivator for developing the ePAD system.

3D Slicer (56, 116), ImageJ (117), and MIPAV (118) are additional freely available desktop applications. 3D Slicer is a cross-platform open-source software for visualization and image computing. It has a plugin architecture to enable researchers to develop their algorithms via C++ plugins and Python scripted modules. It supports DICOM standard for the volumetric annotations and DICOM Structured Report (DICOM-SR) for the measurements collected from the ROIs. ImageJ and MIPAV are both Java applications that can run on any Java-enabled operating system, and researchers can develop their own plugins using Java language. Imagej2, an extended version of ImageJ, supports writing plugin scripts in various programming languages. ImageJ saves the labels and annotations as modified versions of the images or propriety ROI file formats, and MIPAV uses an Extensible Markup Language (XML) format they introduced in an effort to make their format readable by researchers.

Although all 3 applications are cross-platform, they are desktop applications for a single user, which makes multiuser collaboration more difficult.

The Open Health Imaging Foundation (OHIF; http://ohif.org/) is a full-stack Javascript platform, which enables creating a zero-footprint web page and various applications using it. The OHIF Viewer provides web-based image viewing similar to ePAD. The OHIF LesionTracker enables users to annotate and track long-axis and short-axis lesions for oncology workflow; however, it does not save the image annotations in a standard format like AIM.

ePAD was developed to facilitate collecting annotations and measurements on target lesions in compliance with standards in the cancer imaging community. ePAD makes sharing code, data and annotations easy being a web application and saving the collected annotation data in well-documented and standardized formats [DICOM segmentation objects (64) and AIM (63) in particular].

In addition to providing standards-based storage of annotations, ePAD enables user-defined templates for flexible capture of information in the form of data collection templates as part of the annotations. The ePAD platform is also extensible via plugins that lets researchers implement analysis codes as server-side modules in MATLAB or other languages. Many plugins for segmentation and quantitative image biomarker computation are included with ePAD, and users can add additional biomarker modules.

Other functionalities of ePAD that differentiates it from similar existing image viewing applications is that ePAD supports important features unique to image analysis in clinical trial workflow. Specifically, ePAD provides tools enabling oversight of annotations as part of clinical trials, and it lets the users create a collaborative environment by creating projects and assigning users appropriate rights to limit their access facilitating large studies with multiple annotators. ePAD also provides decision support tools—longitudinal annotation summary and waterfall plots—that help researchers evaluate individual patient and cohort population treatment response, respectively. Finally, by computing a variety of image biomarkers on cohorts of patients, ePAD can accumulate a substantial amount of data that can permit studies comparing effectiveness of different imaging biomarkers as indicators of treatment response.

An ultimate metric of the success of ePAD will be increased use of the newer imaging biomarkers in clinical trials. This will require clinical trial groups to include computation of the biomarkers into their study protocols. As the community becomes aware of the potential of these methods and of the facility of tools such as ePAD to include them in clinical trials, we expect these methods will be more commonly used. Certainly, the amount of research studies undertaken to date using ePAD suggests promising future directions.

ACKNOWLEDGMENTS

This work was supported in part by grants from the National Cancer Institute, National Institutes of Health, U01CA142555, 1U01CA190214, and 1U01CA187947.

Disclosures: No disclosures to report.

Conflict of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. Garrett MD, Workman P. Discovering novel chemotherapeutic drugs for the third millennium. Eur J Cancer. 1999;35:2010–2030.
2. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100:57–70.
3. NIH Clinical Trials Working Group. Final CTWG report to the National Cancer Advisory Board, "Restructuring the National Cancer Clinical Trials Enterprise". http://integratedtrials.nci.nih.gov/ict/CTWG_report_June2005.pdf; Accessed: December 21, 2008.
4. El-Maraghi RH, Eisenhauer EA. Review of phase II trial designs used in studies of molecular targeted agents: outcomes and predictors of success in phase III. J Clin Oncol. 2008;26:1346–1354.
5. Taylor PT, Haverstick D. Re: New guidelines to evaluate the response to treatment in solid tumors (ovarian cancer). J Natl Cancer Inst. 2005;97:151.
6. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, Verweij J, Van Glabbeke M, van Oosterom AT, Christian MC, Gwyther SG. New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. J Natl Cancer Inst. 2000;92:205–216.
7. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, Dancey J, Arbuck S, Gwyther S, Mooney M, Rubinstein L, Shankar L, Dodd L, Kaplan R, Lacombe D, Verweij J. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). Eur J Cancer. 2009;45:228–245.
8. Wen PY, Macdonald DR, Reardon DA, Cloughesy TF, Sorensen AG, Galanis E, Degroot J, Wick W, Gilbert MR, Lassman AB, Tsien C, Mikkelsen T, Wong ET, Chamberlain MC, Stupp R, Lamborn KR, Vogelbaum MA, van den Bent MJ, Chang SM. Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. J Clin Oncol. 2010;28:1963–1972.
9. Gallego Perez-Larraya J, Lahutte M, Petrirena G, Reyes-Botero G, Gonzalez-Aguilar A, Houillier C, Guillevin R, Sanson M, Hoang-Xuan K, Delattre JY. Response assessment in recurrent glioblastoma treated with irinotecan-bevacizumab: comparative analysis of the Macdonald, RECIST, RANO, and RECIST + F criteria. Neuro Oncol. 2012;14:667–673.
10. Cheson BD. The International Harmonization Project for response criteria in lymphoma clinical trials. Hematol Oncol Clin North Am. 2007;21:841–854.
11. Escudier B, Eisen T, Stadler WM, Szczylik C, Oudard S, Siebels M, Negrier S, Chevreau C, Solska E, Desai AA, Rolland F, Demkow T, Hutson TE, Gore M, Freeman S, Schwartz B, Shan M, Simantov R, Bukowski RM. Sorafenib in advanced clear-cell renal-cell carcinoma. N Engl J Med. 2007;356:125–134.
12. Ratain MJ, Eisen T, Stadler WM, Flaherty KT, Kaye SB, Rosner GL, Gore M, Desai AA, Patnaik A, Xiong HQ, Rowinsky E, Abbruzzese JL, Xia C, Simantov R, Schwartz B, O'Dwyer PJ. Phase II placebo-controlled randomized discontinuation trial of sorafenib in patients with metastatic renal cell carcinoma. J Clin Oncol. 2006;24:2505–2512.
13. Ratain MJ. Phase II oncology trials: let's be positive. Clin Cancer Res. 2005;11:5661–5662.
14. Ratain MJ, Eckhardt SG. Phase II studies of modern drugs directed against new targets: if you are fazed, too, then resist RECIST. J Clin Oncol. 2004;22:4442–4445.
15. Benjamin RS, Choi H, Macapinlac HA, Burgess MA, Patel SR, Chen LL, Podoloff DA, Charnsangavej C. We should desist using RECIST, at least in GIST. J Clin Oncol. 2007;25:1760–1764.
16. Choi H, Charnsangavej C, de Castro Faria S, Tamm EP, Benjamin RS, Johnson MM, Macapinlac HA, Podoloff DA. CT evaluation of the response of gastrointestinal stromal tumors after imatinib mesylate treatment: a quantitative analysis correlated with FDG PET findings. AJR Am J Roentgenol. 2004;183:1619–1628.

17. Adams VR, Leggas M. Sunitinib malate for the treatment of metastatic renal cell carcinoma and gastrointestinal stromal tumors. Clin Ther. 2007;29:1338–1353.
18. Choi H. Response evaluation of gastrointestinal stromal tumors. Oncologist. 2008;13(*Suppl* 2):4–7.
19. Therasse P, Eisenhauer EA, Verweij J. RECIST revisited: a review of validation studies on tumour assessment. Eur J Cancer. 2006;42:1031–1039.
20. Cyran CC, Paprottka PM, Eisenblatter M, Clevert DA, Rist C, Nikolaou K, Lauber K, Wenz F, Hausmann D, Reiser MF, Belka C, Niyazi M. Visualization, imaging and new preclinical diagnostics in radiation oncology. Radiat Oncol. 2014;9:3.
21. Gatenby RA, Grove O, Gillies RJ. Quantitative imaging in cancer evolution and ecology. Radiology. 2013;269:8–15.
22. Lemasson B, Galban CJ, Boes JL, Li Y, Zhu Y, Heist KA, Johnson TD, Chenevert TL, Galban S, Rehemtulla A, Ross BD. Diffusion-weighted MRI as a biomarker of tumor radiation treatment response heterogeneity: a comparative study of whole-volume histogram analysis versus voxel-based functional diffusion map analysis. Transl Oncol. 2013;6:554–561.
23. Oh D, Lee JE, Huh SJ, Park W, Nam H, Choi JY, Kim BT. Prognostic significance of tumor response as assessed by sequential 18F-fluorodeoxyglucose-positron emission tomography/computed tomography during concurrent chemoradiation therapy for cervical cancer. Int J Radiat Oncol Biol Phys. 2013;87:549–554.
24. Teng FF, Meng X, Sun XD, Yu JM. New strategy for monitoring targeted therapy: molecular imaging. Int J Nanomedicine. 2013;8:3703–3713.
25. Smith AD, Lieber ML, Shah SN. Assessing tumor response and detecting recurrence in metastatic renal cell carcinoma on targeted therapy: importance of size and attenuation on contrast-enhanced CT. AJR Am J Roentgenol. 2010;194:157–165.
26. Weber WA. Assessing tumor response to therapy. J Nucl Med. 2009;(50 *Suppl* 1):1S–10S.
27. Worhunsky DJ, Krampitz GW, Poullos PD, Visser BC, Kunz PL, Fisher GA, Norton JA, Poultsides GA. Pancreatic neuroendocrine tumours: hypoenhancement on arterial phase computed tomography predicts biological aggressiveness. HPB (Oxford). 2013;16:304–311.
28. Workman P, Aboagye EO, Chung YL, Griffiths JR, Hart R, Leach MO, Maxwell RJ, McSheehy PM, Price PM, Zweit J. Minimally invasive pharmacokinetic and pharmacodynamic technologies in hypothesis-testing clinical trials of innovative therapies. J Natl Cancer Inst. 2006;98:580–598.
29. U.S. Department of Health and Human Services Food and Drug Administration. Guidance for Industry: Clinical Trail Endpoints the Approval of Cancer Drugs and Biologics. [cited 2014 Jan 19]. Available from: http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&ved=0CCkQFjAA&url=http%3A%2F%2Fwww.fda.gov%2Fdownloads%2FDrugs%2FGuidanceComplianceRegulatoryInformation%2FGuidances%2FUCM268555.pdf&ei=n2XcUuPoGdDdoASnooGoAw&usg=AFQjCNGiW-duJ8M5E1pqEEVWyyPafAJCqA&sig2=wsdc99GnjhhUNO4qjyCkMg&bvm=bv.59568121,d.cGU
30. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther. 2001;69:89–95.
31. Levy MA, Freymann JB, Kirby JS, Fedorov A, Fennessy FM, Eschrich SA, Berglund AE, Fenstermacher DA, Tan Y, Guo X, Casavant TL, Brown BJ, Braun TA, Dekker A, Roelofs E, Mountz JM, Boada F, Laymon C, Oborski M, Rubin DL. Informatics methods to enable sharing of quantitative imaging research data. Magn Reson Imaging. 2012;30:1249–1256.
32. Kelloff GJ, Hoffman JM, Johnson B, Scher HI, Siegel BA, Cheng EY, Cheson BD, O'Shaughnessy J, Guyton KZ, Mankoff DA, Shankar L, Larson SM, Sigman CC, Schilsky RL, Sullivan DC. Progress and promise of FDG-PET imaging for cancer patient management and oncologic drug development. Clin Cancer Res. 2005;11:2785–2808.
33. Munden RF, Swisher SS, Stevens CW, Stewart DJ. Imaging of the patient with non-small cell lung cancer. Radiology. 2005;237:803–818.
34. Johnson JR, Williams G, Pazdur R. End points and United States Food and Drug Administration approval of oncology drugs. J Clin Oncol. 2003;21:1404–1411.
35. Pien HH, Fischman AJ, Thrall JH, Sorensen AG. Using imaging biomarkers to accelerate drug development and clinical trials. Drug Discov Today. 2005;10:259–266.
36. Schnall M, Rosen M. Primer on imaging technologies for cancer. J Clin Oncol. 2006;24:3225–3233.
37. Shankar LK, Sullivan DC. Functional imaging in lung cancer. J Clin Oncol. 2005;23:3203–3211.
38. Smith JJ, Sorensen AG, Thrall JH. Biomarkers in imaging: realizing radiology's future. Radiology. 2003;227:633–638.
39. O'Connor JP, Jackson A, Asselin MC, Buckley DL, Parker GJ, Jayson GC. Quantitative imaging biomarkers in the clinical development of targeted therapeutics: current and future perspectives. Lancet Oncol. 2008;9:766–776.
40. Galanis E, Buckner JC, Maurer MJ, Sykora R, Castillo R, Ballman KV, Erickson BJ. Validation of neuroradiologic response assessment in gliomas: measurement by RECIST, two-dimensional, computer-assisted tumor area, and computer-assisted tumor volume methods. Neuro Oncol. 2006;8:156–165.
41. Choi H, Charnsangavej C, Faria SC, Macapinlac HA, Burgess MA, Patel SR, Chen LL, Podoloff DA, Benjamin RS. Correlation of computed tomography and positron emission tomography in patients with metastatic gastrointestinal stromal tumor treated at a single institution with imatinib mesylate: proposal of new computed tomography response criteria. J Clin Oncol. 2007;25:1753–1759.
42. Fendler WP, Philippe Tiega DB, Ilhan H, Paprottka PM, Heinemann V, Jakobs TF, Bartenstein P, Hacker M, Haug AR. Validation of several SUV-based parameters derived from 18F-FDG PET for prediction of survival after SIRT of hepatic metastases from colorectal cancer. J Nucl Med. 2013;54:1202–1208.
43. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. J Nucl Med. 2009;50(*Suppl* 1):122S–150S.
44. Downey RJ, Akhurst T, Gonen M, Vincent A, Bains MS, Larson S, Rusch V. Preoperative F-18 fluorodeoxyglucose-positron emission tomography maximal standardized uptake value predicts survival after lung cancer resection. J Clin Oncol. 2004;22:3255–3260. 15310769.
45. Rizk N, Downey RJ, Akhurst T, Gonen M, Bains MS, Larson S, Rusch V. Preoperative 18[F]-fluorodeoxyglucose positron emission tomography standardized uptake values predict survival after esophageal adenocarcinoma resection. Ann Thorac Surg. 2006;81:1076–1081.
46. Guermazi A, Juweid ME. Commentary: PET poised to alter the current paradigm for response assessment of non-Hodgkin's lymphoma. Br J Radiol. 2006;79:365–367.
47. O'Connor JP, Jackson A, Parker GJ, Jayson GC. DCE-MRI biomarkers in the clinical evaluation of antiangiogenic and vascular disrupting agents. Br J Cancer. 2007;96:189–195.
48. Chang EY, Li X, Jerosch-Herold M, Priest RA, Enestvedt CK, Xu J, Springer CS, Jr., Jobe BA. The evaluation of esophageal adenocarcinoma using dynamic contrast-enhanced magnetic resonance imaging. J Gastrointest Surg. 2008;12:166–175.
49. d'Arcy JA, Collins DJ, Padhani AR, Walker-Samuel S, Suckling J, Leach MO. Informatics in Radiology (infoRAD): Magnetic Resonance Imaging Workbench: analysis and visualization of dynamic contrast-enhanced MR imaging data. Radiographics. 2006;26:621–632.
50. Galban CJ, Chenevert TL, Meyer CR, Tsien C, Lawrence TS, Hamstra DA, Junck L, Sundgren PC, Johnson TD, Ross DJ, Rehemtulla A, Ross BD. The parametric response map is an imaging biomarker for early cancer treatment outcome. Nat Med. 2009;15:572–576.
51. Li X, Dawant BM, Welch EB, Chakravarthy AB, Freehardt D, Mayer I, Kelley M, Meszoely I, Gore JC, Yankeelov TE. A nonrigid registration algorithm for longitudinal breast MR images and the analysis of breast tumor response. Magn Reson Imaging. 2009;27:1258–1270.
52. MIM Software. MIMviewer® and PET Edge™. [cited 2014 Jan 28]. Available from: http://www.mimsoftware.com/products/radnuc
53. Mint Medical. Mint Lesion™. [cited 2014 Jan 28]. Available from: http://www.mint-medical.de/productssolutions/mintlesion/mintlesion/
54. Siemens Inc. syngo.via for Oncology. [cited 2014 Jan 28]. Available from: http://www.healthcare.siemens.com/medical-imaging-it/syngoviaspecialtopics/syngo-via-for-oncology/syngo-via-for-oncology-follow-up
55. ClearCanvas Workstation. In, 2014, Available from: http://clearcanvas.ca
56. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin JC, Pujol S, Bauer C, Jennings D, Fennessy F, Sonka M, Buatti J, Aylward S, Miller JV, Pieper S, Kikinis R. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging. 2012;30:1323–1341.
57. Kikinis R, Pieper S. 3D Slicer as a tool for interactive brain tumor segmentation. Conf Proc IEEE Eng Med Biol Soc. 2011;2011:6982–6984.
58. Buckler AJ, Bresolin L, Dunnick NR, Sullivan DC, Aerts HJ, Bendriem B, Bendtsen C, Boellaard R, Boone JM, Cole PE, Conklin JJ, Dorfman GS, Douglas PS, Eidsaunet W, Elsinger C, Frank RA, Gatsonis C, Giger ML, Gupta SN, Gustafson D, Hoekstra OS, Jackson EF, Karam L, Kelloff GJ, Kinahan PE, McLennan G, Miller CG, Mozley PD, Muller KE, Patt R, Raunig D, Rosen M, Rupani H, Schwartz LH, Siegel BA, Sorensen AG, Wahl RL, Waterton JC, Wolf W, Zahlmann G, Zimmerman B. Quantitative imaging test approval and biomarker qualification: interrelated but distinct activities. Radiology. 2011;259:875–884.
59. Institute of Medicine (U.S.). Forum on Drug Discovery Development and Translation. Olson S, Robinson S, Giffin RB. Accelerating the Development of Biomarkers for Drug Safety: Workshop Summary. Washington, DC: National Academies Press; 2009.
60. Waterton JC, Pylkkanen L. Qualification of imaging biomarkers for oncology drug development. Eur J Cancer. 2012;48:409–415.

61. Buckler AJ, Mozley PD, Schwartz L, Petrick N, McNitt-Gray M, Fenimore C, O'Donnell K, Hayes W, Kim HJ, Clarke L, Sullivan D. Volumetric CT in lung cancer: an example for the qualification of imaging as a biomarker. Acad Radiol. 2010;17:107–115.
62. Buckler AJ, Schwartz LH, Petrick N, McNitt-Gray M, Zhao B, Fenimore C, Reeves AP, Mozley PD, Avila RS. Data sets for the qualification of volumetric CT as a quantitative imaging biomarker in lung cancer. Opt Express. 2010;18: 15267–15282.
63. Channin DS, Mongkolwat P, Kleper V, Sepukar K, Rubin DL. The caBIG annotation and image Markup project. J Digit Imaging. 2010;23:217–225.
64. DICOM Standards Committee WG 17 (3D). Supplement 111: segmentation storage SOP class. Digit Imaging Commun Med. 2006;17.
65. caBIG In-vivo Imaging Workspace. Annotation and Image Markup (AIM). https://cabig.nci.nih.gov/tools/AIM; Accessed: December 26, 2008.
66. Rubin DL, Mongkolwat P, Kleper V, Supekar K, Channin DS. Medical imaging on the semantic web: annotation and image markup. In: *2008 AAAI Spring Symposium Series, Semantic Scientific Knowledge Integration*. Stanford, CA: Stanford University; 2008.
67. Rubin DL, Mongkolwat P, Channin DS. A semantic image annotation model to enable integrative translational research. Summit Transl Bioinform 2009;2009: 106–110.
68. DICOM Standards Committee - Working Group 8 - Structured Reporting. Digital Imaging and Communications in Medicine (DICOM); Sup 200 - Transformation of NCI Annotation and Image Markup (AIM) and DICOM SR Measurement Templates. Available from: ftp://medical.nema.org/medical/dicom/Supps/LB/sup200_lb_AIM_DICOMSRTID1500.pdf.
69. DICOM Standards Committee. DICOM PS3.21 2017e - Transformations between DICOM and other Representations; A.6 AIM v4 to DICOM TID 1500 Mapping. Available from: http://dicom.nema.org/medical/Dicom/2017e/output/chtml/part21/sect_A.6.html.
70. Hwang KH, Lee H, Koh G, Willrett D, Rubin DL. Building and querying RDF/OWL database of semantically annotated nuclear medicine images. J Digit Imaging. 2017;30:4–10.
71. Moreira DA, Hage C, Luque EF, Willrett D, Rubin DL. 3D markup of radiological images in ePAD, a web-based image annotation tool. In: *2015 IEEE 28th International Symposium on Computer-Based Medical Systems (CBMS)*, 2015;97–102, Available from: http://ieeexplore.ieee.org/ielx7/7164867/7167433/07167466.pdf?tp=&arnumber=7167466&isnumber=7167433.
72. Rubin DL, Willrett D, O'Connor MJ, Hage C, Kurtz C, Moreira DA. Automated tracking of quantitative assessments of tumor burden in clinical trials. Transl Oncol. 2014;7:23–35.
73. Warnock MJ, Toland C, Evans D, Wallace B, Nagy P. Benefits of using the DCM4CHE DICOM archive. J Digit Imaging. 2007;20(*Suppl* 1):125–129.
74. Hoy MB. HTML5: a new standard for the Web. Med Ref Serv Q. 2011;30: 50–55.
75. Levy MA, Rubin DL. Computational approaches to assist in the evaluation of cancer treatment response. Imaging Med. 2011;3:233–246.
76. Mongkolwat P, Channin DS, Rubin VK. Informatics in radiology: an open-source and open-access cancer biomedical informatics grid annotation and image markup template builder. Radiographics. 2012;32:1223–1232.
77. Langlotz CP. RadLex: a new method for indexing online educational materials. Radiographics. 2006;26:1595–1597.
78. Bruno EJ. SOA, Web services, and RESTful systems—a framework for building RESTful systems. Dr Dobbs J. 2007;32:32.
79. Lipton P, Nagy P, Sevinc G. Leveraging Internet technologies with DICOM WADO. J Digit Imaging. 2012;25:646–652.
80. Meier W. eXist: an open source native XML database. Web Web Serv Database Syst. 2003;2593:169–183.
81. Napel SA, Beaulieu CF, Rodriguez C, Cui J, Xu J, Gupta A, Korenblum D, Greenspan H, Ma Y, Rubin DL. Automated retrieval of CT images of liver lesions on the basis of image similarity: method and preliminary results. Radiology. 2010;256:243–252.
82. Abramson RG, Lakomkin N, Hainline A, Kang H, Hutson MS, Arteaga CL. The attenuation distribution across the long axis of breast cancer liver metastases at CT: a quantitative biomarker for predicting overall survival. AJR Am J Roentgenol. 2018; 210:W1–W7.
83. Lakomkin N, Kang H, Landman B, Hutson MS, Abramson RG. The Attenuation Distribution Across the Long Axis (ADLA): preliminary findings for assessing response to cancer treatment. Acad Radiol. 2016;23:718–723.
84. Yankeelov TE, Gore JC. Dynamic contrast enhanced magnetic resonance imaging in oncology: theory, data acquisition, analysis, and examples. Curr Med Imaging Rev. 2009;3:91–107.
85. Schaer R, Cid YD, Alkim E, John S, Rubin DL, Depeursinge A. Web-based tools for exploring the potential of quantitative imaging biomarkers in radiology intensity and texture analysis on the ePAD platform. In *Biomedical Texture Analysis: Fundamentals, Tools and Challenges*; 2017:379–410.
86. Echegaray S, Bakr S, Rubin DL, Napel S. Quantitative Image Feature Engine (QIFE): an open-source, modular engine for 3D quantitative feature extraction from volumetric medical images. J Digit Imaging. 2018;31:403–414.
87. Depeursinge A, Foncubierta-Rodriguez A, Van de Ville D, Muller H. Rotation-covariant texture learning using steerable riesz wavelets. IEEE Trans Image Process. 2014;23:898–908.
88. John S, Rubin DL, Gude D, Echegaray S, Bakr S, Mattonen S, Napel S. QIFP : A web application to support quantitative imaging pipelines on medical images. In: *3rd Annual Scientific Conference on Machine Intelligence in Medical Imaging (C-MIMI) of the Society for Imaging Informatics in Medicine (SIIM)*. San Francisco, CA; 2018.
89. Fornaciari G, Vitiello A, Giusiani S, Giuffra V, Fornaciari A, Villari N. The Medici Project first anthropological and paleopathological results of the exploration of the Medici tombs in Florence. Med Secoli. 2007;19:521–543.
90. Graves EE, Quon A, Loo BW. RT_Image: an open-source tool for investigating PET in radiation oncology. Technol Cancer Res Treat. 2007;6:111–121.
91. Hoogi A, Beaulieu CF, Cunha GM, Heba E, Sirlin CB, Napel S, Rubin DL. Adaptive local window for level set segmentation of CT and MRI liver lesions. Med Image Anal. 2017;37:46–55.
92. Herz C, Fillion-Robin JC, Onken M, Riesmeier J, Lasso A, Pinter C, Fichtinger G, Pieper S, Clunie D, Kikinis R, Fedorov A. dcmqi: an open source library for standardized communication of quantitative image analysis results using DICOM. Cancer Res. 2017;77:E87–E90.
93. Maastro. Dicom File interface [Internet]. [Cited 2018 Sep 29]. Available from: https://bitbucket.org/maastro/dicomfileinterface
94. Rosset A, Spadola L, Ratib O. OsiriX: an open-source software for navigating in multidimensional DICOM images. J Digit Imaging. 2004;17:205–216.
95. Gevaert O, Mitchell LA, Xu J, Yu C, Rubin D, Zaharchuk G, Napel S, Plevritis S. Radiogenomic analysis indicates MR images are potentially predictive of EGFR mutation status in glioblastoma multiforme. In: *AACR 103nd Annual Meeting*. Chicago, IL; 2012.
96. Gevaert O, Xu J, Hoang C, Leung A, Quon A, Rubin DL, Napel S, Plevritis S. Integrating medical images and transcriptomic data in non-small cell lung cancer. In: *AACR 102nd Annual Meeting*. Orlando, FL; 2011.
97. Gevaert O, Xu JJ, Hoang CD, Leung AN, Xu Y, Quon A, Rubin DL, Napel S, Plevritis SK. Non-small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data-methods and preliminary results. Radiology. 2012;264:387–396.
98. Gimenez F, Xu J, Liu Y, Liu TT, Beaulieu C, Rubin DL, Napel S. On the feasibility of predicting radiological observations from computational imaging features of liver lesions in CT scans. In: *First IEEE Conference on Healthcare Informatics, Imaging, and Systems Biology (HISB)*, IEEE Computer Society: San Jose, CA; 2011.
99. Gimenez F, Xu J, Liu TT, Beaulieu C, Rubin DL, Napel S, Liu Y. Prediction of radiologist observations using computational image features: method and preliminary results. In: *Ninety-Seventy Annual Scientific Meeting of the RSNA*. Chicago, IL; 2011.
100. Hoang C, Napel S, Gevaert O, Xu J, Rubin DL, Leung A, Merritt R, Whyte R, Shrager J, Plevritis S. NSCLC gene profiles correlate with specific CT characteristics: image-omics. In: *American Association for Thoracic Surgery (AATS)*. Philadelphia, PA; 2011.
101. Napel S, Hoang C, Xu J, Gevaert O, Rubin DL, Plevritis S, Xu Y, Leung A, Quon A. Computational and semantic annotation of CT and PET images and integration with genomic assays of tumors in non-small cell lung cancer (NSCLC) for decision support and discovery: method and preliminary results. In: *Ninety-Seventy Annual Scientific Meeting of the RSNA*. Chicago, IL; 2011.
102. Plevritis S, Gevaert O, Xu J, Hoang C, Leung A, Xu Y, Quon A, Rubin DL, Napel S. Rapid Identification of Prognostic Imaging Biomarkers for Non-small Cell Lung Carcinoma (NSCLC) by integrating image features and gene expression and leveraging public gene expression databases. In: *Ninety-Seventy Annual Scientific Meeting of the RSNA*. Chicago, IL; 2011.
103. Korenblum D, Rubin D, Napel S, Rodriguez C, Beaulieu C. Managing biomedical image metadata for search and retrieval of similar images. J Digit Imaging. 2011;24:739–748.
104. Levy MA, O'Connor MJ, Rubin DL. Semantic reasoning with image annotations for tumor assessment. AMIA Annu Symp Proc. 2009;2009:359–363.
105. Levy MA, Rubin DL. Tool support to enable evaluation of the clinical response to treatment. AMIA Annu Symp Proc. 2008:399–403.
106. Echegaray S, Gevaert O, Shah R, Kamaya A, Louie J, Kothary N, Napel S. Core samples for radiomics features that are insensitive to tumor segmentation: method and pilot study using CT images of hepatocellular carcinoma. J Med Imaging. 2015;2:041011.
107. Echegaray S, Nair V, Kadoch M, Leung A, Rubin D, Gevaert O, Napel S. A Rapid segmentation-insensitive "digital biopsy" method for radiomic feature extraction: method and pilot study using CT images of non-small cell lung cancer. Tomography. 2016;2:283–294.

108. Zhou M, Leung A, Echegaray S, Gentles A, Shrager JB, Jensen KC, Berry GJ, Plevritis SK, Rubin DL, Napel S, Gevaert O. Non-small cell lung cancer radiogenomics map identifies relationships between molecular and imaging phenotypes with prognostic implications. Radiology. 2018;286:307–315.
109. Bakr S, Echegaray S, Shah R, Kamaya A, Louie J, Napel S, Kothary N, Gevaert O. Noninvasive radiomics signature based on quantitative analysis of computed tomography images as a surrogate for microvascular invasion in hepatocellular carcinoma: a pilot study. J Med Imaging (Bellingham). 2017; 4:041303.
110. Itakura H, Achrol AS, Mitchell LA, Loya JJ, Liu T, Westbroek EM, Feroze AH, Rodriguez S, Echegaray S, Azad TD, Yeom KW, Napel S, Rubin DL, Chang SD, Harsh GRT, Gevaert O. Magnetic resonance image features identify glioblastoma phenotypic subtypes with distinct molecular pathway activities. Sci Transl Med 2015;7:303ra138.
111. Rubin DL, Willrett D, O'Connor MJ, Hage C, Kurtz C, Moreira DA. Automated tracking of quantitative assessments of tumor burden in clinical trials. Transl Oncol. 2014;7:23–25.
112. Jarrett AM, Hormuth DA, Barnes SL, Feng X, Huang W, Yankeelov TE. Incorporating drug delivery into an imaging-driven, mechanics-coupled reaction diffusion model for predicting the response of breast cancer to neoadjuvant chemotherapy: theory and preliminary clinical results. Phys Med Biol. 2018;63: 105015.
113. Yip SS, Kim J, Coroller TP, Parmar C, Velazquez ER, Huynh E, Mak RH, Aerts HJ. Associations between somatic mutations and metabolic imaging phenotypes in non-small cell lung cancer. J Nucl Med. 2017;58:569–576.
114. American College of Radiology. DART Portal. https://dart.acr.org/https://dart.acr.org/.
115. Strosberg JR, Halfdanarson TR, Bellizzi AM, Chan JA, Dillon JS, Heaney AP, Kunz PL, O'Dorisio TM, Salem R, Segelov E, Howe JR, Pommier RF, Brendtro K, Bashir MA, Singh S, Soulen MC, Tang L, Zacks JS, Yao JC, Bergsland EK. The North American Neuroendocrine Tumor Society Consensus guidelines for surveillance and medical management of midgut neuroendocrine tumors. Pancreas. 2017;46:707–714.
116. Pieper S, Halle M, Kikinis R. 3D SLICER. IEEE International Symposium on Biomedical Imaging ISBI 2004, 2004:632–635.
117. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat Methods. 2012;9:671–675.
118. McAuliffe M. Using MIPAV to label and measure brain components in Talairach space. Tech report. National Institutes of Health. [cited 2018 Sep 15]. Available from: https://mipav.cit.nih.gov/index.php.

TOMOGRAPHY® RESEARCH ARTICLE

The Brain Imaging Collaboration Suite (BrICS): A Cloud Platform for Integrating Whole-Brain Spectroscopic MRI into the Radiation Therapy Planning Workflow

Saumya Gurbani[1,2], Brent Weinberg[3], Lee Cooper[2,4], Eric Mellon[5], Eduard Schreibmann[1], Sulaiman Sheriff[6], Andrew Maudsley[6], Mohammed Goryawala[6], Hui-Kuo Shu[1], and Hyunsuk Shim[1,2,3]

Departments of [1]Radiation Oncology, [2]Biomedical Engineering, [3]Radiology and Imaging Sciences, and [4]Biomedical Informatics, Emory University, Atlanta, GA; Departments of [5]Radiation Oncology and [6]Radiology, University of Miami Miller School of Medicine, Miami, FL

Corresponding Author:
Hyunsuk Shim, PhD
Department of Radiology and Imaging Sciences, Emory University, Atlanta, GA, USA;
E-mail: hshim@emory.edu

Key Words: spectroscopic MRI, radiation therapy, cloud platform

Abbreviations: Radiation therapy (RT), magnetic resonance imaging (MRI), spectroscopic MRI (sMRI), fluid-attenuation inversion recovery (FLAIR), magnetic resonance spectroscopy (MRS), 3-dimensional (3D), N-acetylaspartate (NAA), normal-appearing white matter (NAWM), Digital Imaging and Communication in Medicine (DICOM), echo planar spectroscopic imaging (EPSI)

ABSTRACT

Glioblastoma has poor prognosis with inevitable local recurrence despite aggressive treatment with surgery and chemoradiation. Radiation therapy (RT) is typically guided by contrast-enhanced T1-weighted magnetic resonance imaging (MRI) for defining the high-dose target and T2-weighted fluid-attenuation inversion recovery MRI for defining the moderate-dose target. There is an urgent need for improved imaging methods to better delineate tumors for focal RT. Spectroscopic MRI (sMRI) is a quantitative imaging technique that enables whole-brain analysis of endogenous metabolite levels, such as the ratio of choline-to-N-acetylaspartate. Previous work has shown that choline-to-N-acetylaspartate ratio accurately identifies tissue with high tumor burden beyond what is seen on standard imaging and can predict regions of metabolic abnormality that are at high risk for recurrence. To facilitate efficient clinical implementation of sMRI for RT planning, we developed the Brain Imaging Collaboration Suite (BrICS; https://brainimaging.emory.edu/brics-demo), a cloud platform that integrates sMRI with standard imaging and enables team members from multiple departments and institutions to work together in delineating RT targets. BrICS is being used in a multisite pilot study to assess feasibility and safety of dose-escalated RT based on metabolic abnormalities in patients with glioblastoma (Clinicaltrials.gov NCT03137888). The workflow of analyzing sMRI volumes and preparing RT plans is described. The pipeline achieved rapid turnaround time by enabling team members to perform their delegated tasks independently in BrICS when their clinical schedules allowed. To date, 18 patients have been treated using targets created in BrICS and no severe toxicities have been observed.

INTRODUCTION

The standard-of-care treatment for glioblastoma, the most common adult primary malignant brain tumor, consists of maximal safe surgical resection of tumor followed by high-dose radiation therapy (RT) with concomitant temozolomide chemotherapy (1-4). The standard high-dose prescription of 60 Gy is delivered over 30 fractions to regions of enhancement on T1-weighted contrast-enhanced (CE-T1w) MRI, in which enhancement represents areas of tumor with leaky neovasculature. A lower dose of RT (typically 46–54 Gy) is delivered to areas of hyperintensity on T2-weighted fluid-attenuation inversion recovery (FLAIR) MRI (5). FLAIR hyperintensity corresponds to a nonspecific combination of tumor and nontumor pathologies, including inflammation and vasogenic edema (6). Despite improvements in maximal resection, concurrent and adjuvant chemotherapy, and RT, the median overall survival still remains poor at 15 months (7, 8), with median progression-free survival at only 4–6 months (9). Recurrent glioblastoma is very difficult to treat, often being resistant to further radiation and inaccessible for secondary surgical resection (10). The location of recurrent disease can also vary: within the original 60-Gy RT target, within the intermediate dose area, or to regions several centimeters away, including crossing the midline (11). Both local and distant recurrences need to be addressed to improve progression-free survival. In a phase II study where glioblastomas were treated with high-dose proton therapy up to 90 cobalt-gray equivalent,

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2018.00028

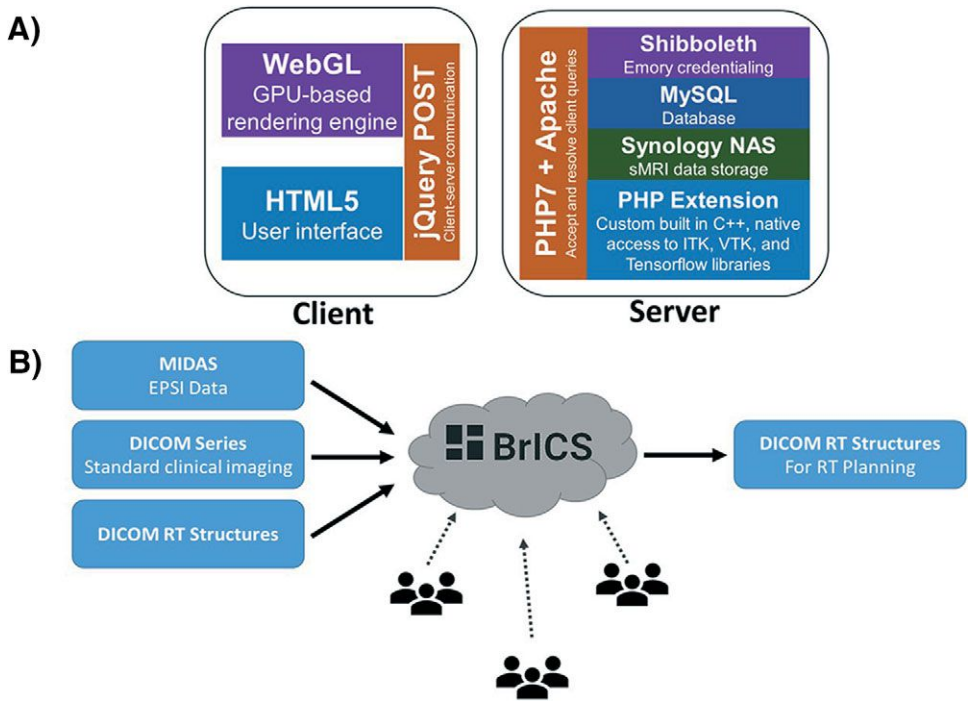


Figure 1. A cloud platform for spectroscopic magnetic resonance imaging (sMRI)-guided radiation therapy (RT). The Brain Imaging Collaboration Suite (BrICS) consists of a centralized server which performs image processing, and a lightweight browser client (A). BrICS imports spectroscopy and Digital Imaging and Communication in Medicine (DICOM)-format MRI volumes, and can export RT targets to other clinical software (B). sMRI volumes are blended with clinical MRI, and users can perform tasks such as evaluating underlying spectra and contouring based on sMRI abnormalities.

it was observed that almost all disease recurred in regions receiving <70 cobalt-gray equivalent (12). Thus, it appears that dose escalation may provide sufficient tumoricidal doses to achieve local control. However, doses >70 Gy need to be applied selectively to prevent toxicity that could result from excess volumes of normal brain receiving doses of that magnitude.

Spectroscopic magnetic resonance imaging (sMRI) is an evolution of magnetic resonance (MR) spectroscopy (MRS) that enables 3-dimensional (3D) whole-brain volumes of metabolite levels to be obtained in vivo without contrast agents or radioactive tracers (13, 14). Two metabolites of particular interest in patients with glioblastoma include choline-containing compounds (Cho), the building blocks of the cell membrane that increase in proliferating tumor cells, and N-acetylaspartate (NAA), a biomarker found in healthy neurons, which diminishes owing to neuronal displacement and death from glial infiltration (13, 15). It has been previously shown via histological correlation that the ratio of Cho to NAA is significantly elevated in glioblastoma owing to the opposing changes in these metabolites; in particular, a two-fold increase in Cho/NAA compared to healthy tissue in contralateral normal-appearing white matter (NAWM) was able to correctly identify tumor in 100% of cases, even when tissue samples were biopsied from regions outside of contrast-enhancement per CE-T1w or FLAIR hyperintensity (16).

A combination of dose escalation guided by sMRI, including regions of occult tumor normally left untreated by high-dose RT, could potentially delay recurrence of disease by delivering a cytotoxic dose of radiation to regions of metabolically abnormal tumor even if these areas are not detected using standard imaging techniques. However, the use of sMRI in clinical practice has been hampered by data processing requirements and limited integration into the RT planning workflow. In previous studies, several time-intensive manual processing steps were required to import metabolite volumes into clinical imaging software so that they could be used in the operating room or for RT planning (17, 18). To enable integration of sMRI into clinical practice, we have developed a software platform designed specifically for the integration of sMRI into the RT planning workflow. In this paper, we describe its architecture and show its features on several sample cases. We show feasibility of this software for collaborative use in a prospective multi-institutional clinical study to target dose-escalated RT based on sMRI. Several challenges in integrating this imaging modality into the clinical workflow are addressed, and a sample case from the ongoing study is presented to show that RT to high-risk regions can be targeted by quantitative imaging techniques such as sMRI.

MATERIALS AND METHODS

Software Architecture

To assist with a collaborative clinical study across institutions, we developed the Brain Imaging Collaboration Suite (BrICS), a web-based software designed specifically to integrate sMRI with clinical MRI volumes, enabling physicians to evaluate relevant metabolite levels and the underlying spectra used for this quantitation, and to delineate target volumes for RT planning based on this information (19). BrICS consists of 2 components: a centralized server and a lightweight browser client (Figure 1A). The server performs computations necessary to analyze and display whole-brain spectroscopy; it consists of modules written in C++ and the PHP server-side scripting language to take advantage of well-established image processing and linear algebra libraries (20, 21). The lightweight browser client written in JavaScript can run on all modern hardware, including thin clients such as laptops and tablets. This browser-based approach offers the following benefits over standalone software clients: (1) improves repeatability and standardization by ensuring data

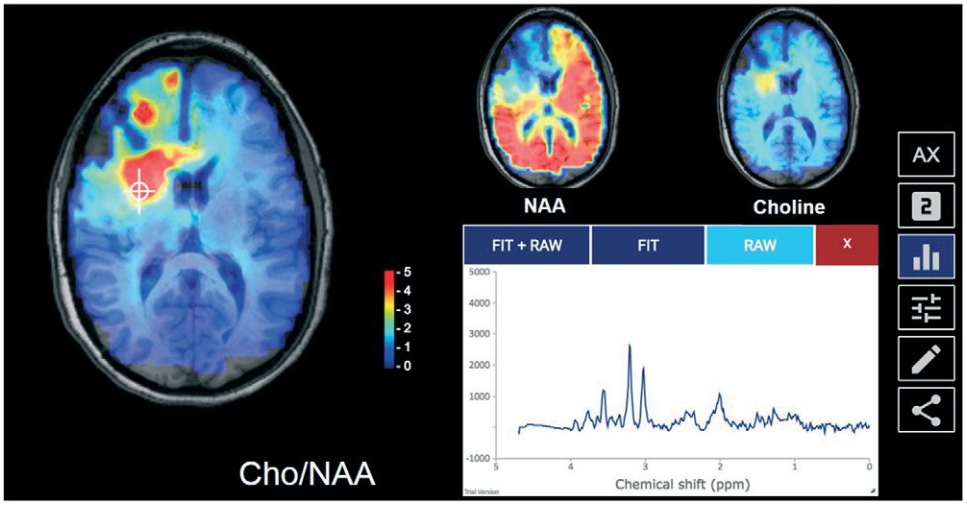


Figure 2. The main user interface for BrICS. sMRI metabolite and metabolite ratio maps are overlaid on top of anatomic magnetic resonance (MR) volumes. Selection of a given voxel brings up the underlying spectrum.

are processed on the same hardware; (2) reduces user variability and bias; (3) enables real-time deployment of software updates across all clients; (4) prevents the need for every end-user to download massive sMRI data sets onto a local computer; (5) runs without the need for the user to download any software beyond a web browser, which is of key importance, as physicians often use restricted hospital workstations; and (6) allows information and images to be easily shared with patients who wish to be better informed of their clinical management.

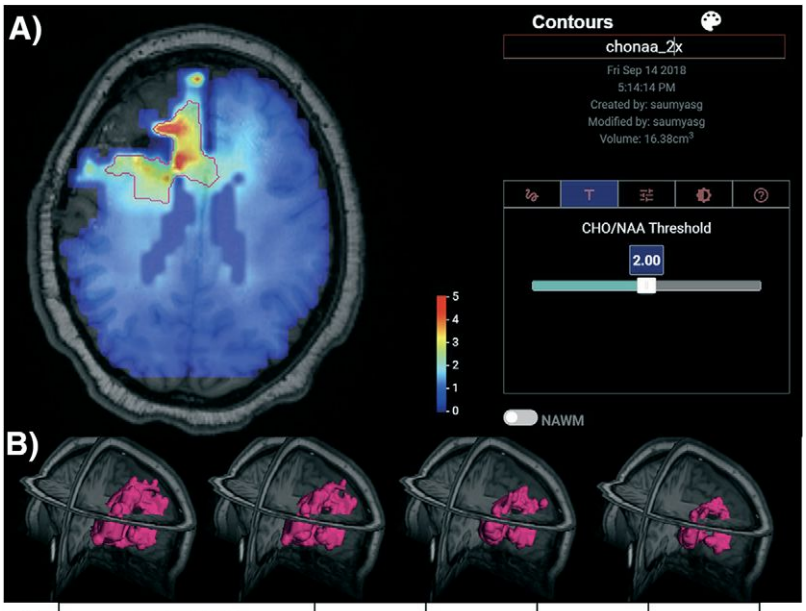
BrICS imports data from spectroscopy processing software, such as the Metabolite Imaging and Data Analysis Software (MIDAS, University of Miami, Miami, FL), and from other imaging systems/software using the Digital Imaging and Communication in Medicine (DICOM) file format. All volumes are coregistered using a rigid transformation and resampled using trilinear interpolation into a high-resolution T1w image space, enabling overlays of metabolic information onto anatomic MRI. Users can then delineate target volumes based on both anatomic and spectroscopic information. These targets can be exported as DICOM RT structure sets (DICOM RT) or binary DICOM masks and imported into RT planning systems to deliver therapy to patients (Figure 1B). A video showing the features of BrICS is available in online Supplemental Video 1 PLAY VIDEO, which are described in detail in the following subsections.

Visualization and Contouring

The main interface of BrICS is shown in Figure 2. sMRI volumes—either individual metabolites or metabolite ratios—are overlaid on anatomic volumes (eg, T1w MRI), enabling visual assessment of metabolic changes in spatially dependent manner. For MR spectroscopists and radiologists familiar with MRS techniques, selection of a voxel will bring up the corresponding spectrum. Because sMRI is a quantitative imaging technique, voxel intensities can be reliably interpreted across subjects, and decision-making can be based on specified thresholds. This ability is built-in to the contouring module; physicians can make contours based on the values in sMRI maps (Figure 3 and online Supplemental Video 1 PLAY VIDEO). For example, the Cho/NAA volume abnormality index (16) can be used, as shown, to generate a contour around all voxels which have a Cho/NAA abnormality index above a given threshold. Users can select this threshold and automatically generate contours of increasing or decreasing sensitivity of disease detection. Radiologists can then review these contours and make changes to them using built-in editing tools (painting, erasing, or selection of connected-components). Once contours are generated, they can be visualized as 3D volumes, enabling visual quality assessment and correspondence with anatomy. Statistics such as contour volume and number of connected components are also reported.

Normalization of Metabolite Values

Cerebral concentrations of several macromolecules, including Cho and NAA, are known to vary based on a subject's age, gender, and anatomic location of brain being sampled (22). To account for these variations in baseline metabolism, metrics such as the Cho/NAA abnormality index (16) and the Cho-NAA index (23) take into account relative changes in these metabolites compared to normal tissue, typically contralateral NAWM (24). For this trial, we use the Cho/NAA abnormality index, defined as the Cho/NAA of a given voxel divided by the mean Cho/NAA value in contralateral NAWM. In previous works (16-18), NAWM was manually contoured on a clinical T1w volume by a neuroradiologist using commercial software, then the mask exported and applied to sMRI data to determine the mean Cho/NAA value. To expedite this process, remove reliance on commercial software, and mitigate user bias, we have implemented an algorithm in BrICS to automatically contour the NAWM based on a Gaussian mixture model (25) (Figure 4). First, all voxels from the cerebrum are masked using an anatomic atlas. Next, all cerebral Cho/NAA voxels are modeled as a bimodal Gaussian distribution, with voxels arising from the second, higher-mean Gaussian population representative of tumor pathology. These voxels are then masked, and the side with largest contiguous abnormal segment is selected as the side of tumor; voxels in the contralateral hemisphere are segmented into gray and white matter based on fractional water content (26) calculated by MIDAS, and then the mean is reported as the normalizing factor for the subject's abnormality index calculations.



C)

Cho/NAA Abnormality Index	1.75X	2.0X	2.5X	3.0X
Contour Volume (cc)	63.98	48.11	29.19	16.38

Figure 3. Contouring of target volumes. The contouring module enables identification of target volumes based on either anatomic or metabolite images (A). For quantitative imaging techniques like sMRI, users can automatically delineate contours using threshold values. A series of targets based on thresholding of the Cho/NAA abnormality index; target volumes can be rendered in 3D for visual inspection prior to being exported to other clinical software (B). A summary of the volumes generated for varying Cho/NAA abnormality indices (C).

Automated Segmentation of Residual Contrast Enhancement

Additional algorithmic modules can be built into BrICS to assist with other routines that are regularly performed by clinicians. One such module automatically contours residual contrast enhancing tissue (Figure 5), so as to differentiate true unresected tumor with leaky neovasculature from blood products owing to surgical resection (27). The module requires a precontrast T1w MRI, a CE-T1w MRI, and a T2w or FLAIR MRI, all of which are coregistered into the same imaging space and resampled to an axial view. The pre- and postcontrast MR images are histogram normalized and subtracted to generate a difference map; Otsu thresholding with four classes is used to identify residual enhancement (28, 29). Otsu thresholding is applied to the FLAIR map to automatically identify hyperintensity; the single largest connected component is used as a bounding mask for the T1w

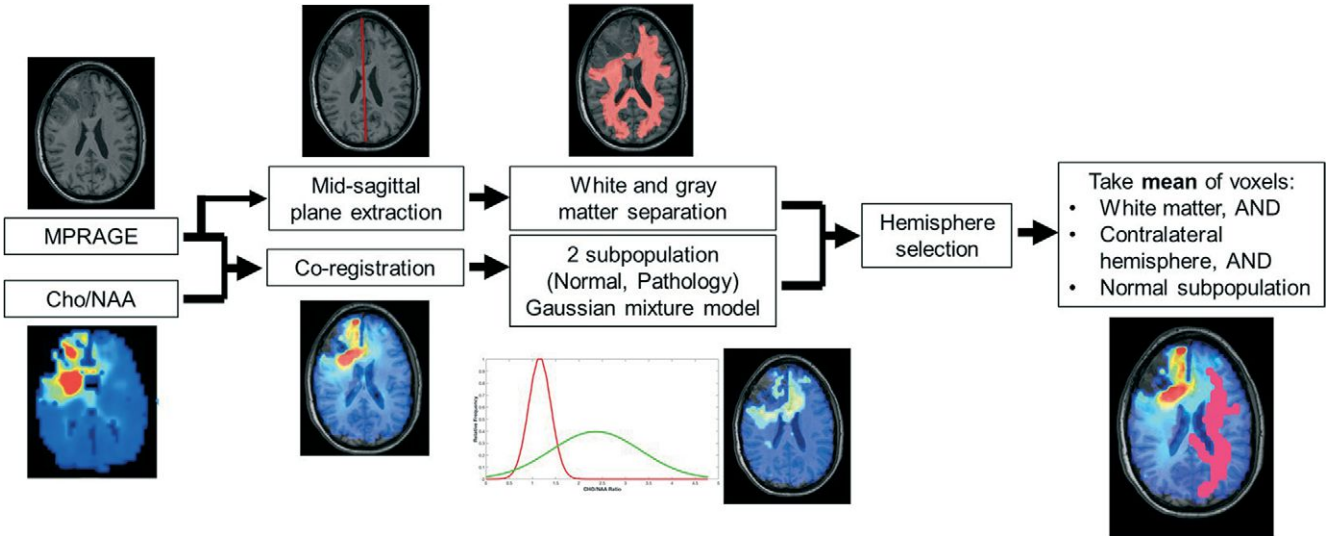


Figure 4. Normalization of metabolite maps by baseline metabolism. High-level schematic of a Gaussian mixture model used to identify regions of normal-appearing white matter (NAWM), which is used as a personal metabolic baseline for the patient. NAWM is typically contoured manually by radiologists; this algorithm can perform the same contouring automatically in just a few seconds.

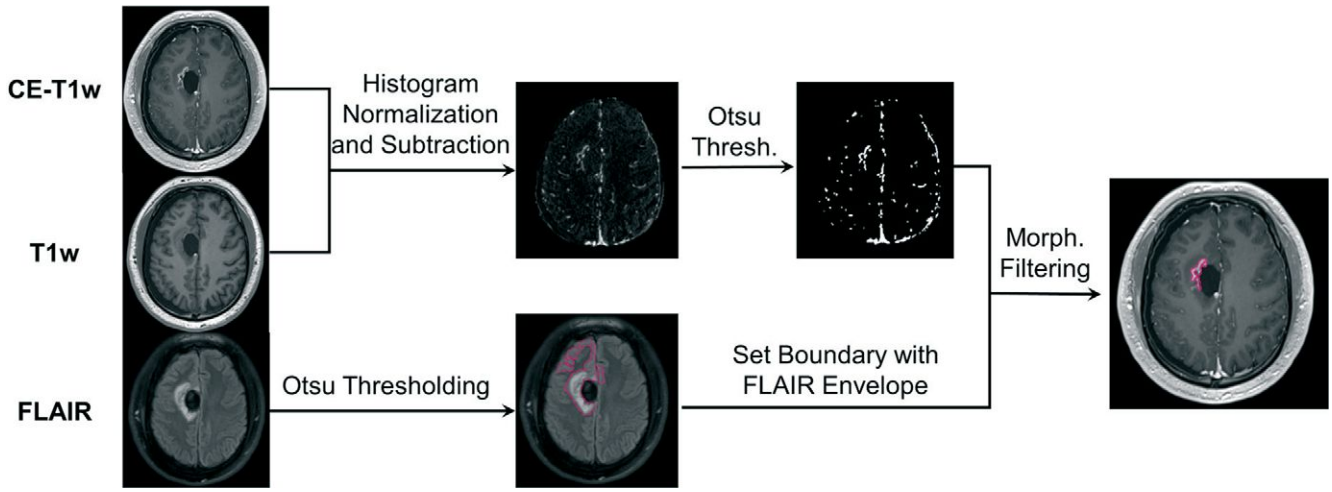


Figure 5. Automated residual contrast enhancement contouring. BrICS takes a postcontrast T1-weighted (T1w) MRI (top), precontrast T1w MRI (middle), and a FLAIR MRI (bottom) volume, and follows the shown algorithm to rapidly contour residual contrast enhancement postsurgical resection. This volume can then be edited manually by the neuroradiologist or radiation oncologist as desired to define a dose-escalated volume.

residual volume. Finally, morphological opening and closing filters are applied to the bounded T1w residual volume to remove islets and thin anisotropic components, e.g. blood vessels. The entire algorithm can be run in <10 seconds on the BrICS server and yields a final contour, which can be evaluated and manually edited, if necessary, by a neuroradiologist—saving valuable clinician time and providing a reproducible starting point for all users.

Patient Enrollment and Imaging

To assess the feasibility and safety of sMRI-guided RT, a multisite clinical study funded by NCI was initiated (Clinicaltrials.gov NCT03137888). Three institutions are participating in this pilot study—Emory University, the Johns Hopkins University, and the University of Miami—and a total of 30 patients with newly diagnosed glioblastoma will be enrolled. Patients are enrolled after undergoing maximal safe surgical resection or biopsy at the discretion of the neurosurgeon. Enrolled patients were ≥18 years of age, had a Karnofsky Performance Score ≥ 60, and were willing to undergo dose-escalated RT to 75 Gy.

An sMRI scan was obtained within 2 weeks prior to starting RT + temozolomide. A 15-minute echo planar spectroscopic imaging (EPSI) pulse sequence combined with GRAPPA [parallel imaging (30)], was performed on a 3 T scanner (Siemens Medical Solutions, Erlangen, Germany) with a 32-channel or a 20-channel head coil array (echo time = 50 milliseconds, repetition time = 1551 milliseconds, flip angle = 71°). During the same session, a high-resolution T1w magnetization prepared rapid acquisition gradient echo (MP-RAGE) sequence was obtained at the same orientation and position as the EPSI. Raw EPSI data were transformed into spatial-spectral data, coregistered with the MP-RAGE volume, and the relative metabolite concentration values were obtained by spectral fitting using MIDAS (22, 30).

RT Planning

An outline of the workflow for patients in this study is shown in Figure 6. The EPSI/GRAPPA and MP-RAGE volumes, in addition to the most recent clinical CE-T1w and FLAIR MRI, were imported into BrICS. Automated contours for Cho/NAA abnormality index of 2.0 and residual contrast-enhancing tissue were generated using the algorithms described above. Using BrICS, 2 MR spectroscopists from different institutions collaboratively reviewed the underlying raw and fitted spectra within the Cho/NAA abnormal contour and removed voxels with poor spectral quality. Meanwhile, a neuroradiologist reviewed and edited the residual contrast-enhancing volume to ensure accurate delineation of the target volume. The 2 contours were then merged to form a single target volume for high-dose RT. Next, an external radiation oncologist (from a nontreating site) edited and approved the volume based on anatomy and dose safety concerns. Finally, the treating-site radiation oncologist made final edits based on his/her discretion and validated the volume for RT treatment. To ensure patient safety and to enable retrospective review of this study, all user edits were tracked in BrICS in a digital audit trail.

The final contour generated in BrICS was defined as gross tumor volume 3 (GTV3). The clinical target volume 3 (CTV3) was defined as equal to GTV3 with no margin. In this pilot feasibility study, a maximum volume of 65 cm^3 was allowed for CTV3, approximately adhering to the 5-cm-diameter boost volume limit used in the NRG Oncology BN001 phase II trial on RT dose escalation for glioblastomas (31). The CTV3 contour was exported from BrICS as a DICOM RT structure set on the high resolution T1w MP-RAGE volume into additional contouring or treatment planning software such as VelocityAI (Varian Medical Systems, Palo Alto, CA), MIM Maestro (MIM Software Inc, Cleveland, OH), Eclipse (Varian Medical Systems, Palo Alto, CA),

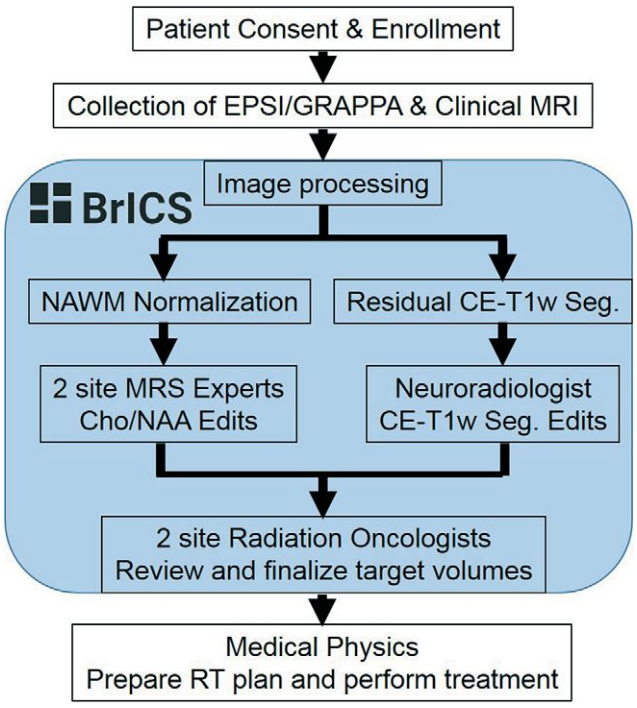


Figure 6. RT planning workflow. After patients are enrolled and consented, their imaging data are processed and edited in BrICS. The centralized platform enables reliable and repeatable processing, with documented edits made by physicians and spectroscopists to prepare the final treatment plan.

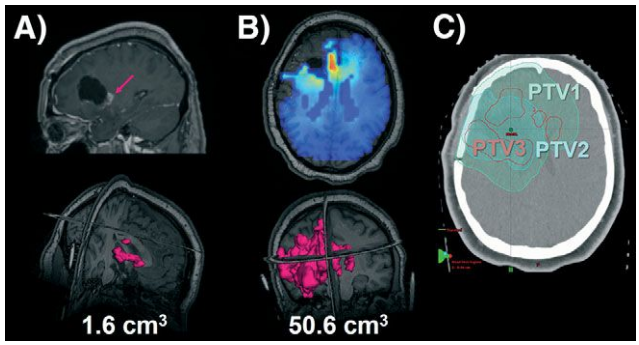


Figure 7. Example treatment plan for study patient. The patient is a 21-year-old woman with newly diagnosed glioblastoma with a near-total resection of the tumor (A). However, the Cho/NAA map indicates metabolically active tumor expanding outward from the resection cavity (B). A boosted dose of 75 Gy (PTV3) was successfully planned and delivered to this patient (C).

Pinnacle (Philips Healthcare, Best, Netherlands), etc., per the routine of the treating site. Additional standard treatment volumes were generated including GTV2, defined as the surgical cavity with residual contrast enhancement, and GTV1, defined as hyperintensity on FLAIR MRI. Five millimeter of anatomically constrained margins were added to GTV2 and GTV1 to produce CTV2 and CTV1, respectively. A 3-mm margin was added to all 3 CTVs to produce the planning target volumes (PTV3, PTV2, and PTV1). A simultaneous in-field boost IMRT plan was generated to treat PTV3, PTV2, and PTV1 to 75 Gy, 60 Gy, and 50.1 Gy, respectively, respecting standard organs-at-risk constraints (Table 1).

RESULTS

A demo of BrICS is available at https://brainimaging.emory.edu/brics-demo with a few curated and deidentified data sets. A video describing the platform and its features is presented in the online Supplemental Video 1 PLAY VIDEO. In addition to the dose-escalated RT study described above, BrICS is currently being used for the following clinical projects internally at Emory University: targeting of biopsies in patients with nonenhancing low-grade gliomas, monitoring therapeutic response of patients with glioblastoma receiving a histone-deacetylase inhibitor in addition to standard chemoradiation, identification of metabolite abnormalities associated with melanoma brain metastasis, and a pilot study evaluating the benefit of sMRI for patients with mild traumatic brain injury. In addition, BrICS served as the platform for testing new imaging processing algorithms such as a neural network for identifying spectral artifacts (32) and autoencoder-based spectral fitting (unpublished data).

RT plans from 1 patient who underwent dose escalation as per this study's protocol are presented in Figure 7. The patient is a 21-year-old woman diagnosed with a frontal glioblastoma and enrolled in the trial 1 month after undergoing surgical

Table 1. Summary of Target Volume Definitions and Dose Prescription for This Clinical Study

Target Name	Definition	CTV Margin (mm)	PTV Margin (mm)	Dose (Gy)
GTV3	Cho/NAA abnormality index ≥ 2 + residual contrast enhancement	0	3	75
GTV2	Contrast enhancing tissue + resection cavity, per standard of care	5	3	60
GTV1	FLAIR hyperintensity, per standard of care	5	3	50.1

In addition to standard chemoradiation (GTV1 and GTV2), a boost is given to areas of sMRI abnormality and residual contrast enhancement (GTV3). All doses are delivered over 30 fractions.

resection of her tumor. sMRI volumes were obtained and processed in MIDAS and in BrICS per the protocol. The neuroradiologists, MR spectroscopists, and radiation oncologists accessed BrICS remotely for several minutes each, when time was available during their busy schedules. Segmentation of residual contrast enhancement by the automated algorithm, followed by neuroradiologist review, identified a 1.6-cm^3 nodular residual contrast-enhancing lesion on the posterior border of the surgical cavity remaining after surgery, typical of patients who underwent near total resections (Figure 7A). However, the Cho/NAA abnormality was much greater with a volume of 50.6 cm^3, expanding laterally, anteriorly, and posteriorly from the surgical cavity (Figure 7B). GTV3 was planned on the union of these 2 contours, and targeted for a 75-Gy boost. The contour for GTV3 was exported as a DICOM RT structure and imported into Eclipse for dose planning (Figure 7C). Dose constraints based on RTOG guidelines to all organs at risk were met, with >95% of the prescribed dose delivered to each PTV.

DISCUSSION

Current treatments for glioblastoma are insufficient in achieving local control. This is felt to be due in part to limitations of standard imaging methods in identification of infiltrating tumor margins, which show no contrast enhancement, potentially leaving these high-risk regions undertreated. Improvements in treatment options, such as with higher radiation doses, can only be beneficial if all high-risk tumor regions (both enhancing and nonenhancing) are properly targeted. In this work, we develop a software platform that successfully enables sMRI integration into the RT planning workflow. The EPSI/GRAPPA sequence can be used on standard 3 T instrumentation, and the current version of the sequence is available for several different Siemens models (eg, PRISMA, Trio, and Skyra); expansion to other vendors is an ongoing project. The data can then be sent to a centralized server for processing. Because it is web-based, BrICS can be used by multiple users and institutions without the need for additional software, data, or processing. BrICS was successfully used as the infrastructure for an ongoing multi-institutional clinical study assessing the feasibility of dose-escalated radiation guided by sMRI in patients with glioblastoma; to date, 18 patients have been treated on this protocol, and no toxicities have been observed. Thus, there is an urgent need for improved quantitative imaging biomarkers that can not only identify these regions but also be readily incorporated into clinical practice.

sMRI has been shown to delineate infiltrating tumor beyond standard MRI but has thus far been used in only retrospective analyses owing to the complexity of integrating it with clinical volumes, the requirement for an on-site MR spectroscopist to manually review spectra in individual voxel, and variability in acquisition and processing across institutions. A web platform such as BrICS provides solutions for these challenges by enabling centralized data storage and analysis, allowing clinicians from multiple institutions to use sMRI without the need for local experts or software. Users will always have the latest version of BrICS without needing to download additional software or data sets and can access BrICS from any computer browser. BrICS is currently being used in a multisite clinical study assessing the feasibility of sMRI guidance for RT and will continue to be developed as infrastructure for future consortium-level trials.

Supplemental Materials

Supplemental Video 1: http://dx.doi.org/10.18383/j.tom.2018.00028.vid.01

ACKNOWLEDGMENTS

We would like to thank the following people for their assistance in collecting and analyzing data: Dr. Peter Barker, Dr. Michal Povazan, Dr. Sarah Dupont, Mr. Michael Larche, Mr. Robert Smith, Ms. Samira Yeboah, and Ms. Sarah Basadre. This work is funded by NIH grants U01CA172027, R01CA214557, U01EB028145, and F30CA206291.

Disclosures: No disclosures to report.

Conflict of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. Wen PY, Macdonald DR, Reardon DA, Cloughesy TF, Sorensen AG, Galanis E, Degroot J, Wick W, Gilbert MR, Lassman AB, Tsien C, Mikkelsen T, Wong ET, Chamberlain MC, Stupp R, Lamborn KR, Vogelbaum MA, van den Bent MJ, Chang SM. Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. J Clin Oncol. 2010;28:1963–1972.
2. Stupp R, Hegi ME, Gilbert MR. Chemoradiotherapy in malignant glioma: standard of care and future directions. J Clin Oncol. 2007;25:4127–4136.
3. Stupp R, Hegi ME, Mason WP, van den Bent MJ, Taphoorn MJ, Janzer RC, Ludwin SK, Allgeier A, Fisher B, Belanger K, Hau P, Brandes AA, Gijtenbeek J, Marosi C, Vecht CJ, Mokhtari K, Wesseling P, Villa S, Eisenhauer E, Gorlia T, Weller M, Lacombe D, Cairncross JG, Mirimanoff RO; European Organisation for Research and Treatment of Cancer Brain Tumour and Radiation Oncology Groups; National Cancer Institute of Canada Clinical Trials Group. Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. Lancet Oncol. 2009;10:459–466.
4. Stupp R, Mason WP, van den Bent MJ, Weller M, Fisher B, Taphoorn MJ, Belanger K, Brandes AA, Marosi C, Bogdahn U, Curschmann J, Janzer RC, Ludwin SK, Gorlia T, Allgeier A, Lacombe D, Cairncross JG, Eisenhauer E, Mirimanoff RO; European Organisation for Research and Treatment of Cancer Brain Tumor and Radiotherapy Groups; National Cancer Institute of Canada Clinical Trials Group. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. N Engl J Med. 2005;352:987–996.
5. Wernicke AG, Smith AW, Taube S, Mehta MP. Glioblastoma: radiation treatment margins, how small is large enough? Pract Radiat Oncol. 2016;6:298–305.
6. Tsuchiya K, Mizutani Y. Preliminary evaluation of fluid-attenuated inversion-recovery MR in the diagnosis of intracranial tumors. Am J Neuroradiol. 1996;17:1081–1086.
7. Stupp R, Hegi ME, Gilbert MR, Chakravarti A. Chemoradiotherapy in malignant glioma: standard of care and future directions. J Clin Oncol. 2007;25:4127–4136.
8. Stupp R, Hegi ME, Mason WP, van den Bent MJ, Taphoorn MJB, Janzer RC, Ludwin SK, Allgeier A, Fisher B, Belanger K, Hau P, Brandes AA, Gijtenbeek J, Marosi C, Vecht CJ, Mokhtari K, Wesseling P, Villa S, Eisenhauer E, Gorlia T, Weller M, Lacombe D, Cairncross JG, Mirimanoff RO; European Organisation for Research and Treatment of Cancer Brain Tumour and Radiation Oncology Groups; National Cancer Institute of Canada Clinical Trials Group. Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy

alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. Lancet Oncol. 2009;10:459–466.
9. Stupp R, Taillibert S, Kanner AA, Kesari S, Steinberg DM, Toms SA, Taylor LP, Lieberman F, Silvani A, Fink KL, Barnett GH, Zhu JJ, Henson JW, Engelhard HH, Chen TC, Tran DD, Sroubek J, Tran ND, Hottinger AF, Landolfi J, Desai R, Caroli M, Kew Y, Honnorat J, Idbaih A, Kirson ED, Weinberg U, Palti Y, Hegi ME, Ram Z. Maintenance therapy with tumor-treating fields plus temozolomide vs temozolomide alone for glioblastoma: a randomized clinical trial. JAMA. 2015;314: 2535–2543.
10. Weller M, Cloughesy T, Perry JR, Wick W. Standards of care for treatment of recurrent glioblastoma—are we there yet? Neuro Oncol. 2013;15:4–27.
11. Campos B, Olsen LR, Urup T, Poulsen HS. A comprehensive profile of recurrent glioblastoma. Oncogene. 2016;35:5819–5825.
12. Fitzek MM, Thornton AF, Rabinov JD, Lev MH, Pardo FS, Munzenrider JE, Okunieff P, Bussière M, Braun I, Hochberg FH, Hedley-Whyte ET, Liebsch NJ, Harsh GR. Accelerated fractionated proton/photon irradiation to 90 cobalt gray equivalent for glioblastoma multiforme: results of a phase II prospective trial. J Neurosurg. 1999;91:251–260.
13. Law M. MR spectroscopy of brain tumors. Top Magn Reson Imaging. 2004;15: 291–313.
14. Maudsley AA, Domenig C, Sheriff S. Reproducibility of serial whole-brain MR Spectroscopic Imaging. NMR Biomed. 2010;23:251–256.
15. Law M, Cha S, Knopp EA, Johnson G, Arnett J, Litt AW. High-grade gliomas and solitary metastases: differentiation by using perfusion and proton spectroscopic MR imaging. Radiology. 2002;222:715–721.
16. Cordova JS, Shu H-KG, Liang Z, Gurbani SS, Cooper LAD, Holder CA, Olson JJ, Kairdolf B, Schreibmann E, Neill SG, Hadjipanayis CG, Shim H. Whole-brain spectroscopic MRI biomarkers identify infiltrating margins in glioblastoma patients. Neuro Oncol. 2016;18:1180–1189.
17. Cordova JS, Gurbani SS, Olson JJ, Liang Z, Cooper LAD, Shu H-KG, Schreibmann E, Neill SG, Hadjipanayis CG, Holder CA, Shim H. A systematic pipeline for the objective comparison of whole-brain spectroscopic MRI with histology in biopsy specimens from grade III glioma. Tomography. 2016;2:106–116.
18. Cordova JS, Kandula S, Gurbani S, Zhong J, Tejani M, Kayode O, Patel K, Prabhu R, Schreibmann E, Crocker I, Holder CA, Shim H, Shu HK. Simulating the effect of spectroscopic MRI as a metric for radiation therapy planning in patients with glioblastoma. Tomography. 2016;2:366–373.
19. Gurbani SS, Schreibmann E, Sheriff S, Cooper LAD, Shu HKG, Holder CA, Maudsley AA, Shim H. A software platform for collaborative radiation therapy planning using spectroscopic MRI. Int J Radiat Oncol Biol Phys. 2017;99:E667.
20. Sanderson C, Curtin R. Armadillo: a template-based C++ library for linear algebra. J Open Source Softw. 2016;1:26.
21. Yoo TS, Ackerman MJ, Lorensen WE, Schroeder W, Chalana V, Aylward S, Metaxas D, Whitaker R. Engineering and algorithm design for an image processing API: a technical report on ITK-the insight toolkit. Stud Health Technol Inform. 2002;85:586–92.
22. Maudsley AA, Domenig C, Govind V, Darkazanli A, Studholme C, Arheart K, Bloomer C. Mapping of brain metabolite distributions by volumetric proton MR spectroscopic imaging (MRSI). Magn Reson Med. 2009;61:548–559.
23. McKnight TR, von dem Bussche MH, Vigneron DB, Lu Y, Berger MS, McDermott MW, Dillon WP, Graves EE, Pirzkall A, Nelson SJ. Histopathological validation of a three-dimensional magnetic resonance spectroscopy index as a predictor of tumor presence. J Neurosurg. 2002;97:794–802.
24. Engwer C, Hillen T, Knappitsch M, Surulescu C. Glioma follow white matter tracts: a multiscale DTI-based model. J Math Biol. 2015;71:551–582.
25. Gurbani SS, Schreibmann E, Sheriff S, Holder CA, Cooper LAD, Maudsley A, et al. editors. Rapid internal normalization of spectroscopic MRI maps using a gaussian mixture model. In Proceedings of the American Association of Physicists in Medicine 59th Annual Meeting; Denver, CO; 2017.
26. Maudsley AA, Domenig C. Signal normalization for MR spectroscopic imaging using an interleaved water-reference. In Proceedings of the 16th Annual Meeting of ISMRM; Toronto, ON; 2008.
27. Keles GE, Chang EF, Lamborn KR, Tihan T, Chang C-J, Chang SM, Berger MS. Volumetric extent of resection and residual contrast enhancement on initial surgery as predictors of outcome in adult patients with hemispheric anaplastic astrocytoma. J Neurosurg. 2006;105:34–40.
28. Kurita T, Otsu N. Texture Classification by Higher Order Local Autocorrelation. In Proceedings of the Asian Conference on Computer Vision; Osaka, Japan; 1993.
29. Cordova JS, Schreibmann E, Hadjipanayis CG, Guo Y, Shu H-KG, Shim H, Holder CA. Quantitative tumor segmentation for evaluation of extent of glioblastoma resection to facilitate multisite clinical trials. Transl Oncol. 2014;7: 40–47.
30. Sabati M, Sheriff S, Gu M, Wei J, Zhu H, Barker PB, Spielman DM, Alger JR, Maudsley AA. Multivendor implementation and comparison of volumetric whole-brain echo-planar MR spectroscopic imaging. Magn Reson Med. 2015;74: 1209–1220.
31. RTOG. Randomized phase II trial of hypofractionated dose-escalated photon IMRT or proton beam therapy versus conventional photon irradiation with concomitant and adjuvant temozolomide in patients with newly diagnosed glioblastoma. Radiat Ther Oncol Gr NRG-BN001 Protoc Inf. 2014.
32. Gurbani SS, Schreibmann E, Maudsley AA, Cordova JS, Soher BJ, Poptani H, Verma G, Barker PB, Shim H, Cooper LAD. A convolutional neural network to filter artifacts in spectroscopic MRI. Magn Reson Med. 2018;80: 1765–1775.

TOMOGRAPHY® RESEARCH ARTICLE

Explaining Deep Features Using Radiologist-Defined Semantic Features and Traditional Quantitative Features

Rahul Paul[1], Matthew Schabath[2], Yoganand Balagurunathan[3], Ying Liu[4], Qian Li[4], Robert Gillies[3], Lawrence O. Hall[1], and Dmitry B. Goldgof[1]

[1]Department of Computer Science and Engineering, University of South Florida, Tampa, FL; [2]Department of Cancer Epidemiology, H. L. Moffitt Cancer Center & Research Institute, Tampa, FL; [3]Department of Cancer Imaging and Metabolism, H. L. Moffitt Cancer Center & Research Institute, Tampa, FL; and [4]Department of Radiology, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center of Cancer, Key Laboratory of Cancer Prevention and Therapy, Tianjin's Clinical Research Center for Cancer, Tianjin

Corresponding Author:
Dmitry B. Goldgof, PhD
Department of Computer Science & Engineering, USF College of Engineering, Building II 4220 E. Fowler Avenue, Tampa, FL 33620, USA;
E-mail: goldgof@mail.usf.edu.

Key Words: deep features, radiomics, semantic features, interpretation of features, CNN, explainable AI, quantitative features
Abbreviations: Convolutional neural network (CNN), low-dose computed tomography (LDCT), screen-detected lung cancer (SDLC)

ABSTRACT

Quantitative features are generated from a tumor phenotype by various data characterization, feature-extraction approaches and have been used successfully as a biomarker. These features give us information about a nodule, for example, nodule size, pixel intensity, histogram-based information, and texture information from wavelets or a convolution kernel. Semantic features, on the other hand, can be generated by an experienced radiologist and consist of the common characteristics of a tumor, for example, location of a tumor, fissure, or pleural wall attachment, presence of fibrosis or emphysema, concave cut on nodule surface. These features have been derived for lung nodules by our group. Semantic features have also shown promise in predicting malignancy. Deep features from images are generally extracted from the last layers before the classification layer of a convolutional neural network (CNN). By training with the use of different types of images, the CNN learns to recognize various patterns and textures. But when we extract deep features, there is no specific naming approach for them, other than denoting them by the feature column number (position of a neuron in a hidden layer). In this study, we tried to relate and explain deep features with respect to traditional quantitative features and semantic features. We discovered that 26 deep features from the Vgg-S neural network and 12 deep features from our trained CNN could be explained by semantic or traditional quantitative features. From this, we concluded that those deep features can have a recognizable definition via semantic or quantitative features.

INTRODUCTION

Lung cancer is one of the most common causes of malignancy worldwide, with a 5-year survival rate of 18% (1). The American Cancer Society estimates 14% of new cancer cases will be lung cancer cases for 2018, making it the second most detected cancer in the United States. They also estimate 154,050 deaths from lung cancer, which is the most in the United States in 2018 (2). As lung cancer typically remains undetected during the initial stages, ~75% of patients with lung cancers are first diagnosed at the advanced stages (III/IV) (3). As a result, early detection and diagnosis is a high priority.

Low-dose computed tomography (LDCT) is a noninvasive and widely used imaging technique for detecting lung nodules. By analyzing CT scans, radiologists can generate specific features from one's lung nodule, which could provide guidance for detection and diagnosis. These distinctive features are named semantic features. They can be categorized into the following different groups: shape (eg, lobulation), location (eg, lobe location), margin (eg, spiculation), external (eg, peripheral emphysema). With CT scans, cavitation is discovered in 22% of primary lung cancers and often the cavities in benign nodules mimic the cavities of malignant nodules, which makes precise diagnosis difficult (4). In another study (5), it was found that the risk of lung cancer can be increased 3- to 4-fold owing to emphysema among heavy smokers. Nodule size also influences cancer diagnosis and treatment (6). Hence, semantic features can be used in creating a predictor of lung cancer.

Using CT scans, quantitative information from a lung nodule can be generated and analyzed using statistics, machine learning, or high-dimensional data analysis. This approach is termed radiomics (7). These quantitative features can be categorized into the following different groups: texture (eg, Law's

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2018.00034

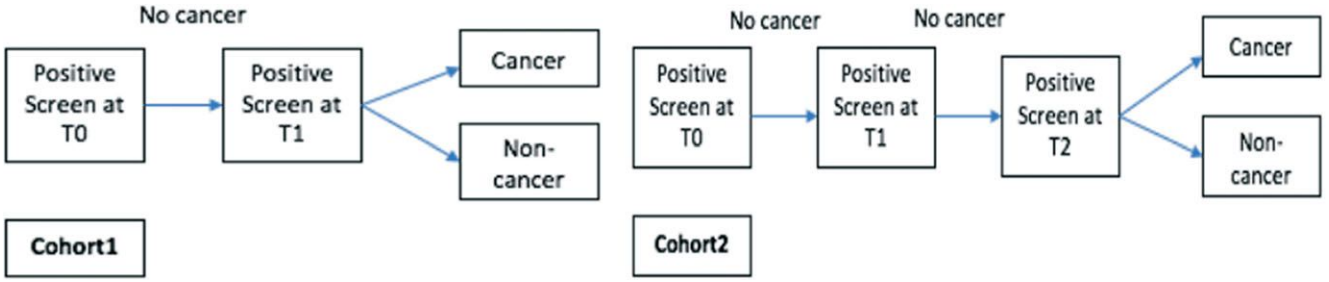


Figure 1. Selection of cohort 1 and cohort 2.

texture features, wavelet features), size (eg, longest diameter, volume), location (eg, attached to the pleural wall, distance from the boundary). These traditional quantitative features can be used to create a biomarker for tumor prognosis, analysis, and prediction (8-10).

Deep learning is an emerging approach mainly applied in recognition-, prediction-, and classification-related tasks. Propagating data through multiple hidden layers will eventually help a neural network to learn and build a representation of data, which can be used further for prediction or classification. For image data, a convolutional neural network (CNN) typically uses several convolutional kernels to extract different textures and edges before propagating the extracted information through multiple hidden layers. For lung nodule analysis, CNNs have been used effectively in recent years (11). In the medical imaging field, data are currently scarce; so, as an alternative to building a new model, transfer learning has been used (12).

Convolution layers of CNNs, after learning, contain representations of edge gradients and textures, and when propagated through fully connected layers, various high-level features are posited to have been learned by the network. From fully connected layers, deep features (the outputs of units in the layer) are extracted and denoted by the number of the feature from the learning tool (the position of a neuron in a hidden layer row vector).

Two pretrained CNNs were used in the work described in this paper for extracting the following deep features: the Vgg-S network (13), which was trained on the ImageNet data set (14) of color camera images and our designed CNN (15), which was trained on lung nodule images. There were 23 traditional quantitative features [RIDER subset features (16)] used in this study along with 20 semantic features, which were generated by an experienced radiologist from Tianjin Medical University Cancer Institute and Hospital, China. This study is an extension of our previous study (17), which analyzes the similarity between deep features and semantic features. In this current study, we also focused on traditional quantitative features, that is, analyzed the similarity of deep feature(s) to traditional quantitative features. The analysis was conducted by replacing ≥ 1 deep features with traditional quantitative or semantic feature(s). The goal was to show that equivalent classification performance can be achieved. That means those deep features contained information similar to that of the semantic or traditional quantitative features. We can equate those deep features with the name of the corresponding semantic or traditional quantitative feature.

We found that location-based semantic features are difficult to replace, but size-, shape-, and texture-based semantic features can be replaced by deep feature(s). Therefore, shape and texture quantitative features can be used to explain deep feature(s). By "explain," we mean the features can replace deep features and a classifier will achieve the same accuracy. We successfully explained 26 deep features from the Vgg-S network out of 4096 features and 12 deep features from our trained CNN by semantic and traditional quantitative features. This provides a semantic meaning for the deep features.

METHODOLOGY

Data Set

A subset of cases from the LDCT-arm of the NLST (National Lung Screening Trial) data set was chosen for this study. The NLST study was conducted over 3 years: 1 baseline scan (T0) and 2 following scans (T1 and T2) in 2 subsequent years with an interval of $\sim$1 year (18) between scans. For this study, a subset of nodule-positive and screen-detected lung cancer (SDLC) cases (years later) from the baseline (T0) scans were chosen, and the patient data were deidentified under an IRB-approved process. These subsets of cases were further divided into the following 2 categories: cohort 1 and cohort 2. Cohort 1 consisted of cases with a baseline scan (T0), which had a follow-up scan after 1 year (T1), wherein some of the nodules became cancerous. Whereas, cohort 2 consisted of nodules that became cancerous after 2 years (T2 scan) from the baseline scan (T0). Selection of cohorts is shown in Figure 1. Only Cohort 2 (SDLC, 85; positive control cases, 152) was chosen for our study. Between the SDLC and control-positive cases, there is no statistically significant difference with respect to sex, race age, ethnicity, and smoking (19). Nodule segmentation was performed using the Definiens software suite (20). From our initial set of cases, 52 cases were excluded owing to ≥ 1 of the following reasons: multiple malignant nodules, inability to identify the nodule, or unknown location of the tumor. So, finally, 185 cases (SDLC, 58; control-positive cases, 127) were selected for our study.

Semantic Features

Semantic features were described from the CT scan of a lung tumor, by an experienced radiologist. They can be used further

Table 1. Description of Semantic Features

Characteristic	Definition	Scoring
Location		
1. Lobe Location	Lobe location of the nodule	Left lower lobe (5), left upper lobe (4), right lower lobe (3), right middle lobe (2), right upper lobe (1)
Size		
2. Long-Axis Diameter	Longest diameter of the nodule	NA
3. Short-Axis Diameter	Longest perpendicular diameter of nodule in the same section	NA
Shape		
4. Contour	Roundness of the nodule	1, round; 2, oval; 3, irregular
5. Lobulation	Wavy nodule's surface	1, none; 2, yes
6. Concavity	Concave cut on nodule surface	1, none; 2, slight concavity; 3, deep concavity
Margin		
7. Border Definition	Edge appearance of the nodule	1, well defined; 2, slight poorly; 3, poorly defined
8. Spiculation	Lines radiating from the margins of tumor	1, none; 2 yes
Attenuation		
9. Texture	Solid, non-solid, part solid	1, non-solid; 2, part solid; 3, solid
10. Cavitation	Presence of air in the tumor at the time of diagnosis	0, no; 1, yes
External		
11. Fissure Attachment	Nodule attaches to the fissure	0, no; 1, yes
12. Pleural Attachment	Nodules attaches to the pleura	0, no; 1, yes
13. Vascular Convergence	Convergence of vessels to nodule	0, no significant convergence; 1, significant
14. Pleural Retraction	Retraction of the pleura towards nodule	0, absence of pleural retraction; 1, present
15. Peripheral Emphysema	Peripheral emphysema caused by nodule	1, absence of emphysema; 2, slight present; 3 severely present
16. Peripheral Fibrosis	Peripheral fibrosis caused by nodule	1, absence of fibrosis; 2, slight present; 3 severely present
17. Vessel Attachment	Nodule attachment to blood vessel	0, no; 1, yes
Associated Findings		
18. Nodules in Primary Lobe	Any nodules suspected to be malignant or intermediate	0, no; 1, yes
19. Nodules in Nonprimary Lobe	Any nodules suspected to be malignant or intermediate	0, no; 1, yes
20. Lymphadenopathy	Lymph nodes with short- axis diameter greater than 1 cm	0, no; 1, yes

for diagnosis. An experienced radiologist (Y.L.) with 7 years of experience from Tianjin Medical University Cancer Institute and Hospital, China, described 20 semantic features (21-24) on a subset of cases that intersected Cohort 2. Semantic features can be categorized into the following groups: shape, size, location, margin, external attenuation, and associated findings. These features have been derived with respect to lung nodules by our group. Table 1 shows a detailed description of our semantic features.

Traditional Quantitative Features

Definiens software (20), along with help from a radiologist, was used to segment lung nodules. Then 23 Rider stable features (16) were extracted using Definiens software. Table 2 shows a detailed description of the "traditional" quantitative features.

Deep Features from Vgg-S Network

Nowadays CNNs are used effectively for image classification and prediction (11, 13). A CNN has many layers of convolution kernels along with multiple hidden layers, which makes the network architecture deeper, and features extracted from such a network are called "deep features." In the medical imaging field, there is typically not enough original data available to train a CNN. As a result, transfer learning (12) is an alternative option. Applying previously learned knowledge from 1 domain to a new task domain is called transfer learning. To extract deep features from a CT scan, the 2-dimensional slice, which has the largest nodule area, was chosen for every case. We extracted only the nodule region by incorporating the largest rectangular box around the nodule. Bicubic interpolation was used to resize the nodule images to 224 × 224, which was the required input size of the Vgg-S network. Figure 2 shows a lung image with nodule

Table 2. Description of Rider Stable Traditional Quantitative Features

Characteristic	Features
Size	1. Long-axis diameter
	2. Short-axis diameter
	3. Long-axis diameter × short-axis diameter
	4. Volume (cm)
	5. Volume (pixel)
	6. Number of pixels
	7. Length/width
Pixel Intensity Histogram	8. Mean (HU)
	9. Stand deviation (HU)
Tumor Location	10. 8a_3D_ is attached to pleural wall
	11. 8b_3D Relative border to lung
	12. 8c_3D_Relative border to pleural wall
	13. 9e_3D_Standard deviation_ COG to border
	14. 9g_3D_max_Dist_COG to border
Tumor Shape (Roundness)	15. 9b-3D circularity
	16. 5a_3D- MacSpic
	17. Asymmetry
	18. Roundness
Run-length and Co-occurrence	19. Avg_RLN
Law's Texture Feature	20. E5 E5 L5 layer 1
	21. E5 E5 R5 layer 1
	22. E5 W5 L5 layer 1
	23. L5 W5 L5 layer 1

and the extracted nodule region. The Vgg-S network was trained using natural camera images, which were 3-channel (R, G, B), but the nodule images were grayscale (no color component and voxel intensities of the CT images were converted to 0-255). So, the same grayscale nodule image was used 3 times to mimic an image with 3 color channels and then normalization was performed using the appropriate color channel image. The deep features were generated from the last fully connected layer after applying the ReLU activation function. The size of the feature vector was 4096.

Deep Features from Our Trained CNN

We also experimented by extracting deep features from our designed CNN network (15). Augmented nodule images of Cohort 1 were used to train our CNN architecture. Each nodule image was augmented first by being flipped horizontally and vertically and then all images were rotated by 15°. Keras (25) with a Tensorflow (26) backend was used to train our CNN. We used the same 2-dimensional slice from a nodule for training the CNN and for transfer learning using the Vgg-S network. The input image size for the CNN architecture was 100 × 100 pixels.

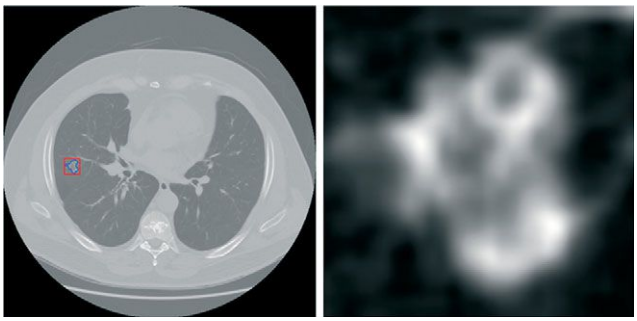


Figure 2. (Left) lung image with nodule inside outlined in blue (nodule pixel size =0.74 mm), with box used for extracted nodule in red, (Right) extracted nodule.

The augmented data set was divided into the following 2 parts: 70% of the data for training and the remaining 30% for validation. The CNN was trained for 100 epochs with 0.0001 learning rate with RMSprop (27) optimization and binary cross-entropy as loss function. A batch size of 16 was chosen for training and validation. L2 regularization (28) along with dropout (29) was used to reduce overfitting of our small and shallow CNN network. Our designed CNN is described in detail in Table 3. The deep features were extracted from the last layer before the

Table 3. Our Designed CNN architecture

Layers	Parameter	Total Parameters
Left branch		
Input Image	100 × 100	
Max Pool 1	10 × 10	
Dropout	0.1	
Right branch		
Input Image	100 × 100	
Conv 1	64 × 5 × 5, pad 0, stride 1	
Leaky ReLU	alpha = 0.01	
Max Pool 2a	3 × 3, pad 0, stride 3	39,553
Conv 2	64 × 2 × 2, pad 0, stride 1	
Leaky ReLU	alpha = 0.01	
Max Pool 2b	3 × 3, pad 0, stride 3	
Dropout	0.1	
Concatenate Left Branch + Right Branch		
Conv 3 + ReLU	64 × 2 × 2, pad 0, stride 1	
Max Pool 3	2 × 2, pad 0, stride 2	
L2 regularizer	0.01	
Dropout	0.1	
Fully Connected 1	1 sigmoid	

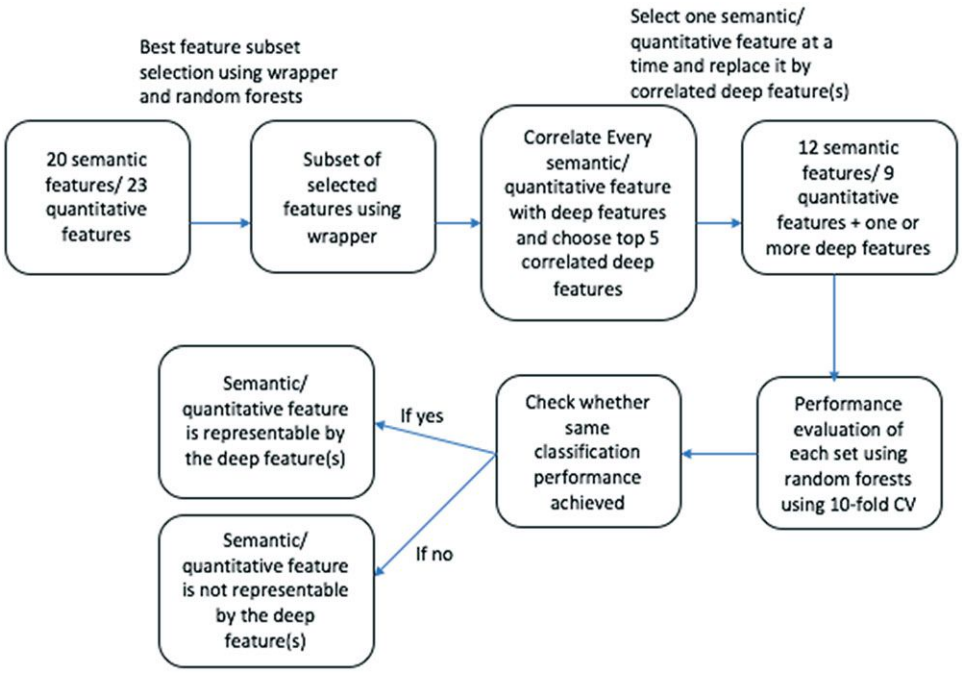


Figure 3. Overview of the approach taken in this study.

classification layer. The size of the feature vector was 1024. After applying the ReLU activation function, some features will be all zeros because ReLU truncates the negative feature values to zero. We removed such features, and as a result, the final number of feature vectors from Vgg-S pretrained CNN and our trained CNN became 3844 and 560, respectively.

Experiments and Results

This section describes the procedure of representing deep feature(s) using semantic or traditional quantitative features.

Wrapper feature selection (30) was applied on traditional quantitative or semantic features of Cohort 2 to select the best subset of features with maximum accuracy. Backward feature selection using the best first strategy and random forests classifier (31) with 200 trees was applied using the wrapper approach. Tenfold cross-validation was used for selecting the best subset of features. We analyzed quantitative features and semantic features separately. A subset of 9 quantitative features was chosen and it enabled a maximum accuracy of 84.32% (AUC 0.87), whereas a subset of 13 semantic features were selected, enabling a maximum accuracy of 83.78% (AUC 0.84). Here, we aim to use semantic features or traditional quantitative features to interpret/explain deep feature(s).

Explaining Deep Features With Respect to Semantic Features

The chosen semantic features (13) were location, long-axis diameter, short-axis diameter, lobulation, concavity, border definition, spiculation, texture, cavitation, vascular convergence, vessel attachment, perinodule fibrosis, and nodules in primary tumor lobe.

After selecting the best subset of semantic features, the correlation coefficient (Pearson correlation coefficient) was calculated for each semantic feature with the deep features, and the 5 most correlated features for each semantic feature were selected. We then replaced each semantic feature with the correlated deep feature(s) and checked whether the same classification accuracy of 83.78% could be achieved.

Our purpose for the study was to determine if semantic features could explain deep features. To do this, we replaced each semantic feature by ≥ 1 deep features to see if the same classification accuracy could be achieved. We replaced 1 semantic feature at a time from the subset of 13 features and substituted that semantic feature by, at first, the most correlated deep feature and, then 2 most correlated deep features and proceeded similarly to add features until the 5 most correlated deep features had been used as replacements. The accuracy was calculated using a random forests classifier with 200 trees using 10-fold cross-validation. Deep features from Vgg-S pretrained CNN and our trained CNN were examined separately. Figure 3 shows the approach taken for the analysis.

After replacing a feature with deep features extracted from the Vgg-S pretrained CNN, we secured the same original classification accuracy of 83.78% for the following 8 semantic features: long-axis diameter, lobulation, concavity, spiculation, texture, cavitation, vascular convergence, and peripheral fibrosis. Using the deep features acquired from our trained CNN, we achieved the same original classification accuracy of 83.78% for the following 4 semantic features: long-axis diameter, concavity, cavitation, nodules in primary tumor lobe. We found that 3 semantic features (long-axis diameter, concavity, cavitation) could be used to explain both deep features from Vgg-S and our trained CNN. Five semantic features could be used to explain only deep features from Vgg-S, and only 1 semantic feature could be used to explain deep features from our trained CNN. The Vgg-S network was trained on camera images from at least 1000 classes of objects, but not lung nodule images. The large training set helped the network to develop general features and which in turn were explained by texture, spiculation, lobulation, vascular convergence, and peripheral fibrosis. The replacement

Table 4. Classification performance After Features Removal

Features	Feature Names	Accuracy
Semantic Features	Long-axis diameter	82.70 (0.82)
	Lobulation	82.70 (0.83)
	Concavity	83.24 (0.83)
	Spiculation	83.24 (0.83)
	Texture	82.70 (0.83)
	Cavitation	82.70 (0.83)
	Vascular convergence	83.24 (0.84)
	Peripheral fibrosis	82.70 (0.83)
	Nodules in primary lobe	81.62 (0.83)
Traditional Quantitative Features	9b-3D circularity	82.16 (0.86)
	Roundness	82.70 (0.87)
	L5W5L5 layer 1	82.70 (0.87)

These features were from our chosen subset of features, leaving 12 features for training/testing.

of the first 3 and the last feature appear to result from training on lots of images of different types.

Table 4 shows the performance of each semantic feature after removing 1 semantic feature at a time from the subset of 13 features. So, we only calculated classification performance of 12 features at a time using random forests classifier using 10-fold cross-validation, to check whether by removing each feature, there was a change in classification accuracy. In Table 4, we show only the semantic features out of the chosen 13 feature subsets that could be used to explain deep feature(s). Table 5 shows the explainable deep features and their equivalent semantic feature(s). We also show the correlation value of each deep feature with a semantic feature in Table 5.

After replacing semantic features with deep feature(s), similar classification performance was obtained for 9 semantic features. For example, 2 deep features (3353 and 526) from the Vgg-S network could achieve the same classification performance of 83.78% if used in place of cavitation. The deep features 3353 and 526 had the correlation of 0.388 and 0.3551, respectively, with the semantic feature cavitation. Whereas, the deep feature 395 from our trained CNN, which had a correlation coefficient of 0.2748, was explained by cavitation. Similarly, 2 deep features (3353 and 2135) from the Vgg-S network and 1 deep feature (230) using the features from our trained CNN were explained long-axis diameter by providing equivalent performance.

Explaining Deep Features Using Traditional Quantitative Features

The 9 traditional quantitative features that enabled the best accuracy were: Mean (HU), 8a-3D_is_attached to pleural wall, 8c-3D_Relative border to pleural wall, 9b-3D circularity, Asymmetry, Roundness, Volume, E5W5L5, and L5W5L5. The Pearson correlation coefficient was calculated for each traditional quantitative feature with the deep features and the top 5 correlated deep features were selected to replace each traditional quantitative feature. We replaced each traditional quantitative feature by $\geq$1 deep features to try to achieve the same classification accuracy of 84.32%. After replacing deep features extracted from the Vgg-S pretrained CNN, we got the same original classification accuracy of 84.32% for the following 3 traditional quantitative features: 9b-3D circularity, roundness, and L5W5L5 layer 1. Hence, they can be used to explain what the deep features that replaced them have learned. Traditional quantitative features consist of tumor size, tumor shape, Law's texture features, tumor location, etc. As we have seen earlier for semantic features, deep features could be explained by shape-based quantitative features.

In Table 4, we only show the 3 quantitative features that can be replaced (used to explain) deep feature(s). Table 5 shows the quantitative features, their equivalent deep feature(s), and correlations.

DISCUSSION

We showed that some deep features can be explained by a semantic feature or traditional quantitative feature. From a lung nodule CT image, experienced radiologists generated semantic features of different types of information regarding a lung nodule, for example, size, shape, location of nodule, the boundary of the nodule, attachment to the vessel, fibrosis information, etc. These features were shown to provide useful information toward the prognosis and diagnosis of lung cancer. From a tumor phenotype, quantitative information can be extracted using various data characterization approaches, and these features are called traditional quantitative features.

Deep features are extracted from a CNN, generally from the last layer before the final classification layer. For this study, deep features were extracted from the last fully connected layer of the following 2 pretrained CNNs: the Vgg-S network, which was trained on the ImageNet data set, and our designed CNN, which was trained on LDCT lung nodule images. The Vgg-S architecture is a network with 5 convolution layers followed by 3 fully connected layers. Our designed CNN is a small and shallow network with 3 convolution layers and 1 fully connected layer. As the Vgg-S network was trained on a large set of classes of camera images, various textures and other features

Table 5. Semantic and Traditional Quantitative Features and Corresponding Deep Feature(s)

Features	Feature Names	Deep Features from Vgg-S With Correlation Value					Deep Features from Our Trained CNN With Correlation Value			
Semantic Features	Long-axis diameter	3353		2135			230			
		0.4334		0.42			0.3055			
	Lobulation	3534		1372	2975	2111	NA			
		0.5742		0.5614	0.5611	0.5520				
	Concavity	3534	2975	1372	2111	3246	547		440	
		0.5	0.4839	0.4837	0.475	0.4612	0.1776		0.1514	
	Spiculation	2811					NA			
		0.4111								
	Texture	1201		3350			NA			
		−0.3119		0.2936						
	Cavitation	3353		526			395			
		0.3888		0.3551			0.2748			
	Vascular convergence	1464		2115			NA			
		0.7052		0.701						
	Peripheral fibrosis	3305		3064			NA			
		0.2076		0.2043						
	Nodules in primary lobe	NA					425		57	
							0.1871		0.1836	
Traditional Quantitative Features	Roundness	1395		2510			160		20	
		0.3		0.27			0.16		0.13	
	9b-3d circularity	1395		1757	3401	2777	160		20	
		0.24		−0.234	−0.2069	−0.2069	0.14		0.13	
	L5W5L5 layer 1	51	66	163	476	928	547	169	265	309
		0.77	0.75	0.69	0.69	0.69	0.28	0.27	0.26	0.26

were extractable, which can be used effectively for tumor classification. Our trained CNN was trained with LDCT lung nodule images and gave us better performance than transfer learning in our previous study (15).

In this study, we attempted to explain deep features using semantic or traditional quantitative features. A subset of features was chosen from the semantic or traditional quantitative features using a wrapper with a random forests classifier. For the semantic features, the best subset had 13 features with an accuracy of 83.78% (AUC 0.84), whereas from traditional quantitative features, the size of the best subset was 9 features with an accuracy of 84.32% (AUC 0.87). The Pearson correlation coefficient was calculated with each of the chosen semantic features or traditional quantitative features and the deep features. For every semantic or traditional quantitative feature, the top 5 most correlated deep features were chosen. Now, from our chosen subset of semantic or traditional quantitative features, 1 feature was removed, and it was substituted by the most correlated deep feature and classification performance was calculated. With a single substituted deep feature, if we can achieve the classification performance then stop; otherwise, substitute that semantic feature or traditional quantitative feature by the 2 most correlated features and continue this process until the 5 most correlated deep features have been used. In total, 26 deep features from the Vgg-S network and 12 deep features from our trained CNN were explained by 9 semantic features and 3 traditional quantitative features. From this, we hypothesized that those deep features can have a recognizable definition from semantic or quantitative features. That is, those deep features can be given some meaningful definition.

We also trained our CNN on cohort 2 (all 237 cases) and then extracted deep features for only the subset of 185 cases for which semantic features were available. The deep feature vector size was 1024. We removed all zero features to get 699 features from cohort 2. We then used these deep features to represent semantic and quantitative features. We found that some additional semantic features could be used to explain deep features from our CNN trained on cohort 1 (shown in Table 5) in addition to the ones previously found useful. Lobulation, spiculation, vascular convergence, perinodule fibrosis and border definition could explain features from our new deep feature set (CNN trained on cohort 2 data only). Among these semantic features, "border definition" was found to explain 4 deep features (147, 160, 504, and 372) and it could not explain any deep features from Vgg-S or our CNN (trained on cohort 1).

For this study, we extracted only the nodule region from a CT slice. As the nodule region was extracted the information regarding pleural wall attachment, fissure attachment, relative

border to the lung, or distance was lost. However, deep features from our trained CNN were explained by only 1 location-based semantic feature (nodules in primary lobe). For training the CNN, we performed data augmentation by rotation and flipping, which enabled the extracted deep features to achieve comparable accuracy. The deep features capture the boundary and shape information quite well because that information could be obtained from the extracted nodule region, and thus, 2 traditional quantitative features (9b-3D-circularity and roundness) and 3 semantic features (lobulation, concavity, and spiculation) were able to explain deep features. Deep features are known to grasp texture-based information as well. As a result, L5W5L5 Law's texture feature and cavitation were useful for explaining deep features. We also found out that deep features 3353, 3534, 1372, 2975, and 2111 from the Vgg-S network were correlated with and explained by >1 semantic features, and feature 1395 was correlated with and explained by 2 traditional quantitative features (roundness and 9b_3D_circularity). Deep features 160 and 20 from our trained CNN network were explained by 2 traditional quantitative features (roundness and 9b_3D_circularity).

In this work, the 5 most correlated features were used to replace a semantic or radiomics feature. Our requirement was some nonzero correlation. Now, with all the comparisons, there will potentially be some spurious correlations. Hence, the Bonferroni correction was used to look at the significance of correlations between deep features and every semantic (or radiomics) feature. As an example, cavitation could be replaced by 2 deep features from the Vgg-S network. Fea 1 (3353) had an original P value = 4.8651e-08 and fea 2 (526) had an original P value = 7.0822e-07. After the Bonferroni correction, the P value of fea 1 was 9.73e-08 and that of fea 2 was 1.4164e-06. Now both Bonferroni-corrected P-values were less than the more rigorous significance level. However, when combined, they added more information to our model and hence appear to be associated with cavitation.

After using the Bonferroni correction, we found some of the features with the 5 highest correlation values did not have a significant correlation with a semantic or radiomics feature. Nonetheless, the weakly correlated features were able to explain some CNN features. We interpret this to mean that insignificant, but nonzero, correlations taken together can provide insight into (some) deep features.

In total, 26 deep features from the Vgg-S network and 12 deep features from our trained CNN were explained by 9 semantic features and three traditional quantitative features.

CONCLUSIONS

The recent success of CNNs in various classification-type tasks leads to the question of what they have learned. Here, deep features are explained with respect to semantic features and traditional quantitative features.

In this study, we found explanations for 26 deep features from the Vgg-S network out of 4096 features and 12 deep features from our trained CNN by semantic and traditional quantitative features. One can also look at this as providing semantic information about deep features. Although there has been some research (32-39) regarding semantic understanding of natural scenes using deep CNN features, to our knowledge, this is the first work to explain deep features with respect to traditional quantitative features and semantic features extracted from a lung nodule. In the future, deep features with semantic meaning can be included in biomarkers for tumor prognosis and diagnosis of lung nodules from CT scans, along with semantic features and traditional quantitative features.

There were 2 limitations in our study, first, only 10-fold cross-validation was used to evaluate the performance as we had a limited set of expensive to obtain semantic information. The second limitation of our study was using a single slice for every patient to extract deep features, whereas semantic information was generated from multiple slices. In the future with more semantic annotated data, we will investigate deep features from a 3D CNN.

ACKNOWLEDGMENTS

This research partially supported by the National Institute of Health under grants (NIH U01 CA143062), (NIH U24 CA180927) and (NIH U01 CA200464), National Science Foundation under award number 1513126 and by the State of Florida Dept. of Health under grant (4KB17). DARPA's SocialSim program under award number FA8650-17-C-7725.

Disclosures: No disclosures to report.

Conflict of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA Cancer J Clin. 2015 Jan;65:5–29.
2. American Cancer Society. Key Statistics for Lung Cancer. [cited 31 Aug 18] Available from: https://www.cancer.org/cancer/non-small-cell-lung-cancer/about/key-statistics.html.
3. Walters S, Maringe C, Coleman MP, Peake MD, Butler J, Young N, Bergström S, Hanna L, Jakobsen E, Kölbeck K, Sundstrøm S. Lung cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK: a population-based study, 2004–2007. Thorax. 2013;68:551–564.
4. Gill RR, Matsusoka S, Hatabu H. Cavities in the lung in oncology patients: imaging overview and differential diagnoses. Appl Radiol. 2010;39:10.
5. Li Y, Swensen SJ, Karabekmez LG, Marks RS, Stoddard SM, Jiang R, Worra JB, Zhang F, Midthun DE, de Andrade M, Song Y. Effect of emphysema on lung cancer risk in smokers: a computed tomography-based assessment. Cancer Prev Res (Phila). 2010;4:43–50.
6. MacMahon H, Naidich DP, Goo JM, Lee KS, Leung AN, Mayo JR, Mehta AC, Ohno Y, Powell CA, Prokop M, Rubin GD. Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society 2017. Radiology. 2017;284:228–243.
7. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology. 2015;278:563–577.
8. Zhang Y, Oikonomou A, Wong A, Haider MA, Khalvati F. Radiomics-based prognosis analysis for non-small cell lung cancer. Sci Rep. 2017;7:46349.
9. Chen CH, Chang CK, Tu CY, Liao WC, Wu BR, Chou KT, Chiou YR, Yang SN, Zhang G, Huang TC. Radiomic features analysis in computed tomography images of lung nodule classification. PloS One. 2018;13:e0192002.
10. Chaddad A, Desrosiers C, Toews M, Abdulkarim B. Predicting survival time of lung cancer patients using radiomic analysis. Oncotarget. 2017;8:104393.
11. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012;1097–1105.

12. Raina R, Battle A, Lee H, Packer B, Ng AY. Self-taught learning: transfer learning from unlabeled data. Proceedings of the 24th International Conference on Machine Learning. 2007;759–766.
13. Chatfield K, Simonyan K, Vedaldi A, Zisserman A. Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531. 2014 May 14.
14. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on 2009 Jun 20 (pp. 248–255). IEEE.
15. Paul R, Hawkins SH, Schabath MB, Gillies RJ, Hall LO, Goldgof DB. Predicting malignant nodules by fusing deep features with classical radiomics features. J Med Imaging (Bellingham). 2018;5:011021.
16. Balagurunathan Y, Kumar V, Gu Y, Kim J, Wang H, Liu Y, Goldgof DB, Hall LO, Korn R, Zhao B, Schwartz LH. Test–retest reproducibility analysis of lung CT image features. J Digit Imaging. 2014;27:805–823.
17. Paul R, Liu Y, Li Q, Hall L, Goldgof D, Balagurunathan Y, Schabath M, Gillies R. Representation of Deep Features using Radiologist defined Semantic Features. Proc Int Jt Conf Neural Netw. 2018;1–7.
18. National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011;365:395–409.
19. Schabath MB, Massion PP, Thompson ZJ, Eschrich SA, Balagurunathan Y, Goldof D, Aberle DR, Gillies RJ. Differences in patient outcomes of prevalence, interval, and screen-detected lung cancers in the CT arm of the national lung screening trial. PloS One. 2016;11:e0159880.
20. Definiens Developer XD. 2.0. 4 User Guide. Definiens AG, Munich, Germany. 2009.
21. Liu Y, Kim J, Qu F, Liu S, Wang H, Balagurunathan Y, Ye Z, Gillies RJ. CT features associated with epidermal growth factor receptor mutation status in patients with lung adenocarcinoma. Radiology. 2016;280:271–280.
22. Li Q, Balagurunathan Y, Liu Y, Qi J, Schabath MB, Ye Z, Gillies RJ. Comparison Between Radiological Semantic Features and Lung-RADS in Predicting Malignancy of Screen-Detected Lung Nodules in the National Lung Screening Trial. Clin Lung Cancer. 2018;19:148–156.
23. Liu Y, Wang H, Li Q, McGettigan MJ, Balagurunathan Y, Garcia AL, Thompson ZJ, Heine JJ, Ye Z, Gillies RJ, Schabath MB. Radiologic features of small pulmonary nodules and lung cancer risk in the National Lung Screening Trial: a nested case-control study. Radiology. 2017;286:298–306.
24. Liu Y, Balagurunathan Y, Atwater T, Antic S, Li Q, Walker RC, Smith G, Massion PP, Schabath MB, Gillies RJ. Radiological image traits predictive of cancer status in pulmonary nodules. Clin Cancer Res. 2016;23:1442–1449.
25. Chollet F. Keras: The python deep learning library. Astrophysics Source Code Library. 2018.
26. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M. Tensorflow: a system for large-scale machine learning. OSDI. 2016;16:265–283.
27. Tieleman T, Hinton G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning. 2012;4:26–31.
28. Ng AY. Feature selection, L 1 vs. L 2 regularization, and rotational invariance. In Proceedings of the Twenty-First International Conference on Machine learning. 2004;Jul 4:78. ACM.
29. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014; 15:1929–1958.
30. Kohavi R, Sommerfield D. Feature Subset Selection Using the Wrapper Method: Overfitting and Dynamic Search Space Topology. In KDD 1995 Aug 20 (pp. 192–197).
31. Ho TK. Random decision forests. In Document analysis and recognition, 1995, Proceedings of the Third International Conference on 1995 Aug 14 (Vol. 1, pp. 278–282). IEEE.
32. Gudi A. Recognizing semantic features in faces using deep learning. arXiv preprint arXiv:1512.00743. 2015 Dec 2.
33. Ufer N, Ommer B. Deep semantic feature matching. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on 2017 Jul 21 (pp. 5929–5938). IEEE.
34. Aubry M, Russell BC. Understanding deep features with computer-generated imagery. In Proceedings of the IEEE International Conference on Computer Vision 2015 (pp. 2875–2883).
35. Zhao RW, Wu Z, Li J, Jiang YG. Learning semantic feature map for visual content recognition. In Proceedings of the 2017 ACM on Multimedia Conference 2017 Oct 23 (pp. 1291–1299). ACM.
36. Chen C, Wu Z, Jiang YG. Emotion in Context: Deep Semantic Feature Fusion for Video Emotion Recognition. In Proceedings of the 2016 ACM on Multimedia Conference 2016 Oct 1 (pp. 127–131). ACM.
37. Li H, Peng J, Tao C, Chen J, Deng M. What do We Learn by Semantic Scene Understanding for Remote Sensing imagery in CNN framework? arXiv preprint arXiv:1705.07077. 2017 May 19.
38. Lynch C, Aryafar K, Attenberg J. Images Don't Lie: Transferring Deep Visual Semantic Features to Large-Scale Multimodal Learning to Rank. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016 Aug 13 (pp. 541–548). ACM.
39. Li S, Zhao Z, Liu T, Hu R, Du X. Initializing convolutional filters with semantic features for text classification. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing 2017 (pp. 1884–1889).

Deep Learning Approach for Assessment of Bladder Cancer Treatment Response

Eric Wu[1], Lubomir M. Hadjiiski[1], Ravi K. Samala[1], Heang-Ping Chan[1], Kenny H. Cha[1], Caleb Richter[1], Richard H. Cohan[1], Elaine M. Caoili[1], Chintana Paramagul[1], Ajjai Alva[2], and Alon Z. Weizer[3]

Departments of [1]Radiology, [2]Internal Medicine-Hematology/Oncology, and [3]Urology, University of Michigan, Ann Arbor, MI

Corresponding Author:
Eric Wu
Department of Radiology, University of Michigan, 1500 E. Medical Center Drive, *MIB C473,* Ann Arbor, MI 48109-5842;
E-mail: ehwu@umich.edu

Key Words: deep-learning, transfer learning, treatment response, segmentation, CT, bladder
Abbreviations: Deep learning-convolutional neural network (DL-CNN), regions of interest (ROIs), area under the ROC curve (AUC), methotrexate, vinblastine, doxorubicin, and cisplatin (MVAC), computed tomography (CT), magnetic resonance imaging (MRI)

ABSTRACT

We compared the performance of different Deep learning-convolutional neural network (DL-CNN) models for bladder cancer treatment response assessment based on transfer learning by freezing different DL-CNN layers and varying the DL-CNN structure. Pre- and posttreatment computed tomography scans of 123 patients (cancers, 129; pre- and posttreatment cancer pairs, 158) undergoing chemotherapy were collected. After chemotherapy 33% of patients had T0 stage cancer (complete response). Regions of interest in pre- and posttreatment scans were extracted from the segmented lesions and combined into hybrid pre -post image pairs (h-ROIs). Training (pairs, 94; h-ROIs, 6209), validation (10 pairs) and test sets (54 pairs) were obtained. The DL-CNN consisted of 2 convolution (C1-C2), 2 locally connected (L3-L4), and 1 fully connected layers. The DL-CNN was trained with h-ROIs to classify cancers as fully responding (stage T0) or not fully responding to chemotherapy. Two radiologists provided lesion likelihood of being stage T0 posttreatment. The test area under the ROC curve (AUC) was 0.73 for T0 prediction by the base DL-CNN structure with randomly initialized weights. The base DL-CNN structure with pretrained weights and transfer learning (no frozen layers) achieved test AUC of 0.79. The test AUCs for 3 modified DL-CNN structures (different C1-C2 max pooling filter sizes, strides, and padding, with transfer learning) were 0.72, 0.86, and 0.69. For the base DL-CNN with (C1) frozen, (C1-C2) frozen, and (C1-C2-L3) frozen, the test AUCs were 0.81, 0.78, and 0.71, respectively. The radiologists' AUCs were 0.76 and 0.77. DL-CNN performed better with pretrained than randomly initialized weights.

INTRODUCTION

Bladder cancer is the fourth most common cancer in men. The American Cancer Society estimates that in 2018, 81 190 (men, 62 380; women, 18 810) new cases of bladder cancer will be diagnosed in the United States, with 17 240 (men, 12 520; women, 4720) deaths (1). Early treatment of bladder cancer is important to reduce morbidity and mortality, as well as reduce costs.

Radical cystectomy is considered the gold standard for treatment of patients with localized muscle-invasive bladder cancer. However, about 50% of such patients develop metastases within 2 years after cystectomy and subsequently die of the disease (2). Neoadjuvant chemotherapy of muscle-invasive operable bladder cancer has been shown to be beneficial for treating micrometastases and improving resectability of larger neoplasms before radical cystectomy (3-5). Chemotherapy involving methotrexate, vinblastine, doxorubicin, and cisplatin (MVAC) followed by radical cystectomy increases the probability of finding no residual cancer at surgery compared with radical cystectomy alone and improves survival among patients with locally advanced bladder cancer (6, 7). In clinical trials, downstaging with drugs before surgery was shown to have significant survival benefits (7, 8). Current standard of care uses the neoadjuvant protocol consisting of 12 weeks of chemotherapy preceding radical cystectomy.

Although patients with advanced disease can benefit from neoadjuvant chemotherapy, there are drawbacks. Chemotherapy with the MVAC regimen has substantial toxicity and side effects (9). Significant toxicities, primarily leucopenia, culture-negative fever at the time of granulocytopenia, sepsis, and mucositis are associated with MVAC combination chemotherapy. Side effects such as nausea, vomiting, malaise, and alopecia are common. In addition, chemotherapy is expensive. However, because no reliable method yet exists for predicting the response of an individual case to chemotherapies such as MVAC, some patients may suffer from adverse reactions to the drugs without achieving beneficial effects, often also missing the opportunity for alternative therapy when their physical condition deteriorates.

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2018.00036

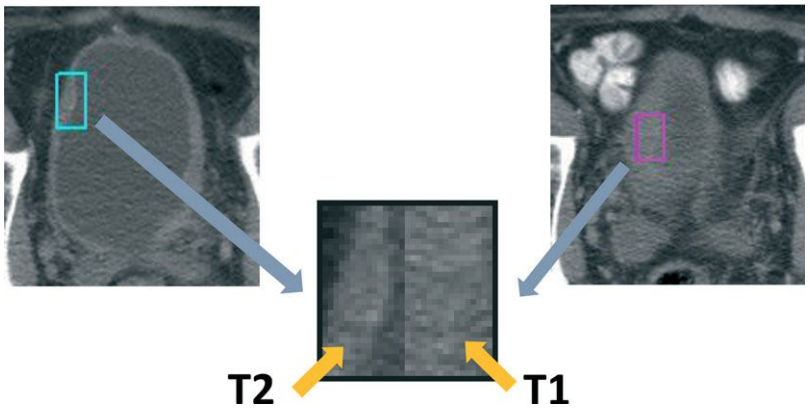


Figure 1. Example of a prepost lesion pair generated ROI. In this example the case was stage T2 in pretreatment, and stage T1 in post-treatment, resulting in a label of >T0.

Early assessment of therapeutic efficacy and prediction of failure of the treatment would help physicians decide whether to discontinue chemotherapy at an early phase and thus reduce unnecessary morbidity and improve the quality of life of the patient, and reduce costs. The ultimate goal is to improve survival for those with a high risk of recurrence while minimizing toxicity to those who will have minimal benefit.

The development of an accurate predictive model for the effectiveness of a specific therapy and clinical evaluation of the predictive model are of critical importance for patients with bladder cancer. In addition, if a patient can be reliably identified as having complete response to treatment, the treatment option of preserving the bladder may be considered, which would drastically reduce the morbidity of the patient and improve his/her quality of life as compared to the current standard treatment by cystectomy.

Pathologic evaluation performed at the time of radical cystectomy is considered a "gold standard" for estimation of treatment response. However, this method cannot be used during the course of chemotherapy. Noninvasive evaluation of the treatment response can be performed during the course of chemotherapy (after 1 or 2 cycles) with computed tomography (CT) or magnetic resonance imaging (MRI) by measuring tumor size. CT provides accurate anatomical images of the tumor and is becoming the main tool for evaluation of bladder cancer.

We are developing a computerized decision support system (CDSS-T) for monitoring of bladder cancer treatment response. Machine learning techniques are used to integrate the image information into an effective predictive model. The purpose of the CDSS-T is to provide noninvasive, objective, and reproducible decision support for identifying nonresponders so that the treatment may be stopped early to preserve their physical condition or to identify full responders for organ preservation.

DL-CNN can be used to build pattern recognition models using large image data sets (10-12). There are an increasing number of DL-CNN applications in medical imaging field for lesion segmentation, characterization, and diagnosis of diseases in different organs (13).

Cha et al. (14) proposed DL-CNN-based method for treatment response assessment of bladder cancers. In their paper, the DL-CNN was trained directly on a pre- and posttreatment set of 82 patients with 87 bladder cancers and deployed on a test pre- and posttreatment set of 41 patients with 43 cancers.

In medical imaging where training image data sets are generally small, a commonly used approach for building robust DL-CNN models is transfer learning (15). This approach uses a large data set from a different domain (for example, natural scene images) to initially train the DL-CNN. Then most of the structures and the parameters of the DL-CNN are kept fixed and only a small part of the DL-CNN is retrained with the smaller data set from the specific domain of the task at hand, for which the model is designed. This approach has shown a lot of promise in a number of medical imaging applications (16-18).

In this study we have explored different DL-CNN models for bladder cancer treatment response assessment based on transfer learning by freezing different DL-CNN layers and varying the DL-CNN structure. We also compared the DL-CNN models to radiomics-based models.

METHODS

Data Set

Pre- and posttreatment CT scans of 123 patients (with 129 total cancers) undergoing chemotherapy were collected with IRB approval. In total, 33% of patients were determined to have T0 stage cancer (complete response) after chemotherapy.

After the chemotherapy treatment, each patient underwent cystectomy. The final cancer stage after treatment was determined on the basis of the pathology obtained from the bladder at the time of the surgery. The pathological cancer stage was used as the reference standard for response to treatment: complete response (stage T0) or not complete response (stage > T0).

The CT scans were acquired with GE Healthcare LightSpeed MDCT scanners (120 kVp; 120–280 mA). The pixel size range was 0.586 to 0.977 mm and the slice thickness range was 0.5 to 7.5 mm.

The lesions on the pre- and posttreatment scans were segmented using our previously developed autoinitialized cascaded level sets system (19). ROIs of pre- and posttreatment scans of these patients were extracted from segmented lesions as 32- × 16-pixel images, and pre- and posttreatment images of patients were combined to make hybrid pre–post image pairs in the form of 32- × 32-pixel image ROIs. Figure 1 gives an example of a >T0 lesion pair and how it is generated. Multiple ROIs were extracted from pre- and posttreatment images of the lesion and combined to obtain a number of hybrid pre–post image pairs for the same lesion. Each hybrid ROI was labeled as T0 (complete

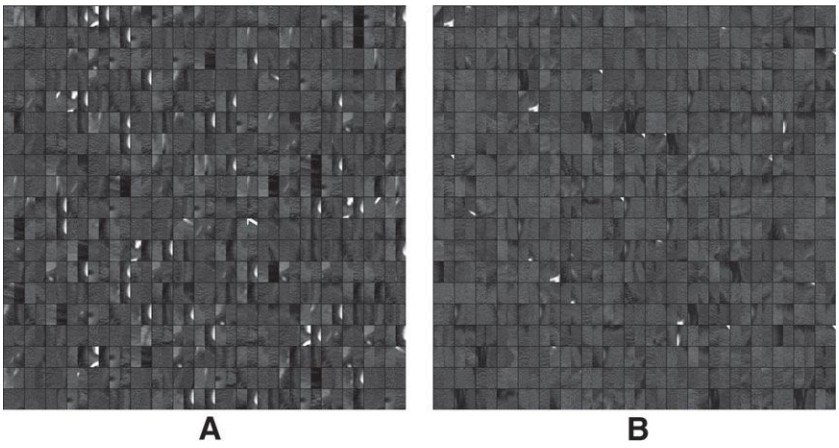


Figure 2. Subset of 6209 total regions of interest (ROIs) used in training set. Cases with complete response (T0) to treatment (A). Cases that did not fully respond (>T0) to treatment (B).

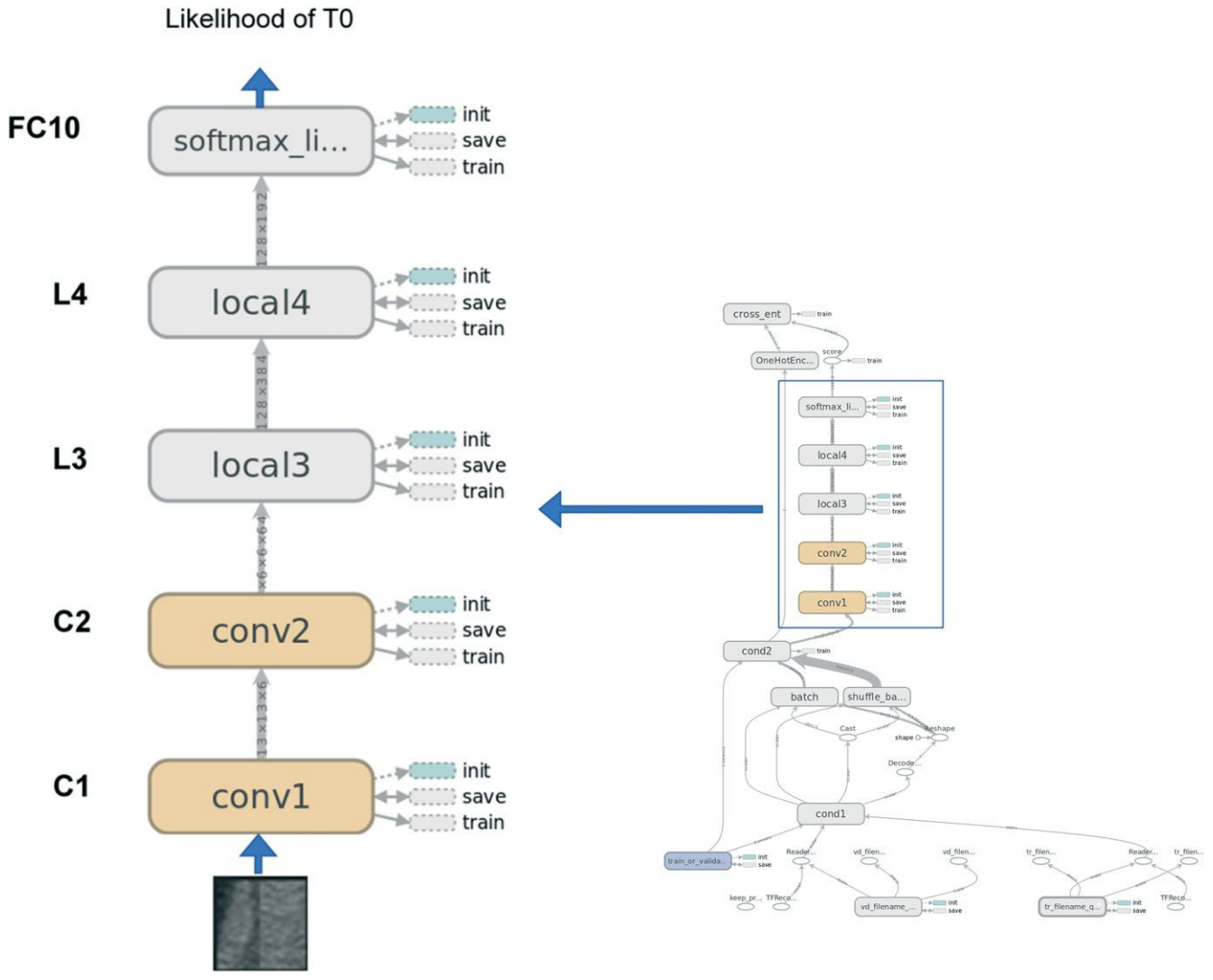


Figure 3. TensorFlow graph of base deep learning-convolutional neural network (DL-CNN) structure with different layers marked. The hybrid pre–post lesion pair ROIs were input to the DL-CNN, which then predicted a likelihood score of a complete response (T0) as an output.

Table 1. Modifications in Layers C1 and C2 for Each Structure Variation

	Base	DL-CNN-1	DL-CNN-2	DL-CNN-3
C1				
Convolution				
Size	5 × 5	5 × 5	5 × 5	5 × 5
Stride	1	1	2	1
Max Pooling				
Size	3 × 3	5 × 5	3 × 3	3 × 3
Stride	2	2	2	2
Padding	Valid	Valid	Valid	Same
C2				
Convolution				
Size	5 × 5	5 × 5	5 × 5	5 × 5
Stride	1	1	1	1
Max Pooling				
Size	3 × 3	2 × 2	2 × 2	4 × 4
Stride	2	1	1	2

response after treatment) or >T0 (the cancer did not respond completely after treatment) as determined by pathology.

The data set was split into training, validation, and test sets. The training set consisted of 77 lesions from 73 patients, where 19 lesions were stage T0, and 58 lesions were stage >T0. The 77 lesions formed 94 lesion pairs, and 6209 hybrid ROIs were generated. The validation set consisted of 10 lesions (stage T0, 5; stage >T0, 5) that formed 10 pre- and posttreatment cancer pairs and generated 521 hybrid ROIs. The test set was composed of 42 lesions from 41 patients, where 12 lesions were stage T0, and 30 lesions were stage >T0. The 42 lesions formed 54 pre- and posttreatment cancer pairs. Figure 2 displays 2 mosaics of different pre–post lesion pairs used in the training, with the left mosaic (Figure 2A) containing T0 pairs and the right (Figure 2B) containing >T0 pairs.

Two experienced radiologists, blinded to the clinical treatment outcome, also evaluated each pair of pre- and posttreatment CT scans in the test data set, displayed on 2 medical-grade monitors side by side, and provided ratings for the likelihood of the posttreatment lesions being stage T0 cancer.

Network Structures

The DL-CNN structure used in this study was based on AlexNet (10) and implemented and validated in the TensorFlow framework. The base structure of the DL-CNN consisted of 2 convolution layers (C1 and C2) followed by 2 locally connected layers (L3 and L4) and a fully connected layer (FC10). The output from the DL-CNN was trained to classify cases as fully responding (stage T0) or not fully responding (stage > T0) to chemotherapy based on the hybrid ROIs. Within C1 and C2, convolution filtering with 64 "5 × 5" kernels and a stride of 1 was performed, followed by local response normalization and max pooling with a 3 × 3 filter of stride 2. Layer L3 consisted of 64 "3 × 3" kernels, and L4 consisted of 32 "3 × 3" kernels. The output from L4 was input to the FC10, which was a softmax linear layer. The FC10 layer produced a numerical likelihood score from 0 to 1, with 0 corresponding to a stage > T0 case, and 1 corresponding to a stage T0 case. Figure 3 shows a labeled map of the DL-CNN generated by TensorBoard, a visualization tool for TensorFlow.

We first trained the DL-CNN with randomly initialized weights. We then explored the use of transfer learning. The DL-CNN with pretrained weights from the CIFAR10 image set were used. The CIFAR10 image set consists of 10 classes (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck) and 60 000 total 32 × 32 images collected by Krizhevsky et al. Each class contains 6000 images (20). We also performed alterations to the DL-CNN structure to study its effect on the DL-CNN performance. The modifications of the structures took place in layers C1 and C2, and these involved the filter size, filter

Table 2. Test AUC Values for DL-CNN Models with Modified Structures

DL-CNN Type	Base DL-CNN Structure (Random Weights)	Base DL-CNN Structure (Pretrained Weights)	DL-CNN-1	DL-CNN-2	DL-CNN-3
AUC	0.73 ± 0.08	0.79 ± 0.07	0.72 ± 0.08	0.86 ± 0.06	0.69 ± 0.09

Table 3. Test AUC Values for DL-CNN Models with Transfer Learning and Different Frozen Layers

DL-CNN Type	Base DL-CNN Structure (Pretrained Weights)	C1 Frozen	C1, C2 Frozen	C1, C2, L3 Frozen
AUC	0.79 ± 0.07	0.81 ± 0.07	0.78 ± 0.08	0.71 ± 0.08

stride, and padding type of the convolutions and max pooling performed in each layer. Three different structures were studied (DL-CNN-1, DL-CNN-2, and DL-CNN-3), and the modifications performed can be observed in Table 1.

In addition, we trained the network with one (C1) or more (C1, C2, L3) layers frozen. Freezing a layer during training prevents its weights from being altered, and it may be necessary to preserve the starting weights for some layers of the network to optimize training results (21). All of the experiments with frozen layers used the CIFAR10 transfer learning and the original DL-CNN network structure.

Training and Testing Process

The DL-CNN models were trained first for 10 000 epochs by using the training data set. For every 100 epochs, the trained DL-CNN model was deployed on the validation set. The area under the ROC curve (AUC) was calculated as a performance measure, and the validation AUC results were recorded. To reduce the likelihood of overfitting, a line plot of the validation AUC results was created and a training epoch number around where the validation AUCs peaked (usually around 2000 epochs) was selected. The final DL-CNN model was trained on the combined training set (comprising the merged training and validation sets) up to the selected epoch. The trained DL-CNN model was then deployed on the test set and the AUC was estimated. Training for 10 000 epochs for 1 experiment typically took about 8.3 hours with an NVidia GeForce GTX 1080 Ti GPU. Final training with the combined set took about 1.7 hours. Deployment on the test set took less than 1 minute per case.

Evaluation

The AUC results of our experiments were compared with those of the 2 radiologists, as well as those from 2 radiomics feature-based classification methods (RF-SL and RF-ROI) by Cha et al. (14). The radiomics-based methods involved predicting the response of cases based on the estimated changes in automatically extracted features (including morphological, gray level, and texture features) between lesions in pre- and posttreatment scans. Cha et al. (14) also evaluated the performance of a similarly structured DL-CNN. The results of the variations in the DL-CNN structure and the transfer learning schemes were compared with those of the base structure. We generated ROC curves for each experiment and used 2 statistical significance tests, ROC-kit from the University of Chicago, and the DeLong Test, to estimate the statistical significance of the differences between AUC values of the corresponding experiments. In addition, using the ROC curves, we calculated the sensitivity and accuracy of the test results at specificity of 80%, and statistical significance of the differences was also estimated. The specificity of

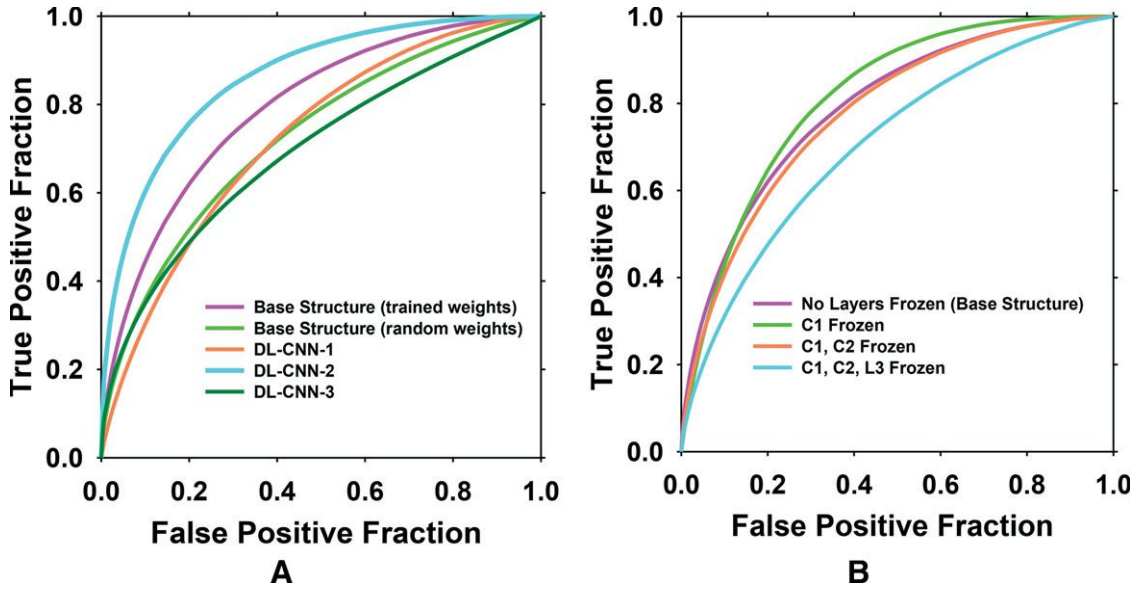


Figure 4. Test ROC curves of different DL-CNN models. ROC graph comparing base DL-CNN model (base structure) to DL-CNN models with modified structure (A). ROC graph comparing base DL-CNN model (base structure) with pretrained weights but no frozen layers to DL-CNN models with frozen layers (B).

Table 4. Test AUC Values for Radiologists and Methods Used in Cha et al. Study

DL-CNN Type	Base DL-CNN Structure (Random Weights)	Radiologist 1	Radiologist 2	DL-CNN (Cha)	RF-SL	RF-ROI
AUC	0.73 ± 0.08	0.76 ± 0.08	0.77 ± 0.08	0.73 ± 0.08	0.77 ± 0.08	0.69 ± 0.08

80% was selected by an experienced urologist (A.W.), as a possible clinically meaningful value.

RESULTS

The AUCs for our experiments are shown in Tables 2 and 3, and the ROC curves are shown in Figure 4. For the base DL-CNN structure with randomly initialized weights, the test AUC for T0 prediction was 0.73 ± 0.08. For the base DL-CNN structure, with transfer learning using CIFAR10 pretrained weights and no frozen training layers, the test AUC was 0.79 ± 0.07. The test AUCs for the DL-CNN-1, DL-CNN-2, and DL-CNN-3 modified structures (with transfer learning and no frozen layers) were 0.72 ± 0.07, 0.86 ± 0.06, and 0.69 ± 0.09, respectively. The only statistical significance difference observed was between DL-CNN-2 and DL-CNN-3 ($P = .007$, DeLong; $P = .006$, ROC-kit).

With the first layer (C1) of the base DL-CNN frozen, the test AUC was 0.81 ± 0.07. With the first 2 layers (C1 and C2) frozen, the test AUC was 0.78 ± 0.08. With the first 3 layers (C1, C2, and L3) frozen, the test AUC was 0.71 ± 0.08. None of the differences in AUC between the DL-CNN with frozen layers and the base structure with no layers frozen reached statistical significance.

Table 4 shows the AUC of the base DL-CNN with randomly initialized weights versus the radiologists and methods from the Cha et al. study (14). The AUCs of radiologist 1 and radiologist 2 were 0.76 ± 0.08 and 0.77 ± 0.08, respectively. The AUCs of the radiomics-based methods RF-SL and RF-ROI were 0.77 ± 0.08 and 0.69 ± 0.08, respectively. The network structure used in the study by Cha et al. achieved an AUC of 0.73 ± 0.08.

Table 5 shows the sensitivity and accuracy of each model at a specificity of 80%. The corresponding sensitivities ranged from 41.7% to 75.0%, while the corresponding accuracies ranged from 64.1% to 78.9%. Neither of the differences in sensitivities and accuracies between models reached statistical significance.

DISCUSSION

The results of this study show the feasibility of DL-CNN in estimating bladder cancer treatment response in CT. The DL-CNN performed better with pretrained weights from the CIFAR-10 image set than with randomly initialized weights, while the AUC from the randomly initialized weights matched that of the network structure used in the previous Cha et al. study (14). The base DL-CNN and its modified structures all performed similarly to the radiologists, and in a few cases, performing better with higher AUCs. The AUCs of the base DL-CNN and its variations were comparable to the AUCs of the radiomics-based methods from the Cha et al. study. Only 1 network variation (DL-CNN-2) resulted in a statistically significant improvement in performance compared to the base structure.

Figure 5 shows examples of pre- and postlesion pairs predicted correctly and incorrectly by the base DL-CNN with CIFAR10 weights.

The performance of the DL-CNN generally decreased as more training layers were frozen. Freezing layer C1 resulted in a slight, but not statistically significant, improvement in performance. According to a study by Yosinski et al. (22), the first layer of neural networks trained on natural images aims, in general, to capture more universal features (such as edges and curves), while proceeding layers aim to capture features more specific to the input image set (in this case, bladder lesions). As a result, allowing the first layer to train and change its weights may have minimal or adverse effects on the results of the training. Such a phenomenon may have been observed in our experiments, given the performance increase in our network with layer C1 frozen.

Similar trends were observed by Samala et al. (23) for the task of classification of malignant and benign breast masses on mammograms and tomosynthesis.

In our statistical significance tests, we found that one of our structure modifications, DL-CNN-2 (with the highest AUC value of all structures), achieved statistically significant improvement in performance compared to DL-CNN-3 (with the lowest AUC value of all structures). We will perform further testing to confirm the validity of our results and measure the performance of the structure with a larger data set.

There are limitations in this study. We are currently working with a relatively small data set in training, validation and testing of our DL-CNN models, which may also be a reason for achiev-

Table 5. Test Sensitivity and Accuracy of DL-CNN Models at a Specificity of 80%

	Sensitivity (%)	Accuracy (%)
Base Structure (Pretrained weights)	59.5%	64.1%
Base Structure (Random Weights)	41.7%	71.5%
Structure Modifications		
DL-CNN-1	50.0%	73.3%
DL-CNN-2	75.0%	78.9%
DL-CNN-3	50.0%	73.3%
Layer Freezing		
C1 Frozen	58.3%	75.2%
C1, C2 Frozen	58.3%	75.2%
C1, C2, L3 Frozen	58.3%	75.2%

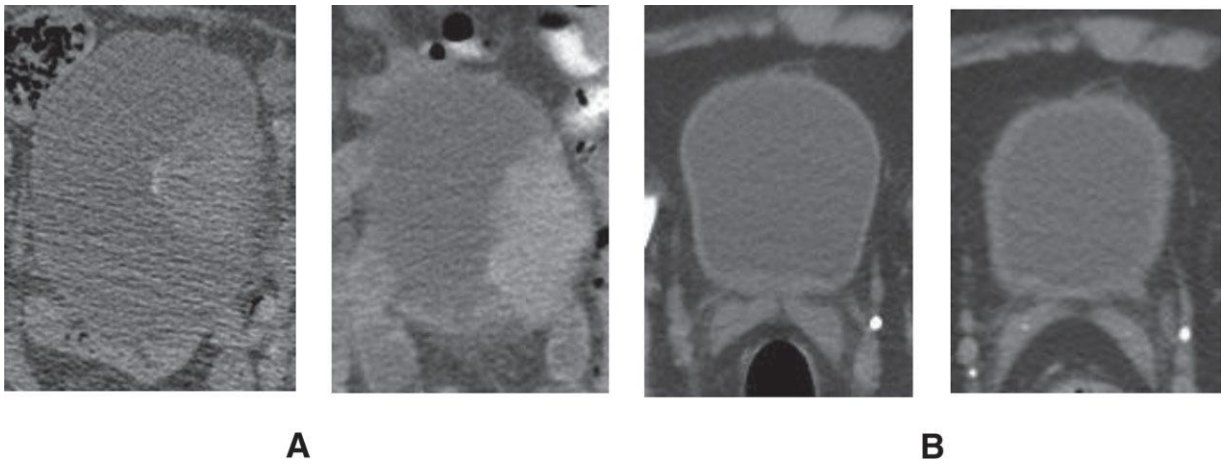


Figure 5. Examples of cases that the DL-CNN predicted correctly and incorrectly. The base DL-CNN with transferred weights correctly predicted this lesion as >T0 (A). The base DL-CNN with transfer learning incorrectly predicted this T0 lesion as >T0 (B).

ing statistical significance for only 1 comparison. In the future, we will continue to collect a larger data set with new cases (both T0 and non-T0) in our networks. Another limitation is that we have evaluations from only 2 radiologists on the test set. Additional classifications from different radiologists would be needed to study the variability in the accuracy of such readings.

Our network was trained using the CIFAR-10 data set, which produces favorable results, but is not relevant in the field of medical imaging. A better approach for training with transfer learning would be to use CT scan images, ideally bladder scans, as pretrained weights. Several networks pretrained using CT scans exist, and we may, in the future, explore the use of such networks in training with our data set.

The pixel sizes of the CT scans used in our data set vary in the range of 0.586 to 0.977 mm^2, and slice thicknesses vary from 0.5 to 7.5 mm. While the nonuniform nature of the scans may be seen as a limitation, in that it may bias the training results, learning different sizes would help the network better handle variability which would be present in real clinical applications. While scans would ideally take place under the same conditions using the same scanner, this is very difficult to achieve in clinical settings. Nevertheless, we may try in the future to match voxel sizes of scans using methods such as interpolation.

It is important to accurately assess a bladder cancer's response to treatment based on pre- and posttreatment lesion scans to determine what further treatment a patient will require, if any at all. While our current network structure has shown to classify cases with considerable accuracy, we will further improve the model and validate its generalizability in unknown cases. Because of the small data set, we used DL-CNNs of relatively small structures in this study. We will investigate if deeper DL-CNN models such as GoogLeNet Inception (24) and ResNet (25) may provide better performance when a large data set becomes available.

In conclusion, our results showed that DL-CNN can effectively predict the response of a bladder cancer lesion to chemotherapy, with many of our experiments comparing favorably to the performance of the radiologists. Adjusting the structure of the base network and freezing certain layers of the network during training may further improve the performance. This study suggests that the DL-CNN may be useful in conjunction with medical professionals as decision support for bladder cancer treatment response assessment.

ACKNOWLEDGMENTS

This work is supported by National Institutes of Health grant number U01CA179106.

Disclosures: No disclosures to report.

Conflict of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. *Key Statistics for Bladder Cancer*. 2018. https://www.cancer.org/cancer/bladder-cancer/about/key-statistics.html.
2. Sternberg CN. The treatment of advanced bladder cancer. Ann Oncol. 1995;6:113–126.
3. Fagg SL, Dawson-Edwards P, Hughes MA, Latief TN, Rolfe EB, Fielding JW. CIS-Diamminedichloroplatinum (DDP) as initial treatment of invasive bladder cancer. Br J Urol. 1984;56:296–300.
4. Raghavan D, Pearson B, Coorey G, Woods W, Arnold D, Smith J, Donovan J, Langdon P. Intravenous CIS-platinum for invasive bladder cancer – safety and feasibility of a new approach. Med J Aust. 1984;140:276–278.
5. Meeks JJ, Bellmunt J, Bochner BH, Clarke NW, Daneshmand S, Galsky MD, Hahn NM, Lerner SP, Mason M, Powles T, Sternberg CN, Sonpavde G. A systematic review of neoadjuvant and adjuvant chemotherapy for muscle-invasive bladder cancer. Eur Urol. 2012;62:523–533.

6. Advanced Bladder Cancer Meta-analysis Collaboration. Neoadjuvant chemotherapy in invasive bladder cancer: a systematic review and meta-analysis. Lancet. 2003;361:1927–1934.
7. Grossman HB, Natale RB, Tangen CM, Speights VO, Vogelzang NJ, Trump DL, deVere White RW, Sarosdy MF, Wood DP Jr., Raghavan D, Crawford ED. Neoadjuvant chemotherapy plus cystectomy compared with cystectomy alone for locally advanced bladder cancer. N Engl J Med. 2003;349:859–866.
8. Splinter TAW, Scher HI, Denis L, Bukowski R, Simon S, Klimberg I, Soloway M, Vogelzang NJ, van Tinteren H, Herr H. The prognostic value of the pathological response to combination chemotherapy before cystectomy in patients with invasive bladder cancer. J Urol. 1992;147:606–608.
9. Witjes JA, Wullink M, Oosterhof GO, de Mulder P. Toxicity and results of MVAC (methotrexate, vinblastine, adriamycin and cisplatin) chemotherapy in advanced urothelial carcinoma. Eur Urol. 1997;31:414–419.
10. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep ConvolutionalNeural Networks. Paper presented at: 26th Annual Conference on Neural Information Processing Systems 25 (NIPS 2012); December 03–08, 2012; Advances in Neural Information Processing Systems. 2012;2:1097–1105. Lake Tahoe, NV.
11. LeCun Y, Bengio Y, Hinton GE. Deep learning. Nature. 2015;521:436–444.
12. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw. 2015;61:85–117.
13. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88.
14. Cha KH, Hadjiiski L, Chan HP, Weizer AZ, Alva A, Cohan RH, Caoili EM, Paramagul C, Samala RK. Bladder cancer treatment response assessment in CT using radiomics with deep-learning. Sci Rep. 2017;7:8738.
15. Pan SJ, Yang QA. A survey on transfer learning. IEEE Trans Knowl Data Eng. 2010;22:1345–1359.
16. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging. 2016;35:1285–1298.
17. Samala RK, Chan HP, Hadjiiski LM, Helvie MA, Cha KH, Richter CD. Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms. Phys Med Biol. 2017;62:8894–8908.
18. Huynh BQ, Li H, Giger ML. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. J Med Imaging. 2016; 3:034501.
19. Hadjiiski LM, Chan HP, Caoili EM, Cohan RH, Wei J, Zhou C. Auto-initialized cascaded level set (AI-CALS) segmentation of bladder lesions on multi-detector row CT urography. Acad Radiol. 2013;20:148–155.
20. Krizhevsky A. Learning Multiple Layers of Features from Tiny Images. Toronto: University of Toronto; 2009:60.
21. Samala RK, Chan HP, Hadjiiski L, Helvie MA, Wei J, Cha K. Mass detection in digital breast tomosynthesis: deep convolutional neural network with transfer learning from mammography. Med Phys. 2016;43:6654–6666.
22. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems 27 (NIPS '14). 2014: 3320–3328.
23. Samala RK, Chan HP, Hadjiiski L, Helvie MA, Richter CD, Cha KH. Breast cancer diagnosis in digital breast tomosynthesis: effects of training sample size on multi-stage transfer learning using deep neural nets. IEEE Trans Med Imaging. 2018:1.
24. Szegedy C, Lui W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. IEEE Computer Vision and Pattern Recognition (CVPR). 2015:1–9.
25. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27–30, 2016:770–778. Las Vegas, NV.

Accuracy and Performance of Functional Parameter Estimation Using a Novel Numerical Optimization Approach for GPU-Based Kinetic Compartmental Modeling

Igor Svistoun[1], Brandon Driscoll[1], and Catherine Coolens[1,2,3]

[1]Department of Medical Physics, Princess Margaret Cancer Centre and University Health Network, Toronto, Canada; [2]Departments of Radiation Oncology and IBBME, University of Toronto, Toronto, Canada; and [3]TECHNA Institute, University Health Network, Toronto, Canada

Corresponding Author:
Catherine Coolens, PhD
Princess Margaret Cancer Centre, Department of Medical Physics,
Rm 6:306 - 700 University Avenue, Toronto, ON M5G 1Z5, Canada;
E-mail: Catherine.Coolens@rmp.uhn.ca

Key Words: DCE imaging, numerical optimization, functional analysis, GPU

Abbreviations: Dynamic contrast-enhanced (DCE), graphical processing unit (GPU), magnetic resonance imaging (MRI), arterial input function (AIF), computer processing unit (CPU), finite impulse response (FIR), infinite impulse response (IIR), pattern search (PS), differential evolution (DE), fast Fourier transform (FFT)

ABSTRACT

Quantitative kinetic parameters derived from dynamic contrast-enhanced (DCE) data are dependent on signal measurement quality and choice of pharmacokinetic model. However, the fundamental optimization analysis method is equally important and its impact on pharmacokinetic parameters has been mostly overlooked. We examine the effects of those choices on accuracy and performance of parameter estimation using both computer processing unit and graphical processing unit (GPU) numerical optimization implementations and evaluate the improvements offered by a novel optimization approach. A test framework was developed where experimentally derived population-average arterial input function and randomly sampled parameter sets $\{K_{trans}, K_{ep}, V_b, \tau\}$ were used to generate known tissue curves. Five numerical optimization algorithms were evaluated: sequential quadratic programming, downhill simplex (Nelder–Mead), pattern search, simulated annealing, and differential evolution. This was combined with various objective function implementation details: delay approximation, discretization and varying sampling rates. Then, impact of noise and CPU/GPU implementation was tested for speed and accuracy. Finally, the optimal method was compared to conventional implementation as applied to clinical DCE computed tomography. Nelder–Mead, differential evolution and sequential quadratic programming produced good results on clean and noisy input data outperforming simulated annealing and pattern search in terms of speed and accuracy in the respective order of 10^{-8}%, 10^{-7}%, and $\times 10^{-6}$%). A novel approach for DCE numerical optimization (infinite impulse response with fractional delay approximation) was implemented on GPU for speed increase of at least 2 orders of magnitude. Applied to clinical data, the magnitude of overall parameter error was $<$10%.

INTRODUCTION

Obtaining a better understanding of a (personalized) tumor or disease microenvironment is quickly becoming a driving force in a whole range of medical scenarios from earlier disease diagnosis to image-based assessment of treatment efficacy (1). In this context, dynamic contrast-enhanced (DCE) imaging is increasingly used to help quantify vascular and tissue properties as to inform on the functionality and dynamic behavior of the disease and/or normal tissue. In terms of tissue perfusion and permeability, this is typically achieved with the additional use of tracer kinetic models that describe the flow of contrast agents through the tissue (2).

DCE computed tomography (CT) and magnetic resonance imaging (MRI) have been widely investigated, and despite their obvious differences in methodology to measure dynamic contrast enhancement curves, they share the same parametric analysis approach: both use low-molecular-weight contrast agents and as such they share mostly the same pharmacokinetic models that are applied after the imaging signal is converted to contrast concentration data (3). The delivery of the contrast agent to the organ or region of interest (eg, a tumor) is reflected in the arterial input function (AIF). Using the contrast enhancement curves in the organ or region of interest as a response on the AIF, an estimation of the tracer kinetic model parameters can be obtained. An example of this would be the widely used 2-compartmental modified Tofts model (2).

Whereas increasing efforts are in place to help standardize the acquisition and analysis methods of DCE imaging in both CT

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2018.00048

Table 1. Tofts Model Parameters

Variable	Description	Units
C_t	Tissue concentration of contrast agent as a function of time	HU
C_a	AIF representing the arterial concentration of contrast agent as a function of time	HU
K_{trans}	Transfer constant from blood plasma into the EES	mL/g/min
K_{ep}	Transfer constant from EES back to the blood plasma	mL/g/min
V_b	Blood volume per unit of tissue	mL/g
t	Time variable	second
τ	Time delay from time of contrast injection to contrast arriving at region of interest	second
HCT	Hematocrit—fraction of red blood cells in blood. Value of 0.4 is used during this investigation.	Fraction

(4) and MRI (5), the solution of these tracer kinetic models is not necessarily trivial and requires an optimization method to solve for parameters in heterogeneous volumetric data. The effect of image noise and voxel-based analysis has also been reported on, showing a marked improvement in parameter robustness that can be achieved by balancing preprocess filtering with information loss (6). Regardless, parameters must be extracted given the nonuniform, discrete, limited-time measurements. Implementing the parameter estimation algorithm involves many other design decisions including choice of data processing rate, continuous-to-discrete system mapping approach, and numerical optimization algorithm. To the best of our knowledge, no investigations have been reported on the impact of the optimization method used on resulting parametric maps. Yet, it is well-known from other areas of research that significant differences can be found in between optimization methods in their ability to adequately resolve multiple variables simultaneously.

Given the large amount of data involved in processing DCE parametric maps, it is further increasingly important that these processes are as automated as possible to allow for useful integration into clinical workflows with a nearly real-time experience. Current implementations of kinetic models rely on manual or semiautomated estimations of the fractional delay in contrast arrival time at the region of interest. Not only is this a time-limiting factor for a fully automated workflow, it will be shown that lack of inclusion of this parameter in the optimization process creates larger estimation errors. For this reason, moving the optimization processes to a graphical processing unit (GPU) offers known speed improvements over standard computer processing unit (CPU) implementation of a fully inclusive optimization approach.

Having recently shown the improved correlation between CT- and MRI-based perfusion parameters (7) when using a common analysis platform to process DCE data regardless of the imaging modality, the purpose of this paper is now to (1) quantify the effects of system design choices (eg, processing sampling rate) and noise (both aliasing and background) present in the data on accuracy and speed of various CPU and GPU numerical optimization implementations and (2) to obtain a better understanding of parameter accuracy in clinically relevant DCE-CT data.

METHODS

Continuous Time Model and Problem Statement

Various models exist to describe contrast solute exchange of iodine or gadolinium-based DCE imaging methods. The modified Tofts model is by far the most widely implemented technique and as such it was felt worthwhile to investigate the design variations to better understand the largest available literature of pharmacokinetic metrics reported. The modified Tofts model describes a linear time-invariant first-order system. Data are acquired by the scanner, which can be expressed as tissue concentration function $C_t[n]W_c[n]$ and AIF $C_a[n]W_c[n]$ for $n \in \{\text{nonuniform discrete time points}\}$. $W_c[n]$ is a rectangular window function that takes on the value 1 at $0 < n \leq c \,/\, T$ and 0 otherwise, where T is the sampling period. The window function represents the fact that acquisition of measurements stops after a certain time $= c$ seconds.

The 2-compartmental model of tissue enhancement that takes into account contributions from intravascular and the interstitial space (which is what's measured by the scanner) is given by the following linear time-invariant system (2):

$$\begin{aligned} C_t(t) &= C_a(t) \times \left[\left(\frac{K_{trans}}{1 - HCT} \right) e^{-K_{ep}(t-\tau)} u(t-\tau) + V_b \delta(t-\tau) \right] \\ &= C_a(t) \times H(t) \end{aligned} \quad (1)$$

The parameters used in the model are summarized in Table 1. The continuous-time system must be approximated by a discrete-time system to carry out the computation of the output, making use of the discrete measurements – like the ones acquired from a scanner – as input to the system. The field of digital signal processing (DSP) offers many methods to accomplish this. It therefore helps to examine the model in the frequency domain by applying the continuous-time Fourier transform resulting in equation (2).

$$\begin{aligned} C_t(j\Omega) &= C_a(j\Omega) H(j\Omega) \\ H(j\Omega) &= \left[\frac{\left(\frac{K_{trans}}{1 - HCT} \right)}{K_{ep} + j\Omega} + V_b \right] e^{-\tau j\Omega} \\ H(j\Omega) &= [H_1(j\Omega) + H_2(j\Omega)] H_3(j\Omega) \end{aligned} \quad (2)$$

Examining the model in frequency allows the overall system to be broken down into the following 3 simpler parts: summation of constant gain V_b with a first-order system, $H_1(j\Omega)$ and an overall delay element $H_3(j\Omega)$. Note, the delay element is required because the measurement site is upstream to the input and it will take some amount of time for the contrast agent to arrive at the measurement site. Frequency domain analysis offers several discretization approaches—mainly finite impulse response (FIR) approximation and infinite impulse response (IIR).

The objective is to find parameters K_{trans}, K_{ep}, V_b, τ given the measurements $C_t[n]W_c[n]$ and $C_a[n]W_c[n]$. This is done using constrained nonlinear numerical optimization attempting to minimize the sum of square errors.

$$f(K_{trans}, K_{ep}, V_b, \tau) = \sum_{n=0}^{c/T} \left(\hat{C}_t[n]W_c[n] - C_t[n]W_c[n]\right)^2$$
$$\begin{aligned} 0 < \ & K_{trans} \leq 5 \\ 0 < \ & K_{ep} \leq 10 \\ 0 \leq \ & V_b \leq 1 \\ 0 \leq \ & \tau \leq c \end{aligned} \qquad (3)$$

Where $\hat{C}_t[n]$ represents samples of system output for a given set of parameters K_{trans}, K_{ep}, V_b, τ and a particular AIF$C_a[n]W_c[n]$. The summation limits reflect the fact that our measurements are cut off after $n = c/T$ samples. The optimization constraints were chosen to be within reasonable physical limits, and to aid certain optimization algorithms converge quicker.

Note that to compute $\hat{C}_t[n]$ the model (2) must be discretized. The discretization step introduces its own set of errors. In particular the choice of sampling rate and continuous-to-discrete mapping approach affect how well the discrete-time system resembles the continuous-time system at the range of frequencies of interest. The accuracy of fitted parameters depends greatly on the accuracy of the system approximating $\hat{C}_t[n]$.

Discrete Approximation Methods and Sampling Rates

There are 2 main methods evaluated in this paper to approximating the continuous-time system by a discrete system. The first method is the FIR using the window approach to filter design and the second is IIR using bilinear transformation (also known as Tustin's method). How well the discrete system approximates the continuous-time system depends largely on the sampling rate used during approximation (see online supplemental Appendix).

Although acquiring data at very high sampling rates is not clinically feasible, this section discusses the ideal signal processing case. Two factors affect the selection of appropriate sampling rate, both of which depend on the cutoff frequency - i.e., the point in the frequency domain where the signal is zero. Nyquist requires sampling rate to be at least 2x the cutoff frequency to avoid aliasing error (8). The second factor for selecting sampling rate is to ensure the discrete-time system matches the continuous system closely up to the cutoff frequency. Even if Nyquist rate criteria is satisfied, the discrete approximation may not match the continuous system up to the cutoff frequency and additional error may be introduced. In certain circumstances the acquired data should be up-sampled and processed at a higher rate to avoid introducing this additional error.

When the signals are not band limited and do not reach zero past any frequency, like in this case, a cutoff frequency is selected based on desired precision and computational feasibility. A low pass filter (LPF) is used prior to digitizing the signal to attenuate components past the cutoff frequency. The degree of attenuation in the stop band of the LPF depends on the noise floor, which is the background noise that is technically infeasible to get rid of in the system.

In the ideal simulation case where population average AIF is computed and then in turn used to generate signals, the noise floor is due to errors in floating point arithmetic. Studying the signals involved in the Tofts model, the cutoff frequency for the ideal case can be determined based on when the frequency components reach below the noise floor level (as if the low pass filter was applied). It was determined that to achieve precision on the order of single floating point arithmetic error, sampling rate of 3500 Hz is required (more detail can be found in the online supplemental Appendix).

Efficient Fractional Delay Approximation

As mentioned earlier, there is a delay between the time when the contrast agent is injected and when it arrives at the measurement site. This can be expressed as a continuous-time system $H_3(j\Omega)$. To account for this delay, the DCE analysis implementation could ask the user to visually evaluate the curves and supply the delay value when the tissue response curve begins to increase and optimize the other 3 kinetic parameters of the model; this approach would be tedious for a user to perform repeatedly for each voxel, error prone, as visual analysis could differ between users, and error prone if the user specifies the same delay value for a large physical area, which does not account for the fractions of seconds that it took for tracer to arrive at a further upstream site. Another approach to account for the delay could involve analyzing tissue response curves automatically based on the curve slope to determine the onset time, and then optimize the other 3 kinetic parameters (6). Heuristic search based on slope is susceptible to noise if there are noisy spikes before the true onset or if the onset occurs between samples. For this DCE analysis implementation, it was decided to numerically optimize all 4 kinetic model parameters, including the delay.

The discretization approaches, FIR and IIR, described in previous sections can deal with only delay by whole number of samples. For example if the system's sampling period is 1 s, only integer delay may be computed. This coarse approximation of delay can lead to poor fit in other parameters—K_{trans}, K_{ep}, V_b. The sampling rate can be increased to allow for a broader range of delay values—for example, 10 Hz would allow for any delay that is a multiple of 0.1 s—but at a proportional cost to memory requirements and processing time. This problem can be alleviated with the use of fractional delay approximation, which allows for estimation of the output signal for any floating point delay value (9). In our investigation the first-order Thiran filter considerably improved the results with negligible additional run-time cost. The delay in seconds can be implemented by the following 2 operations: Delay By Whole # of Samples

Table 2. Data Sets Analyzed

Name	Samples	Duration	Gaussian Noise
Data set 1	200 samples 1-second interval	200 seconds	None
Data set 2	9 samples 2-second interval	209 seconds	Added: $\mu = 0$
	19 samples 5-second interval		$\sigma = 6HU$
	9 samples 10-second interval		
DCE-CT Brain Scan	9 samples 2-second interval	209 seconds	Estimated: $\mu = 0$
	19 samples 5-second interval		$\sigma = 6HU$
	9 samples 10-second interval		

$N = \lfloor \tau / T \rfloor$ followed by Fractional Delay $FD = \tau / T - \lfloor \tau / T \rfloor$. The first-order filter is provided in equation (4)

$$H_{thiran}(z) = \frac{a_1 + z^{-1}}{1 + a_1 z^{-1}} \qquad a_1 = \frac{1 - FD}{1 + FD} \tag{4}$$

Testing Framework Design and Investigation Goals

The following were the investigation goals when designing the test framework:

1. Derive theoretical background for the ideal case to validate algorithm implementation and calibrate values for the basic numerical optimization algorithm parameters.
2. Investigate and demonstrate the effects of discretization method, sampling rates used during processing, noise, and fractional delay approximation filters on the resulting accuracy of the kinetic model parameters.
3. Investigate achievable accuracy of kinetic parameters extracted from clinical data set.

An experimentally derived functional form of population-average AIF (10) was sampled at 3500 Hz based on theoretical discussion in the section with the heading "Discrete Approximation Methods and Sampling Rates" in this paper. A uniformly distributed pseudorandom number generated was used to sample parameters K_{trans}, K_{ep}, V_b, τ from the minimization constraints range (8). The tissue curves were then calculated for each parameter set by a discrete-time system approximating the continuous model at 3500 Hz.

The ideal generated tissue curves proceed to a measurement stage where ideal high sampling rate signals are decimated and additional Gaussian white noise may be added. A summary of data sets analyzed and their canonical names used throughout the paper are summarized in Table 2.

In this setup, the ground truth parameters for data sets 1 and 2 are known. The generated signals at 3500 Hz represent the ideal case and it should be possible to recover the original parameters used to generate the signals to within tolerances of single floating point precision arithmetic. Running numerical optimization on the ideal signals was used to calibrate and configure the algorithms, as well as validate all additional custom code. The optimization algorithms evaluated in the simulation include: sequential quadratic programming (SQP) (11), downhill simplex (Nelder–Mead) (12), pattern search (PS) (13), simulated annealing (SA) (14), and differential evolution (DE) (15). Matlab (v2015b) optimization and global optimization toolbox's implementation of SQP, Nelder–Mead, PS, and SA were used. Price et al. implementation of DE was used for the experiments (15).

The algorithm parameters and values configured during calibration are described in Table 3. To overcome problems of local minima, SQP, Nelder–Mead, PS, and SA were initialized to

Table 3. Algorithm Parameters

Algorithm	# Start Points	Max Iterations	Exit Criteria	
			TolFun	TolX
SQP	32	1000	10^{-8}	10^{-8}
Nelder–Mead	32	1000	10^{-8}	10^{-8}
CUDA Nelder–Mead	32	1000	10^{-8}	10^{-8}
PS	32	1000	10^{-8}	NA
SA	32	1000	10^{-8}	NA
DE	64	1000	10^{-8}	NA
CUDA DE	512	1000	10^{-8}	NA

Table 4. Algorithm Calibration at 3500 Hz: Median of Percent Error and Timing

Algorithm	Overall %Error	Time (sec./voxel)
SQP	$8.97 \times 10^{-6} \pm 4.66 \times 10^{-7}$	1030±16
Nelder–Mead	$5.69 \times 10^{-8} \pm 2.32 \times 10^{-9}$	522 ± 23.7
CUDA Nelder–Mead (IIR)	$1.07 \times 10^{-7} \pm 1.27 \times 10^{-8}$	$(14.5 \pm 9.82) \times 10^{-3}$
DE	$3.27 \times 10^{-7} \pm 2.20 \times 10^{-8}$	1230 ± 12.3
CUDA DE (IIR)	$3.35 \times 10^{-7} \pm 2.59 \times 10^{-8}$	$(34.0 \pm 5.33) \times 10^{-3}$
PS	2.79 ± 1.04	13300 ± 284
SA	3.85 ± 1.23	2960 ± 32.5

quasi-random starting points generated using the Halton sequence (16). A quasi-random sequence was used to avoid the probability of generating tight clusters of starting points that could arise when using a distribution generated by a pseudo-random number generator. Each algorithm was configured to exit based on the maximum number of iterations, a minimum change in objective function (TolFun), and a minimum change in estimated parameter (TolX) to avoid infinite run-time. DE operates on a population of candidates that can conceptually be considered as the number of starting points. Furthermore, the DE objective function–based exit criteria was chosen such that the algorithm would exit when the difference between minimum and maximum values of the current objection function across the population was found to be below the TolFun threshold. All algorithm parameters were tweaked experimentally until the accuracy of the results were within the maximum accuracy allowable by a single floating point precision arithmetic or the results produced by the algorithm did not show any further improvement indicating numerical optimization algorithm limitations.

After calibration of the algorithm parameters (TolX, TolFun, etc.) and after having established an accuracy baseline, changes to objective function calculation in the form of adding fractional delay, changing discretization methods, and sampling rate were implemented. The validity of such code changes was verified by ensuring that at ideal processing rates, the accuracy matched the baseline accuracy. Then, data sets 1 and 2 were processed and the performance of each change was analyzed for its impact on accuracy and speed.

Analyzing the impact results, 2 algorithms were ported to CUDA to run on the GPU. In case of DE, the population size was increased to 512 compared to its CPU counterpart to take advantage of the multithreaded GPU architecture and have each optimization converge faster. Data sets 1 and 2 and an additional clinical DCE-CT brain scan were analyzed using this numerical optimization implementation under an institutionally approved REB protocol. The analysis was performed on CPU and GPU.

In terms of underlying hardware and timing analysis, the simulations were performed on several Xeon E5-2690 CPUs, and for comparison, on Tesla K40m GPU. A high-throughput computing cluster HTCondor was used; however, to narrow the analysis to only the algorithm performance, the overhead of data serialization, network transfer, and start-up time on remote nodes were discarded—only the main algorithm run-time was recorded.

In summary, earlier theoretical discussion led us to design for the ideal case under a single floating point precision. The algorithms were calibrated to perform within tolerances specified by the ideal case. With established confidence in correctness of implementation and calibration parameters, 2 artificial data sets were generated and run through the testing framework, while several other parameters were changed including the sampling rate and discretization method used on the Tofts model and the use of fractional delay approximation versus rounded delay for estimating the contrast arrival time at the site. Because, the second data set had the same sampling and noise profile of a scanned DCE-CT brain scan data set, when numerical optimization was carried out on the clinical data set, a conclusion on the accuracy of the extracted parameters could be determined.

RESULTS

Algorithm Calibration

The percent relative error for each parameter is defined as $\in = 100 \times |x_{true} - x_{approx}| \; / \; |x_{true}|$. The percent relative errors for each of the 4 parameters was combined into a single array of errors and the mean statistic along with 95% confidence interval was calculated and summarized in Table 4. Note that these calibrations are processed at very large sampling rates as discussed in the section with the heading "Discrete Approximation Methods and Sampling Rates" in this paper.

The SQP algorithm hits its optimization accuracy limit at percentage errors 1 and 2 orders of magnitude below DE and Nelder–Mead algorithms; decreasing tolerances and increasing sampling rates did not produce better results for SQP. The likely reason for this has to do with the fact that SQP is a gradient approach and the function is quite flat around the optimal point. This conclusion lead us to investigate nongradient-based approaches. From these approaches, Nelder–Mead and DE performed quite well. However, PS and SA could not be configured to achieve optimization values anywhere close to other algorithms; further modifications of algorithm parameters (such as increasing the number of starting points) produced marginally better results at a cost of much higher run-times. Because of these calibration results, long run-times and poor-accuracy PS and SA algorithm were discarded as viable numerical optimization candidates for this particular problem.

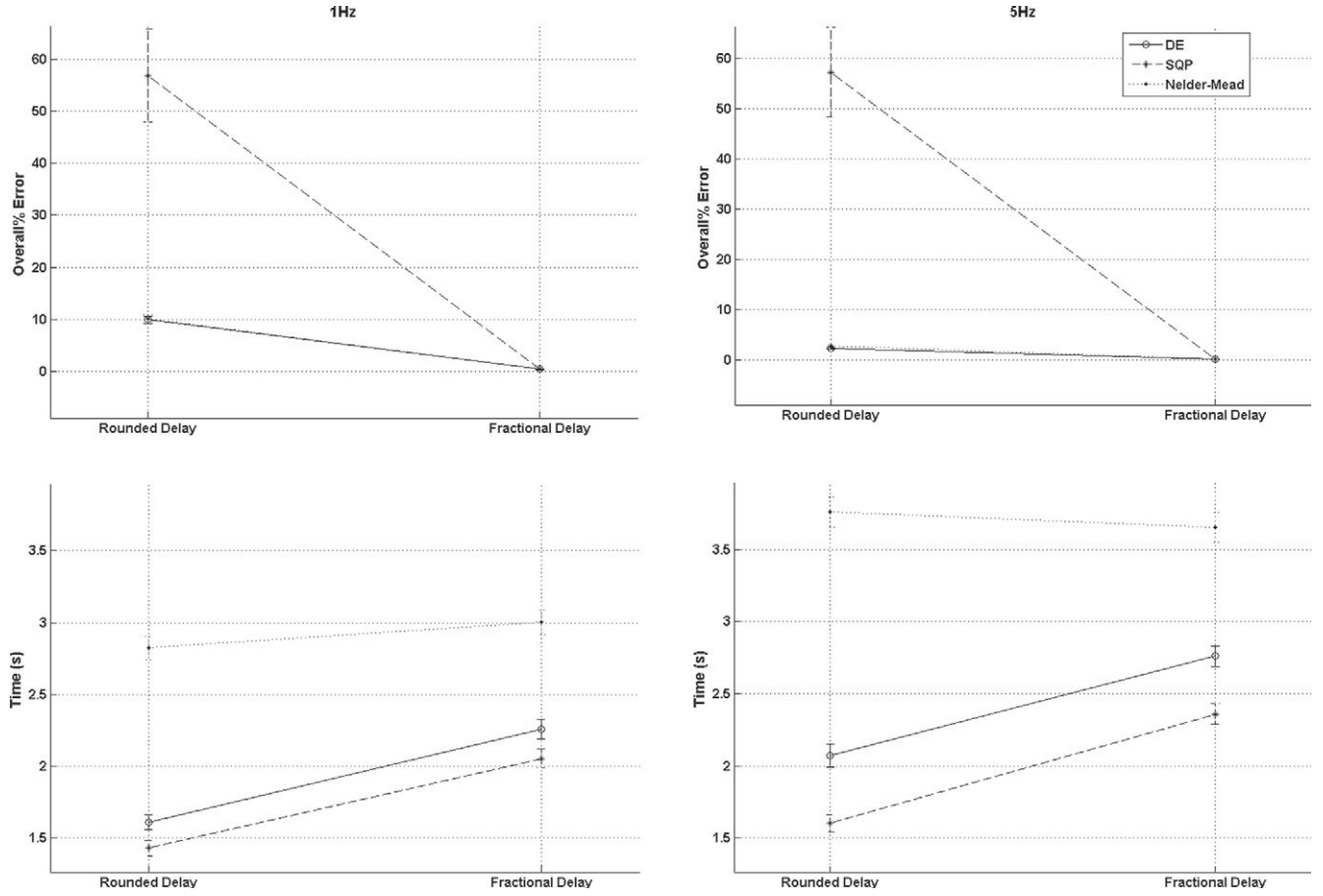


Figure 1. Data set 1. Impact of rounded delay vs fractional delay analysis processed at 1-Hz and 5-Hz infinite impulse response (IIR) on the mean overall %error and mean run-time per voxel.

Fractional Delay Analysis

Figure 1 shows the mean relative percent errors and the mean run-time sec./voxel with 95% confidence interval for results extracted from data set 1. IIR approximations at 1 Hz and 5 Hz were used. Using rounded delay approximation, SQP performs very poorly with the mean of the overall relative error at 56.9% regardless of the sampling rate used to process the data. One explanation for this is SQP exits criteria based on the objective function change is triggered because the gradient is constant for a range of delay values when rounding is used. Similar problems with rounded delay can be seen with DE and Nelder–Mead algorithm. With fractional delay approximation, instead of rounding, the error was reduced from 10% to 0.4% for DE and Nelder–Mead algorithms, and from 56.9% to 0.4% for the SQP algorithm.

An alternative to approximating the fractional delay is to use higher sampling such as 5 Hz. Somewhat surprisingly, SQP showed no improvement when using rounded delay compared to 1 Hz with the error still at 56.9%. The other numerical optimization algorithms did show a significant improvement where the overall error was 2.83%. However it should be noted that increasing the sampling rate by some factor increases the memory requirement by the same factor. Better accuracy can be achieved at 1 Hz with fractional delay approximation (0.4%) than at 5 Hz and using rounded delay (2.83%).

Figure 2 shows the fractional delay analysis run on data set 2, which has coarse, nonuniform sampling and additional $\mu = 0, \sigma = 6HU$ Gaussian noise added. Similar behavior can be observed for the SQP algorithm—it exits prematurely, causing very large errors (56.9%). Because of large amount of noise there (aliasing and artificial), there was no significant improvement in accuracy when using fractional delay approximation. It should be noted that in this case, the addition of the fractional delay approximation did not add significant amount of overall computation time.

In general, fractional delay approximation greatly improves accuracy of gradient-based numerical optimization algorithms such as SQP. When the noise profile of the data permits, it also improves accuracy significantly without having to process at higher sampling rates. Because of this, fractional delay approximation was added to all further analysis simulations and to the algorithms used to analyze clinical data.

Discrete Approximation and Sampling Impact Analysis

Figure 3 shows the means of relative percent errors across all parameters, as well as the mean logarithm of sec./voxel with

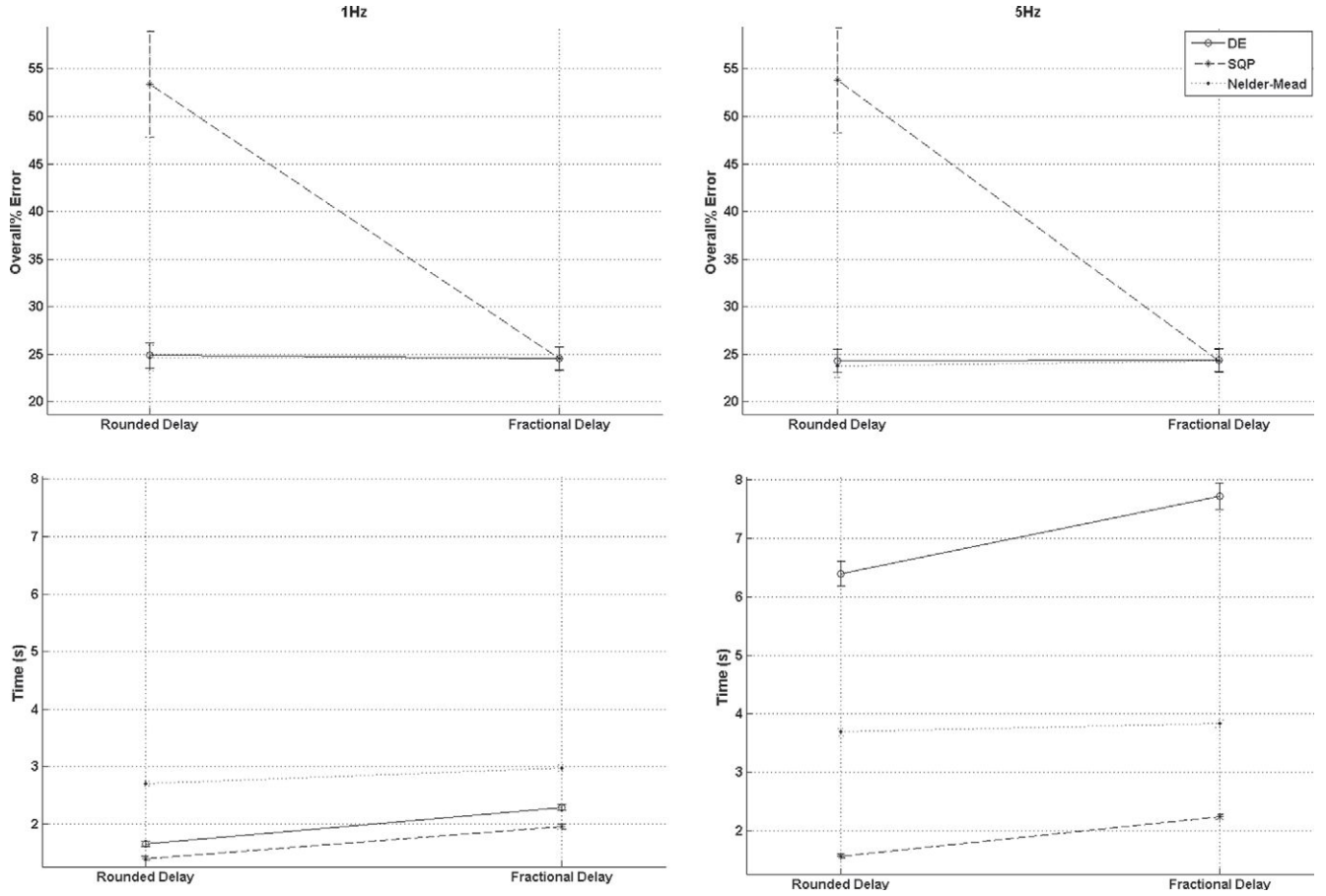


Figure 2. Data set 2. Impact of rounded delay vs fractional delay analysis processed at 1-Hz and 5-Hz IIR on the mean overall % error and mean run time per voxel.

95% confidence interval. The simulation compares 2 discrete approximation methods—FIR and IIR—and the effect of up-sampling data set 1 and using the more accurate discrete approximations that are a direct result of higher sampling rate. Fractional delay approximation was used during this analysis.

In terms of accuracy, the algorithms perform almost identically across sampling rates and discretization methods. Data set 1 was sampled at 1 Hz; the high-frequency information is lost forever regardless of how much the signals are up-sampled. However, if the signals were processed at 1 Hz, additional error would be introduced owing to the discrete-system poorly approximating the continuous-time system at this low rate. Figure 3 shows that accuracy can be increased by up-sampling the data and processing at higher rates. It is also evident that IIR approximation of the Tofts continuous-time system is more accurate than the FIR approximation at lower sampling rates, as the accuracy achieved by IIR approximation at 1 Hz is slightly better than the overall accuracy achieved by FIR approximation at 5 Hz. The mean of errors for each individual parameter when using IIR approximation at 1 Hz is {0.27%, 0.10%, 8.81%, 0.13%} for the parameters $\{K_{trans}, K_{ep}, V_b, \tau\}$, respectively. The overall mean error across all parameters is 2.33%. By switching to IIR approximation at 5 Hz, the overall mean of errors reduces to 0.40%, or individually, the error for each parameter becomes {0.46%, 0.16%, 0.84%, 0.12%}, showing large improvements for *Vb* parameter as a result of changing discretization method and increasing the sampling rate.

The run-time for the algorithm is shown as a log plot. For all sampling rates, IIR runs faster than FIR. The reason for this has largely to do with the fact that for this particular system, the IIR can be implemented in a single loop over the input data, so the complexity is $O(M)$, where M is the size of the signal. On the other hand, direct convolution requires 2 nested loops and has complexity $O(M^2)$. When signal size is large (such as when higher sampling rate is used), convolution implementation can be sped up by zero-padding the signals, computing the fast Fourier transform (FFT), multiplication of frequency bin values, and IFFT (17), in which case the complexity is $O(Nlog(N))$, where N is the size of padded signals. The implementation used during simulation uses the FFT approach, which handles larger signals much better than convolution. The algorithm complexity related to input size is evident in the timing plot, where FIR versions increase steadily as the sampling rate (and hence signal size) grows, whereas the IIR versions remain relatively flat.

The combination of better scalability as a result of algorithm complexity and the lower memory footprint requirement

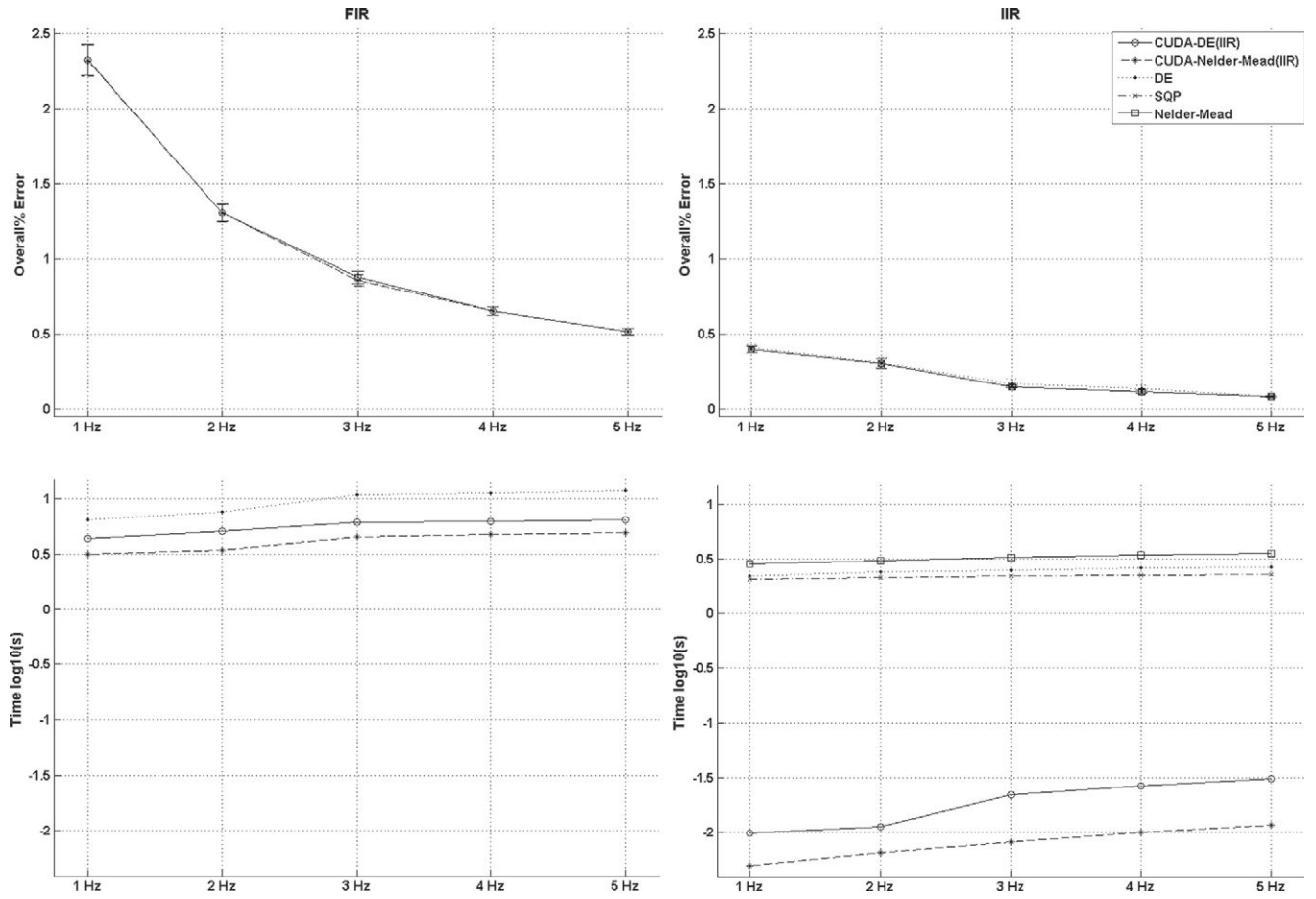


Figure 3. Overview of the impact of choice of sampling and discretization method on mean percentage overall error and mean run-time per voxel for Data set 1.

owing to better accuracy at lower sampling rates were the main reasons for using IIR approximation of the system in the CUDA implementation of DE and Nelder–Mead algorithms. The highly optimized CUDA implementation of the numerical optimization algorithms ran 2 orders of magnitude faster than their CPU counterparts.

Data set 2 was sampled nonuniformly, coarsely (average sampling rate 0.18) and had additional Gaussian noise ($\mu = 0, \sigma = 6HU$). Although the accuracy improvements from increased sampling and IIR approximation are very small, they are still evident. This analysis conveys the fact that data sets such as these need to be processed at only 1 Hz, as no further accuracy improvements can be gained by up-sampling to ensure the discrete-time system better approximates the continuous-time system. As a result of this analysis, the IIR approximation was chosen as the best discretization approach for this problem.

Figure 4 shows the results of the error analysis for data set 2 as a result of a changing the data sampling times. The resulting mean percentage error in parameter estimation was the smallest for the 1-s interval sampling interval and it increased with the increasing sampling rate. The clinical scan intervals varied depending on which part of the enhancement curve was being measured and the percentage errors therefore roughly correspond to the error values closest to the 3- and 5-s sampling intervals.

GPU Implementation and Clinical Data Analysis

Discrete approximation and sampling impact analysis showed that regardless of the optimization algorithm, IIR filter approximation produced more accurate results at lower sampling rates. In addition, fractional delay approximation allows for greater accuracy at lower sampling rates. Owing to excellent calibration accuracy, Nelder–Mead and DE, using IIR approximation and fractional delay filter, were chosen to be implemented in CUDA to run on the GPU. The calibration results from Table 4, along with identical accuracy compared to CPU counterparts (Figures 3 and 4), serve as verification that the algorithm implementation in CUDA is correct.

The best and fastest implementation (CUDA Nelder–Mead, with IIR filter and fractional delay approximation) was used to analyze a clinical DCE-CT brain scan. By analyzing CT scan areas that should contain a uniform CT number value, it was determined that the scanner may be adding as much as $\sigma = 6HU$ noise to the data. The noise was assumed to be Gaussian distributed (18) and the same population AIF was used as for the simulated curves. From earlier analysis on data set 2, which had

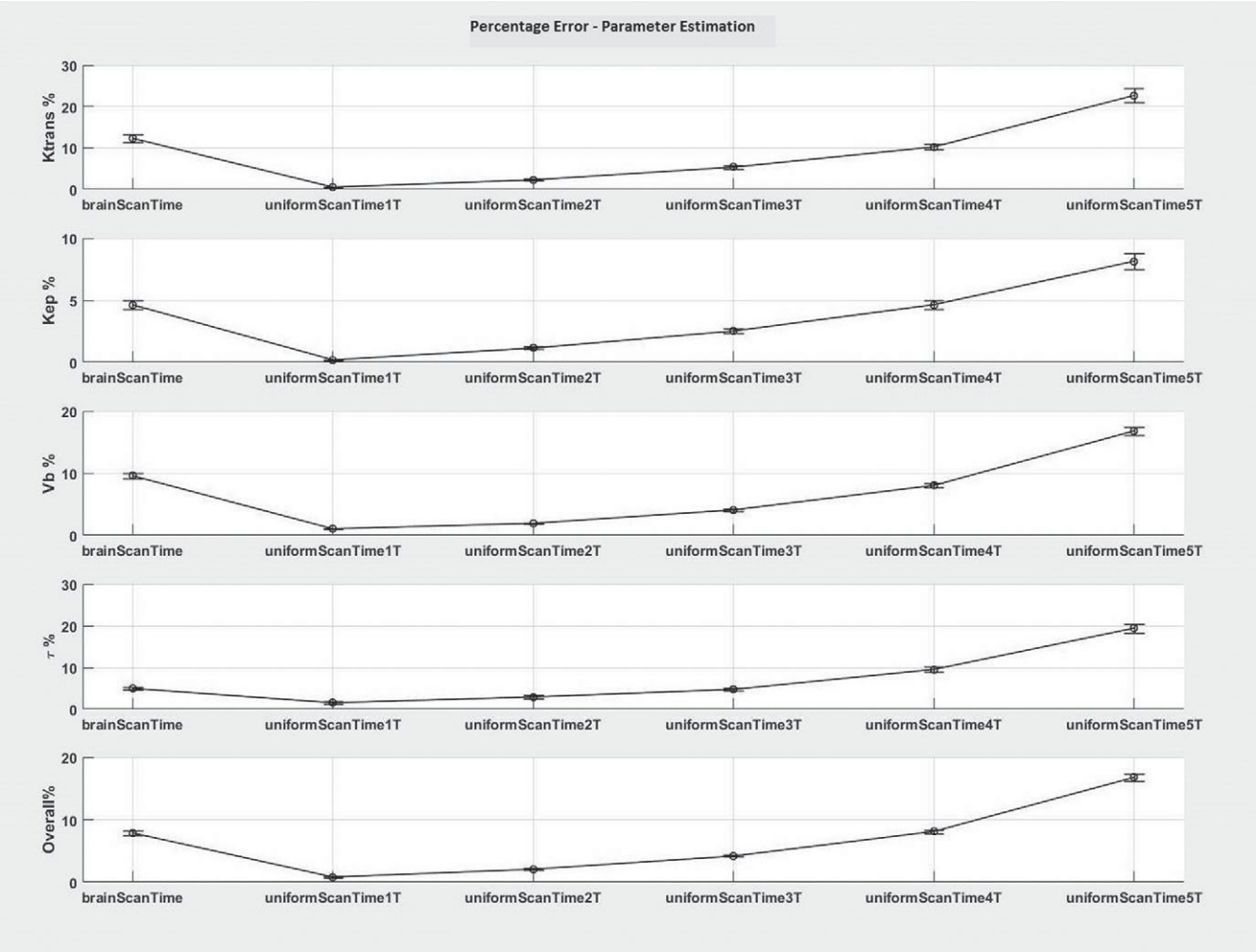


Figure 4. Data set 2, sampling analysis. Impact of data sampling on parameter estimation accuracy for (K_{trans}, K_{ep}, V_b, τ).

the same sampling and noise profiles as this CT data set, it can be concluded that the overall accuracy of parameters estimated from the CT data set is less than 10%.

Figure 5 shows a volume rendering of Vb parameter on the left, and the onset delay parameter rendering color coded such that red corresponds to earlier onset time and blue corresponds to later onset time.

Figures 3 and 4 show the GPU-based algorithm achieves speed improvements of 2 orders of magnitude compared with their CPU counterparts when run on generated data. Tables 5 and 6 show speed improvement when processing CT brain scan data. The first row is the baseline CPU implementation that uses FIR discretization of the Tofts model. The second row shows a modest speed increase because of changing the discretization to IIR. Finally the benefits of implementing the algorithm to run on a GPU are shown in the last row.

DISCUSSION AND CONCLUSIONS

Numerical optimization algorithms were carried out by designing for the ideal signal processing case at single floating point precision accuracy limits. Nelder–Mead, DE, and SQP produced good results under ideal conditions, achieving overall relative error 5.69×10^{-8}%, 3.27×10^{-8}%, and 8.97×10^{-6}%, respectively. SA and PS were found to be unsuitable for this problem because the lowest overall relative error that could be achieved was 3.85% and 2.79%, respectively.

The algorithms were designed and implemented to extract parameters from data sets with a wide range of sampling and noise profiles—ranging from the ideal and clinically infeasible data sets without noise to noisy and sparsely sampled CT brain data sets. To accomplish this, the thresholds for exit criteria were chosen to be of the order of 10^{-8}%. For very noisy data sets, this most likely creates a large amount of unnecessary processing that costs extra time; however, that is the trade-off to be able to achieve high accuracy for low-noise data sets as well. In cases of high-noise data sets, the numerical optimization exit is triggered when change in candidate parameter drops below threshold, rather than objective function target threshold. This is why for DE, the exit criteria were based on thresholding the difference between minimum/maximum objective function values across

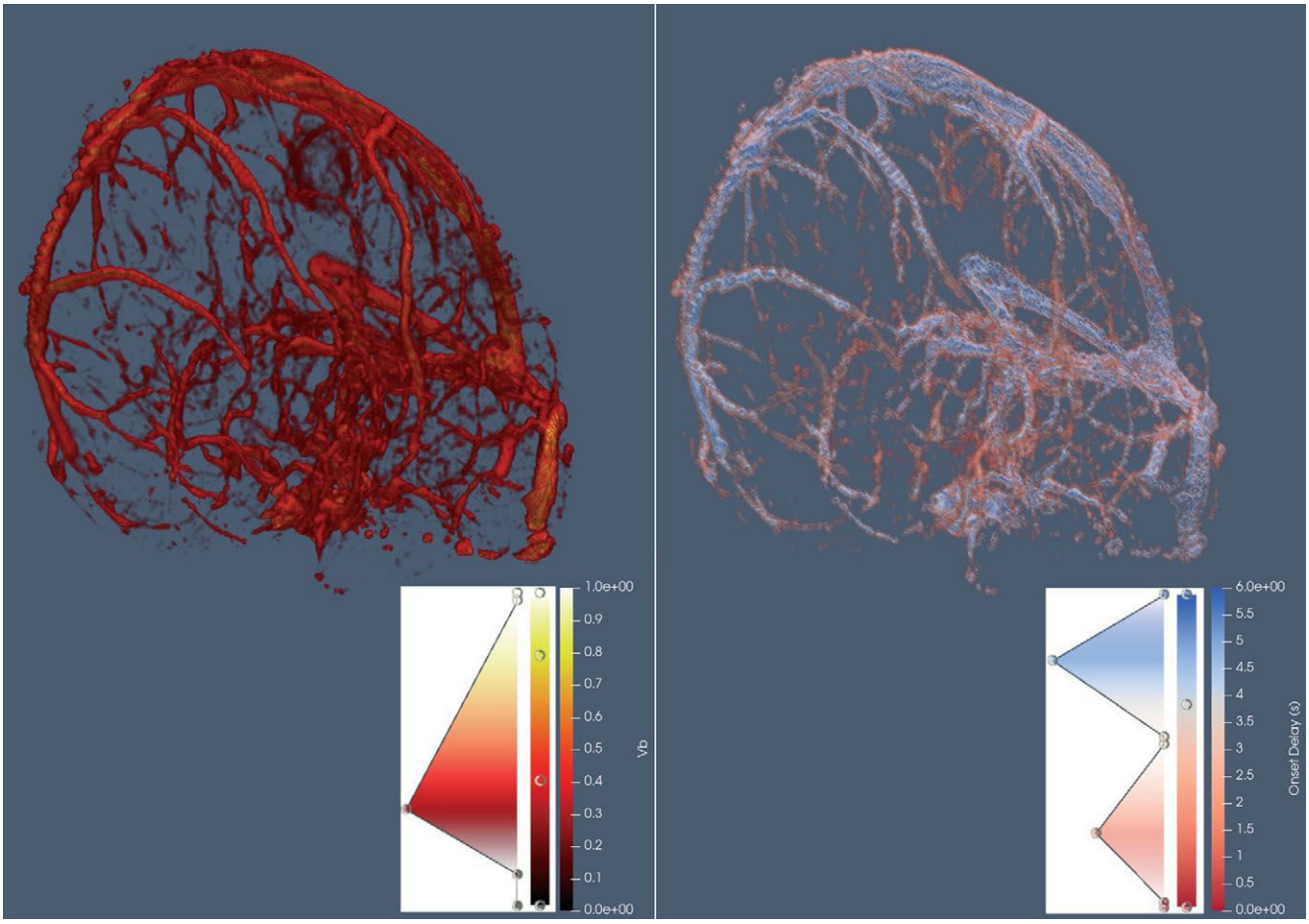


Figure 5. Volume rendering of Vb (left) and onset delay (right) parameters.

the population. Furthermore, numerical optimization algorithms that find local minima (compared to algorithms designed with global optimization in mind such as DE) were restarted many times at different initial starting points. Although the continuous objective function described in equation (3) may not have multiple minima, the discrete implementation of the objective function has many regions that would cause a numerical optimization algorithm to exit without reaching a point that would result in a better fit. For example, if rounded delay is used at 1-Hz sampling, the objective function is constant for all $\tau \in (0, 0.5)$, creating a saddle point which could cause numerical optimization to exit. This is especially evident in the gradient-based approach early termination summarized in Figures 1 and 2. Therefore as many as 32 starting points were used; using fewer starting points yielded poorer accuracy in the ideal optimization case. Having designed an algorithm that is capable of achieving best results in terms of accuracy for a very wide range of data, and a framework under which to conduct tests, it is possible to design a faster algorithm (by increasing thresholds of the exit criteria) that is able to achieve best results for the specific clinical data set.

Once numerical optimization algorithms were working to within designed tolerances of single floating point precision,

Table 5. Nelder–Mead Numerical Optimization CPU vs GPU Run-Time CT Brain Scan

Algorithm	Mean Time sec./Voxel	Relative Speed
CPU FIR 1 Hz	4.37	1.0
CPU IIR 1 Hz	3.05	1.4
CUDA IIR 1 Hz	0.0026	1680.8

Table 6. DE Numerical Optimization CPU vs GPU Run-Time CT Brain Scan

Algorithm	Mean Time sec./Voxel	Relative Speed
CPU FIR 1 Hz	2.51	1.0
CPU IIR 1 Hz	1.93	1.3
CUDA IIR 1 Hz	0.0068	369.1

experiments were conducted to vary other data processing steps and digital signal processing filters. It was shown that using fractional delay approximation filter stabilized gradient-based numerical optimization approaches and allowed the algorithm to produce accurate results instead of terminating early. Furthermore, fractional delay approximation allowed the discrete-time approximation for the Tofts model at lower sampling rates.

It was also shown that IIR discrete approximation of continuous-time Tofts model produces more accurate results at lower sampling rates. The recursive filter implementation has lower complexity compared to FIR discrete approximation, which requires convolution. This translates to lower memory footprint and faster processing times.

The clinical DCE-CT brain scan volume of interest contains just over 6 million voxels to analyze, after delineating and discarding areas outside the patient and bone. Combination of the 2 conclusions above led to an efficient port of the CPU-based algorithms into CUDA to run on the GPU. The framework can be used independent of image segmentation and run on every voxel or within a specific region of interest. The improvements in correlation between CT- and MRI-based measurements of tumor perfusion patients when a common analysis platform is used falls outside the scope of this article but is being reported on elsewhere (4).

To obtain entire brain perfusion maps required 4.3 hours (based on run-times in Table 5) on a single GPU; the same computation would take 179 days when processing on a single CPU (based on run-times reported in Table 6). If volume of interest is narrowed down further, for example, to only the tumor and surrounding tissue, which span 5 cc or just over 100,000 voxels, then kinetic model parameters can be computed in 4.3 min. Several orders of magnitude improvements such as these were also reported by Wang et al. (17) who achieved an even better 0.00025 s/voxel (compared to 0.0026 s/voxel) computation times using the block-FFT approach (FIR approximation of the Tofts model) on a less powerful GPU than Tesla K40. It should be noted that the implementation used for this paper used 32 starting points (effectively attempting to optimize each voxel 32 times to ensure global minimum) and stringent exit criteria. During CUDA code optimization attempts, it was found that the largest remaining barrier to even further speed optimization was noncoalesced memory access as a result of the delay parameter τ. In particular, on NVIDIA GPUs, the best speed can be achieved when the following holds: if a thread N reads memory location M, then thread N + 1 reads memory location M + 1 for all threads executing within a scheduled block. When implementing the delay which offsets the index of variables being read/written, coalesced memory access optimization does not apply, causing performance decrease.

A test framework such as this can further be used to determine the sampling rate required to process clinical data and gauge the magnitude of error that should be expected from the computed parameters, as well as calibrate numerical optimization algorithms to ensure best possible accuracy has been achieved.

Supplemental Materials

Supplemental Appendix: http://dx.doi.org/10.18383/j.tom.2018.00048.sup.01

ACKNOWLEDGMENTS

This work was supported by NSERC Discovery Grant #354701 and OICR operating grant #P.IT.020.

Disclosures: No disclosures to report.

Conflict of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. Jaffray DA, Chung C, Coolens C, Foltz W, Keller H, Menard C, Milosevic M, Publicover J, Yeung I. Quantitative imaging in radiation oncology: an emerging science and clinical service. Semin Radiat Oncol. 2015;25:292–304.
2. Tofts PS, Brix G, Buckley DL, Evelhoch JL, Henderson E, Knopp MV, Larsson HB, Lee TY, Mayr NA, Parker GJ, Port RE, Taylor J, Weisskoff RM. Estimating kinetic parameters from dynamic contrast-enhanced T(1)-weighted MRI of a diffusable tracer: standardized quantities and symbols. J Magn Reson Imaging. 1999;10:223–232.
3. Driscoll B, Keller H, Jaffray D, Coolens C. Development of a dynamic quality assurance testing protocol for multisite clinical trial DCE-CT accreditation. Med Phys. 2013;40:081906.
4. Coolens C, Driscoll B, Foltz W, Svistoun I, Sinno N, Chung C. Unified platform for multimodal voxel-based analysis to evaluate tumour perfusion and diffusion characteristics before and after radiation treatment evaluated in metastatic brain cancer. Br J Radiol. 2018:20170461.
5. Huang W, Chen Y, Fedorov A, Li X, Jajamovich GH, Malyarenko DI, et al. The impact of arterial input function determination variations on prostate dynamic contrast-enhanced magnetic resonance imaging pharmacokinetic modeling: a multicenter data analysis challenge. Tomography. 2016;2:56–66.
6. Coolens C, Driscoll B, Chung C, Shek T, Gorjizadeh A, Menard C, Jaffray D. Automated voxel-based analysis of volumetric dynamic contrast-enhanced CT data improves measurement of serial changes in tumor vascular biomarkers. Int J Radiat Oncol Biol Phys. 2015;91:48–57.
7. Coolens C, Driscoll B, Foltz W, Pellow C, Menard C, Chung C. Comparison of voxel-wise tumor perfusion changes measured with dynamic contrast-enhanced (DCE) MRI and volumetric DCE CT in patients with metatstatic brain cancer treated with radiosurgery. Tomography. 2016;2:325–233.
8. Mani R, Oppenheim AV, Willsky AS, Nawab SH. Solutions manual, Signals & systems, Second edition. 462 p.
9. Laakso TI, Valimaki V, Karjalainen M, Laine UK. Splitting the unit delay [fir/all pass filters design]. Signal Processing Magazine IEEE. 1196;13:30–60.
10. Parker GJM, Roberts C, Macdonald A, Buonaccorsi GA, Cheung S, Buckley DL, et al. Experimentally-derived functional form for a population-averaged high-temporal-resolution arterial input function for dynamic contrast-enhanced MRI. Magnetic Resonance in Medicine. 2006;56(5):993–1000.
11. Nocedal J, Wright SJ. Numerical Optimization. 2nd ed. New York: Springer; 2006.
12. Nelder JA, Mead R. A simplex method for function minimization. Comput J. 1965;7:308–313.
13. Audet C, Dennis JEJ. Analysis of generalized pattern searches. SIAM J Optim. 2002;13:889–903.
14. Press WH. Numerical Recipes: The Art of Scientific Computing. 3rd ed. Cambridge: Cambridge University Press; 2007.
15. Price K, Storn RM, Lampinen JA. Differential Evolution: A Practical Approach to Global Optimization: Verlag Berlin Heidelberg: Springer; 2006.
16. Halton JH. Radical-inverse quasi-random point sequence. Commun ACM. 1964; 7:701–702.
17. Wang H, Cao Y. GPU-accelerated voxelwise hepatic perfusion quantification. Phys Med Biol. 2012;57:5601–5616.
18. Coolens C, Breen S, Purdie TG, Owrangi A, Publicover J, Bartolac S, Jaffray DA. Implementation and characterization of a 320-slice volumetric CT scanner for simulation in radiation oncology. Med Phys. 2009;36:5120–5127.

TOMOGRAPHY® RESEARCH ARTICLE

A Web-Based Response-Assessment System for Development and Validation of Imaging Biomarkers in Oncology

Hao Yang, Xiaotao Guo, Lawrence H. Schwartz, and Binsheng Zhao

Department of Radiology, Columbia University Medical Center, New York, NY

Corresponding Author:
Hao Yang, MA
Department of Radiology, Columbia University Medical Center,
710 West 168th Street, B26, New York, NY 10032, USA;
E-mail: yh2588@cumc.columbia.edu

Key Words: quantitative imaging biomarkers, response assessment, web-based image platform

Abbreviations: Picture archiving and communication system (PACS), personal computer (PC), digital imaging and communications in medicine (DICOM), dynamic link library (DLL), java native interface (JNI), Health Level Seven (HL7), java message service (JMS), enterprise java beans (EJB), web access to DICOM objects (WADO), retrieve information for display (RID), internet protocol (IP), unique identifier (UID), annotation and image markup (AIM), extensible markup language (XML)

ABSTRACT

Quantitative imaging biomarkers are increasingly used in oncology clinical trials to assist the evaluation of tumor responses to novel therapies. To identify these biomarkers and ensure smooth clinical translation once they have been validated, it is critical to develop a reliable workflow-efficient imaging platform for integration in clinical settings. Here we will present a web-based volumetric response-assessment system that we developed based on an open-source image viewing platform (WEASIS) and a DICOM image archive (DCM4CHEE). Our web-based response-assessment system offers a DICOM imaging archiving function, standard imaging viewing and manipulation functions, efficient tumor segmentation and quantification algorithms, and a reliable database containing tumor segmentation and measurement results. The prototype system is currently used in our research lab to foster the development and validation of new quantitative imaging biomarkers, including the volumetric computed tomography technique, as a more accurate and early assessment method of solid tumor responses to targeted and immunotherapies.

INTRODUCTION

The use of imaging biomarkers to monitor responses of tumors to treatment has attracted increasing interest in recent years. Despite the accelerated pace of new drug discoveries, and the availability of treatment options in oncology, methods for assessing tumor responses remain almost unchanged over the past few decades, that is, using tumor diameter to gauge tumor change with therapy (1, 2). This is particularly challenging for targeted and immune therapies, as efficiencies of these therapies may be better reflected by tumor density changes than by tumor size shrinkage.

Researchers, including us, have been developing novel quantitative response-assessment methods including volume measurements and the use of more complex radiomic features to measure tumor changes (3). To foster the development and validation of quantitative imaging biomarkers, we developed a portable response-assessment system, based on an open-source WEASIS (4, 5). This system has a PACS-like user interface, which allows radiographic images to be viewed and manipulated efficiently. We integrated our homegrown segmentation tools into this system to facilitate efficient and accurate tumor segmentation and quantification.

The portable response-assessment system consists of a database server and a response-assessment application. The database server stores and manages tumor segmentation and measurement results, whereas the application consists of the following key components: (1) a WEASIS viewer module that allows the program's user to open, display, and manipulate radiological images, (2) an algorithm module that integrated our tumor segmentation algorithms and editing tools, and (3) a database module that allows users to communicate with the database server. The WEASIS response-assessment application is installed on each of the PCs in the lab.

However, shortcomings of the portable response-assessment system are obvious: (a) the application is hard to maintain, upgrade, and distribute and (b) (deidentified) DICOM images need to be transferred to and stored in each PC for tumor measurements. To address these 2 shortcomings of the portable response-assessment system, we upgraded our response-assessment system by tuning the response-assessment application so that it was web-based, and by adopting DCM4CHEE as the DICOM image archive. DCM4CHEE is a free, stable, feature-rich DICOM image archive (6). In the Methods section of this report, we will explain in detail how we designed and implemented our web-based response-assessment system.

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2019.00006

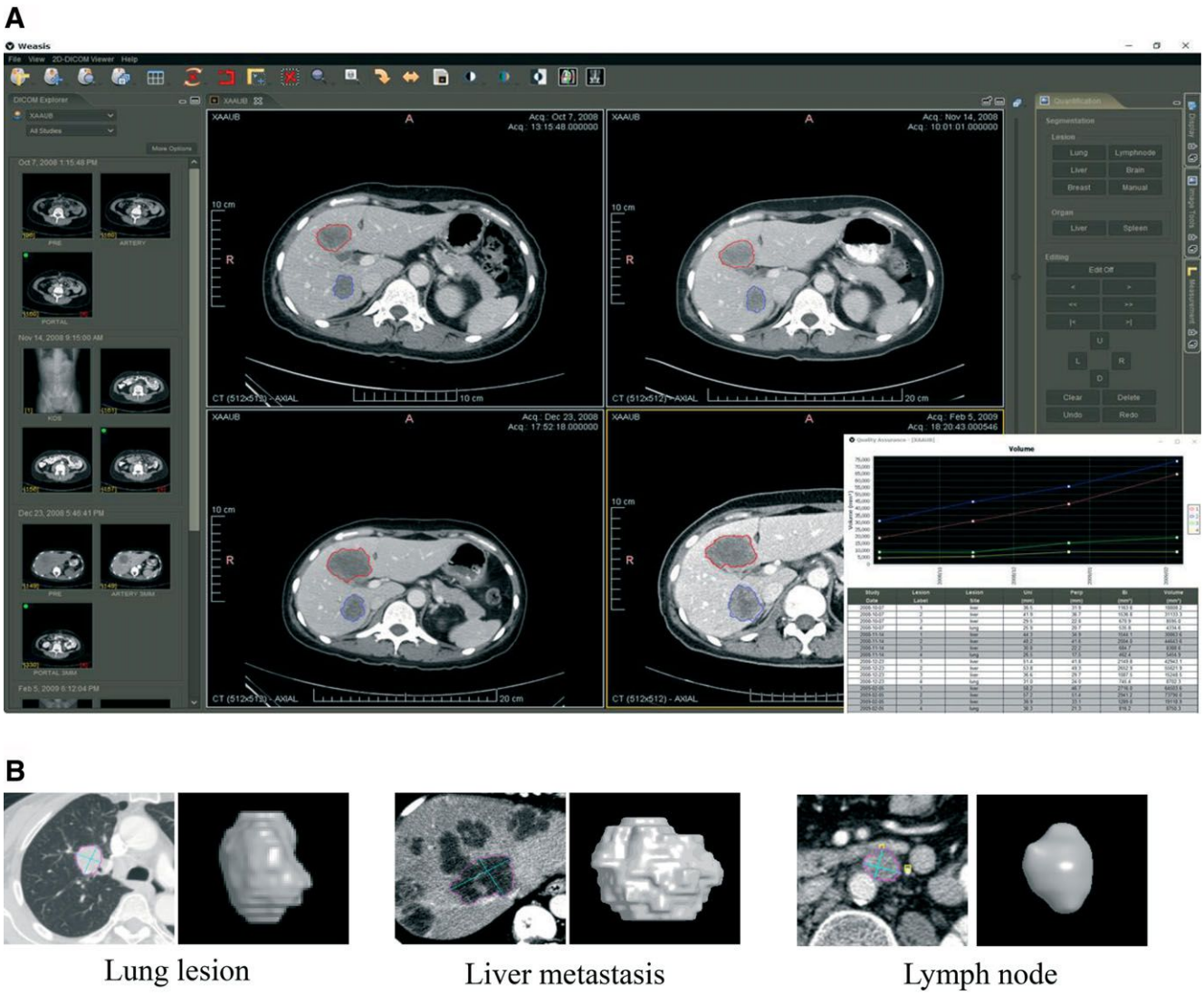


Figure 1. Web-based response assessment system with integrated tumor segmentation and quantification algorithms (A). Lung, liver, and lymph node segmentation algorithms integrated into the WEASIS response assessment system (B).

METHODS

Segmentation Algorithms and Editing Tools

As shown in Figure 1, our web-based response-assessment system has several segmentation algorithms, such as: a lung nodule segmentation algorithm (7); a liver tumor segmentation algorithm (8); a lymph nodule segmentation algorithm (9); and a brain tumor segmentation algorithm (10). The initialization procedure for each algorithm has been implemented and integrated into the system. The system also provides editing tools that are designed to efficiently modify and optimize the segmentation contours. It consists of some basis operators, such as refine, smooth, expand, and shrink functions, which can automatically modify the segmentation contours based on energy-minimization segmentation methods. It also provides propagation functions, which can propagate a segmentation result from 1 slice to the neighboring slices while automatically finding the optimal contour for each slice.

All the algorithms are written in C++ or Matlab computer language and compiled into dynamic link libraries (DLLs). Because the WEASIS platform is programmed in JAVA language, we use the Java Native Interface (JNI) to call these algorithms/functions in the libraries.

Segmentation Results and Storage Methods

Lesions will be semiautomatically delineated using the integrated 3D lesion segmentation tools developed in our research lab. Once the lesions are segmented, lesion diameters and volumes (as well as other radiomic features once the feature extraction methods are integrated into the platform) can be computed automatically. Our tumor segmentation results are binary images in which a background voxel has a value of 0 (zero) and a lesion voxel has a value of 1 (one). The compressed binary images and tumor measurement values of UNI (the maximal in-plane diameters), BI (product of the maximal diameter and its maximal perpendicular diameter), and

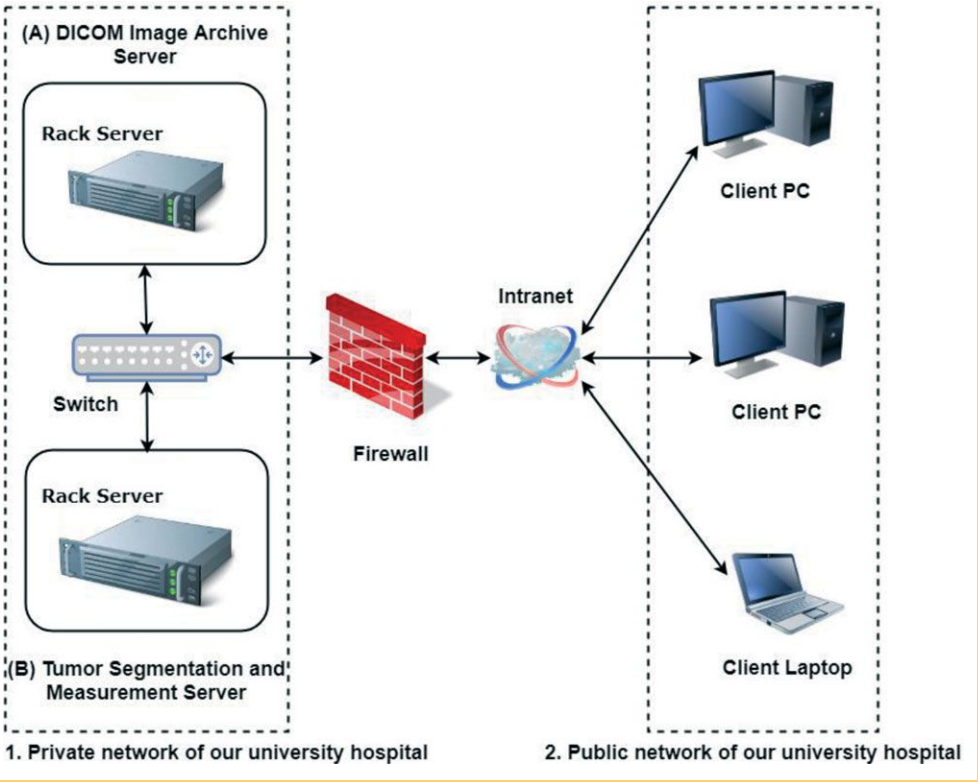


Figure 2. The hardware architecture diagram of our web-based response assessment system working in the intranet of our university hospital.

volumetry are sent to and stored in a relational database that is an industry standard. The database structure we designed is published in our previous paper, which reported our portable response-assessment system (4).

System Hardware Architecture

Figure 2 shows the hardware architecture diagram of our web-based response-assessment system. It consists of the following 2 rack servers: (1) a DICOM image archive server, hosting a DCM4CHEE and a web-based response-assessment application and (2) a dedicated database server storing tumor segmentation and measurement results in a MySQL database. The web-based response-assessment system resides in the intranet of our university hospital. The two servers are inside the private network of our university hospital, and public access is prevented by a firewall; while the client PCs, where users of the response-assessment system work, are within the public network of our university hospital. The deidentified DICOM images are stored in the image archive server and are remotely accessed by client PCs through a web browser and a web-based response-assessment application.

System Software Framework

Figure 3 shows the major components of the web-based response-assessment system, which is divided into the following 4 layers: an interface layer, an application layer, a service layer,

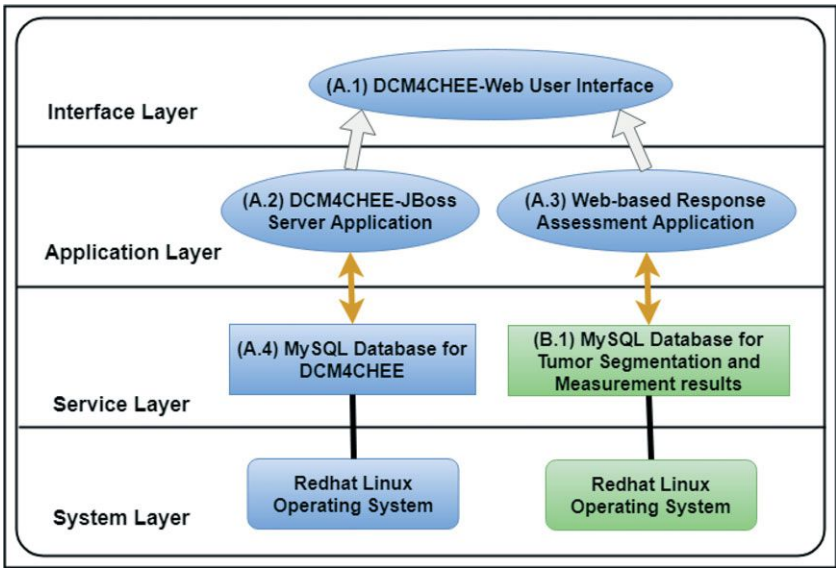


Figure 3. The layered software framework of the web-based response assessment system.

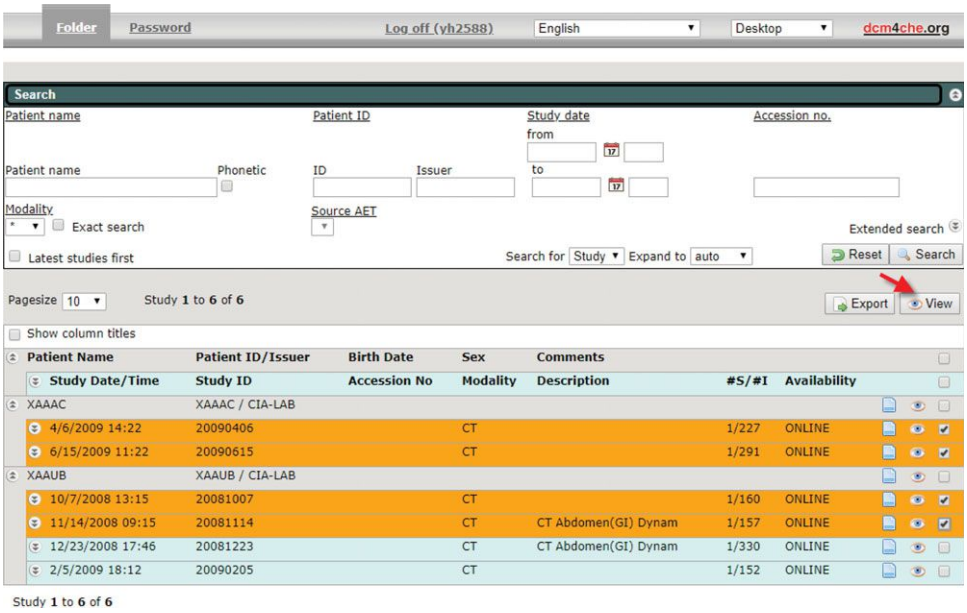


Figure 4. User interface of DCM4CHEE-Web and the link (arrow) to launch response assessment application.

and a system layer. The components in the DICOM archive server are: (A.1) a DCM4CHEE–web user interface, (A.2) a DCM4CHEE-JBoss server application, (A.3) a web-based response-assessment application, and (A.4) a MySQL database for DCM4CHEE. The component in the tumor segmentation and measurement server is a MySQL database for tumor segmentation and measurement results (B.1). We will now describe the details of each component in the web-based response-assessment system.

DCM4CHEE–Web User Interface (A.1). DCM4CHEE is a collection of open-source applications and utilities for managing and archiving DICOM imaging. It was developed in the Java programming language. The DCM4CHEE application uses the DICOM, HL7 services and interfaces to provide storage, retrieval, and workflow of DICOM imaging.

The DCM4CHEE–web user interface (shown in Figure 4) runs entirely in web browsers of client PCs. It can search for patients or studies, browse the archived DICOM information listed in a patient-study-series-image layout, and launch the response-assessment application.

DCM4CHEE-JBoss Server Application (A.2). The DCM4CHEE-JBoss server application consists of a collection of open-source applications and utilities that have been developed in the Java programming language for improved performance and portability. It contains the Health Level 7 (HL7) and Digital Imaging Communication in Medicine (DICOM) services and interfaces that are required to provide storage, retrieval, and workflow to a healthcare environment. A DCM4CHEE-JBoss server application is prepackaged and deployed within the JBoss application server. By taking advantage of many JBoss features, such as JMS (Java Message Service), EJB (Enterprise Java Beans), Servlet Engine, etc., and assuming the role of several IHE (Integrating the Healthcare Enterprise) actors for the sake of interoperability, the DCM4CHEE-JBoss server application provides the following services: (1) DICOM Storage, acting as an archive to store DICOM images to standard file systems, with compression if necessary; (2) DICOM Query/Retrieve, querying the archive for DICOM images, and retrieving them; and (3) WADO (Web Access to DICOM Objects) and RID (Retrieve Information for Display), supporting web access to the archived data.

Web-Based Response-Assessment Application (A.3). The web-based response-assessment application is based on WEASIS, a versatile open-source DIOM viewer. The framework of the response-assessment application is explained in detail in Yang et al. (4). The response-assessment application can be easily packaged for portable distribution or web-based distribution. In our system, we use the web-based distribution.

The web-based response-assessment application is hosted by the DCM4CHEE-JBoss in the DICOM archive server, and launched by the DCM4CHEE-Web. It does not persistently retain user information. Thus, the database of DCM4CHEE and the database of tumor segmentation and measurement keep exactly the same user information. In other words, to register a user, the user's information should be saved into both databases. This is a prerequisite for the web-based response-assessment application to be able to retrieve tumor segmentation and measurement results and review tumor contours using the WEASIS viewer.

MySQL Database for DCM4CHEE (A.4). The MySQL database for DCM4CHEE manages all the user information of DCM4CHEE and DICOM information, for example, user credentials, user access rights, and the path of a DICOM image in the DICOM storage.

MySQL Database for Tumor Segmentation and Measurement Results (B.1). The MySQL database storing and managing tumor segmentation and measurement results in a dedicated server with superior data protection and twice-daily data backup, rather than in the DICOM image archive sever with inferior data protection and monthly data backup.

We keep the tumor segmentation database and the DCM4CHEE database independent for the purpose of easy upgrade and backup. Also, the user table of the DCM4CHEE database is used more often than the tumor segmentation database. We thus store the user registration information in two databases. We synchronize the 2 user registration databases by a program so that a change in one registration database will be automatically made to the other registration database.

Workflow

Figure 5 shows a sequence diagram of the workflow that users use to login to the system and views selected DICOM images.

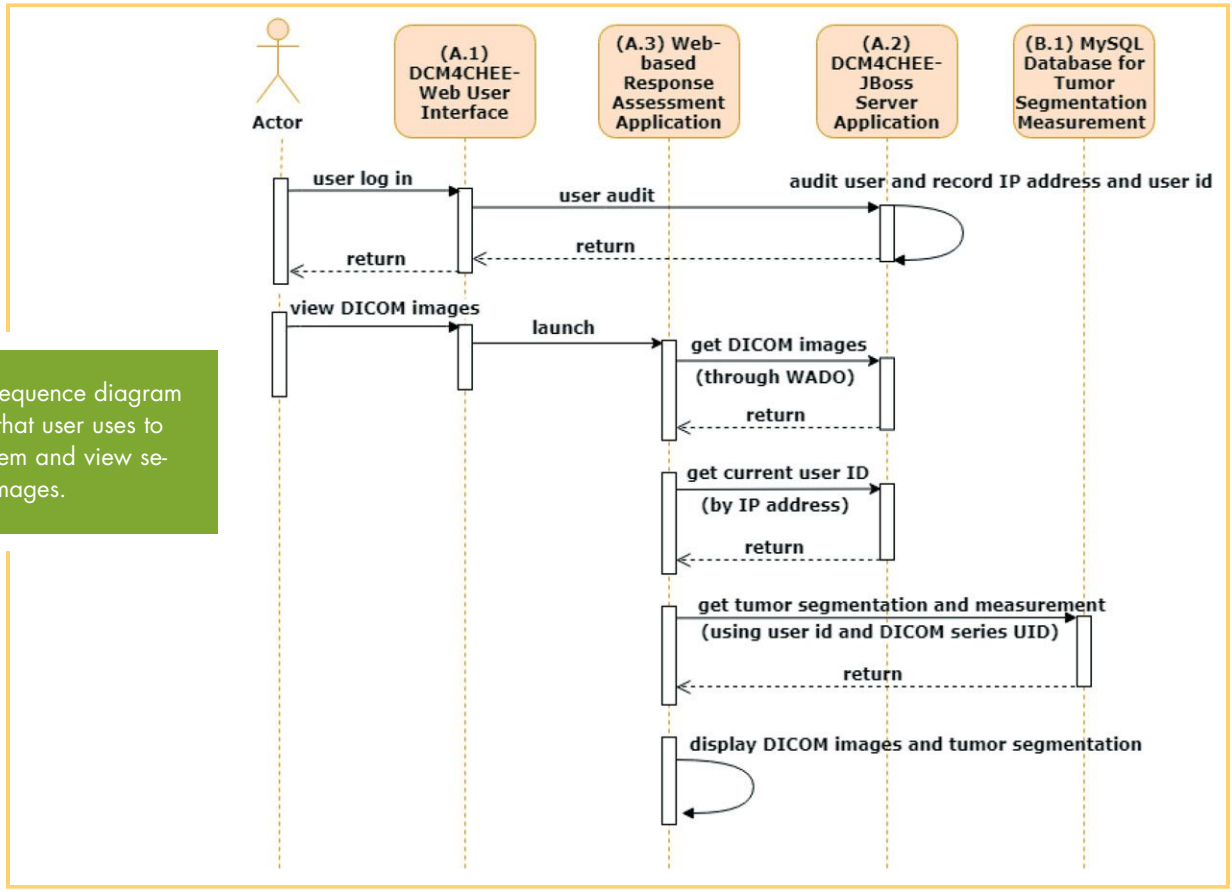


Figure 5. The sequence diagram of the workflow that user uses to log in to the system and view selected DICOM images.

When a user logs in to the system in a web browser, the DCM4CHEE-JBoss server application records the IP address of the user's PC and the user's credentials in the MySQL database for DCM4CHEE, as shown in the upper part of Figure 5. Once approved by the DCM4CHEE-JBoss server application, the user will see the DCM4CHEE–web user interface shown in Figure 4. Then the user may choose DICOM images of patients, studies, or series and click the link (highlighted in Figure 4) to web-based response-assessment application to view them.

The web-based response-assessment application can retrieve DICOM images archived by DCM4CHEE, because the DCM4CHEE-JBoss server application supports web access of DICOM images. A middle ware, called a WEASIS-PACS-connector, streamlines the process of retrieving DICOM images. When the user requests viewing the selected DICOM images, using the DICOM Query service of DCM4CHEE-JBoss, the middle ware collects the necessary information to launch the web-based response-assessment application.

After the response-assessment application is launched, it gets the DICOM image archived by DCM4CHEE, through the WADO Service of DCM4CHEE-JBoss. Next, the response-assessment application gets a user ID from the client PC's IP address from the DCM4CHEE-JBoss server application, as both user ID and client PC's IP address have been recorded by the DCM4CHEE-JBoss server application. Later, using user id and DICOM series UIDs of DICOM images, the response-assessment application gets tumor segmentation and measurement from the MySQL database for tumor segmentation and measurement results. Last, the response-assessment application displays DICOM images and their associated tumor segmentation.

Finally, the response-assessment application shows the selected DICOM images. The user can use response-assessment applications to segment tumors immediately. Only response-assessment applications and the database of tumor segmentation and measurement results are involved in the process of tumor segmentation, which was detailed in our previous paper (4).

DISCUSSION

We have developed a web-based imaging system to support the development and validation of quantitative imaging biomarkers for improved assessment of (solid) tumor responses to therapies, particularly novel targeted therapy and immunotherapy. The web-based response-assessment application of this system is based on the open-source DICOM image viewer WEASIS. To manage DICOM images, the response-assessment imaging system incorporates DCM4CHEE, an open-source DICOM image archive. The system consists of 2 interdependent servers with a Linux operating system: 1 hosts the DICOM image archive and web-based response-assessment application and the other hosts image biomarkers, for instance, tumor segmentation and unidimensional and volumetric measurement results. Users can log in to the web-based response-assessment imaging system using a web browser, and browse data on patients or on studies, and remotely access DICOM images and tumor segmentation on them.

The web-based response-assessment has many advantages over the previous portable response-assessment system: (1) archiving the DICOM images in a server rather than on a local hard disk of the client's PC, the system promotes the management of

DICOM images, for example, access control and storage of the DICOM images increase and (2) the system facilitates the distributing, updating, and upgrading of the response-assessment application by configuring the application to be web-based and to be hosted in a server.

As mentioned earlier, our objective is to develop an advanced imaging platform to accelerate the development and validation of novel quantitative imaging biomarkers for tumor response assessment by providing efficient tumor measurement tools. Our research system to assess tumor response is built based on an open-source, the WEASIS, platform. It is a PACS-like workstation that has basic image-viewing and manipulation functions. We customized it specifically for the assessment of advanced quantitative imaging biomarkers by (1) developing an industrial standard, novel relational database structure to store segmented tumor contours and measurements (4); (2) integrating our homegrown advanced tumor segmentation and editing tools so that tumor contours can be delineated more accurately and efficiently; (3) providing lesion tracking tools to reduce human error in tumor measurements at multiple scan time-points; and (4) making the system more user-friendly across multiple platforms and various screen sizes, and more accessible from different locations. Most importantly, our system is designed with an extendable architecture, so that other image-based quantitative tasks (eg, body fat quantification) can be easily added to the system.

We are aware that there exist many open-source and commercially available tools that provide similar functionality for lesion segmentation and/or lesion tracking. For example, there are 3D Slicer (11), ePAD (12), ITK-SNAP (13), OsiriX MD (14), and ClearCanvas (15). 3D Slicer is an open-source software platform and widely used by researchers worldwide for medical image processing (eg, lesion segmentation) and 3-dimensional visualization. However, 3D Slicer is not developed specifically for tumor response assessment, and thus, when using it for this purpose, it will not be as efficient as ours. For example, the 3D Slicer does not provide any lesion tracking tools that would be important when measuring lesions on longitudinal scan time points. ePAD is a web-based image viewer and annotator for quantitative image analysis. The system uses Annotation and Image Markup (AIM) file for tumor segmentation and measurement results and stores these files in an AIM Annotation Database. The AIM Annotation Database is an XML database that is known to be inefficient and unreliable for storing and maintaining large volumes of data. As to commercially available systems, such as OsiriX MD and ClearCanvas, the significant advantage of our system over them lies in its great capability to be extended. For example, we can add radiomic feature extraction methods easily to our system, whereas a commercial system has a hard time doing so.

Our response-assessment system has shown value in its ability to efficiently obtain/measure tumor size, particularly tumor volume, at serial scan time points in clinical trial settings to help monitor changes in total tumor burden—a potentially better imaging biomarker of response.

We will integrate our custom-developed radiomics features into our response-assessment system, so that it can be used to explore tumor imaging phenotypes for therapy response predictions and patient stratification for future clinical trials. We also plan to extend this system for exploring the data of DICOM images and tumor segmentation results, such as using artificial intelligence to automatically identify target lesions on baseline scans and new lesions on follow-up scans.

ACKNOWLEDGMENTS

This work was supported in part by Grant U01 CA225431 from the National Cancer Institute (NCI). The content is solely the responsibility of the authors and does not necessarily represent the funding sources.

Disclosure: No disclosures to report.

Conflict of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, Verweij J, Van Glabbeke M, van Oosterom AT, Christian MC, Gwyther SG. New guidelines to evaluate response to treatment in solid tumors. J Natl Cancer Inst. 2000;92:205–216.
2. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, Dancey J, Arbuck S, Gwyther S, Mooney M, Rubinstein L, Shankar L, Dodd L, Kaplan R, Lacombe D, Verweij J. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009;45:228–247.
3. Zhao B, Oxnard GR, Moskowitz CS, Kris MG, Pao W, Guo P, Rusch VM, Ladanyi M, Rizvi NA, Schwartz LH. A pilot study of volume measurement as a method of tumor response evaluation to aid biomarker development. Clin Cancer Res. 2010;16:4647–4653.
4. Yang H, Schwartz LH, Zhao B. A response assessment system for development and validation of imaging biomarkers in oncology. Tomography. 2016;2:406–410.
5. Weasis. https://nroduit.github.io/en/
6. DCM4CHEE. https://www.dcm4che.org/
7. Tan Y, Schwartz LH, Zhao B. Segmentation of lung lesions on CT scans using watershed, active contours, and Markov random field. Med Phys. 2013;40:043502.
8. Guo X, Zhao B, Schwartz LH. Methods and systems for segmentation of organs and tumors and objects. U.S. Patent Application 14/394,097, filed March 19, 2015.
9. Tan Y, Lu L, Bonde A, Wang D, Qi J, Schwartz LH, Zhao B. Lymph node segmentation by dynamic programming and active contours. Med Phys. 2018;45:2054–2062.
10. Guo X, Schwartz LH, Zhao B. Semi-automatic segmentation of multimodal brain tumor using active contours. In *Proceedings of Workshop on Brain Tumor Segmentation MICCAI,* 2013;27–30.
11. 3D Slicer. https://www.slicer.org/
12. Rubin DL, Willrett D, O'Connor MJ, Hage C, Kurtz C, Moreira DA. Automated tracking of quantitative assessments of tumor burden in clinical trials. Transl Oncol. 2014;7:23–35.
13. ITK-SNAP. http://www.itksnap.org/
14. Osirix MD. https://www.osirix-viewer.com/osirix/osirix-md/
15. ClearCanvas. https://www.clearcanvas.ca/

TOMOGRAPHY® RESEARCH ARTICLE

Reliability of Radiomic Features Across Multiple Abdominal CT Image Acquisition Settings: A Pilot Study Using ACR CT Phantom

Lin Lu, Yongguang Liang, Lawrence H. Schwartz, and Binsheng Zhao

Department of Radiology, Columbia University Medical Center, New York, NY

Corresponding Author:
Binsheng Zhao, DSc
Department of Radiology, Columbia University Medical Center,
710 West 168th Street, B26, New York, NY 10032;
E-mail: bz2166@cumc.columbia.edu

Key Words: quantitative imaging biomarkers, Radiomic Features, Reliability, Abdominal CT
Abbreviations: Computed tomography (CT), radiomic features (RFs), gray-level co-occurrence matrix (GLCM), regions of interest (ROIs)

ABSTRACT

We studied the reliability of radiomic features on abdominal computed tomography (CT) images reconstructed with multiple CT image acquisition settings using the ACR (American College of Radiology) CT Phantom. Twenty-four sets of CT images of the ACR CT phantom were attained from a GE Discovery 750HD scanner using 24 different image acquisition settings, combinations of 4 tube currents (25, 50, 100, 200 Effective mAs), 3 slice thicknesses (1.25, 2.5, 5 mm), and 2 convolution kernels (STANDARD and SOFT). Polyethylene (−95 HU) and acrylic (120 HU) of the phantom model were selected for calculating real feature value; a noise-free, computer-generated phantom image series that reproduced the 2 objects and the background was used for calculating reference feature value. Feature reliability was defined as the degree of predicting reference feature value from real feature value. Radiomic features *mean, std, skewness, kurtosis,* gray-level co-occurrence matrix (*GLCM*)*-energy, GLCM-contrast, GLCM-correlation, GLCM-homogeneity* were investigated. The value of $R^2 \geq 0.85$ was considered to be of high reliability. The reliability of *mean* and *std* were high across all image acquisition settings. At 200 Effective mAs, all features except *GLCM-homogeneity* showed high reliability, whereas at 25 Effective mAs, most features (except *mean* and *std*) showed low reliability. From high to low, reliability was ranked in the following order: *mean, std, skewness, kurtosis, GLCM-energy, correlation, contrast* and *homogeneity*. CT image acquisition settings affected the reliability of radiomic features. High reliable features were attained from images reconstructed at high tube current and thick slice thickness.

INTRODUCTION

Medical imaging plays an ever greater role in disease diagnosis and patient care. One of the most exciting new areas related to cancer diagnosis, treatment planning, and response assessment is the field of radiomics, which involves the extraction and analysis of a large number of quantitative imaging features from medical images for characterization of tumor and tissue phenotypes (1, 2).

Owing to the associations between tumor phenotypes and underlying biological processes, radiomic features (RFs) or RF-derived phenotypes can act as biomarkers that convey information about disease to help with the management of therapies. To date, radiomics has shown promise in improving cancer diagnosis and prognostic assessment in several tumor types including lung (3-5), brain (6), breast (7), liver (8-10), kidney (11), and esophagus (12) cancers. Moreover, RFs also exhibit correlations with genetic mutation status (5) and disease recurrence (13), as well as therapeutic response (14) and survival (15) in lung cancer.

While serving as an imaging biomarker for oncology, the influence of image acquisition settings on RFs should be well understood before the biomarker can be fully utilized (16). Until now, numerous studies have been conducted on the "reproducibility" of RFs (17-20), which refers to whether feature values could remain the same when reimaged using different equipment and different image acquisition settings. To the best of our knowledge, with the exception of studies on the accuracy of volume measurements (21, 22), there has been no report to date exploring the "reliability" of RFs. "Reliability" refers to whether true feature value could be maintained when imaged using different scanners and image acquisition settings. The true feature value in our study was defined as the feature value that was calculated on computed tomography (CT) image within which the CT number of each tissue composition was equal to its theoretical CT number at 120 kVp, for example, air equals to −1000 HU, and water equals to 0 HU. Thus, true feature value was also called as reference value in our study.

ISSN 2379-1381 http://dx.doi.org/10.18383/j.tom.2019.00005

Table 1. Image Acquisition Parameters

Scanner	GE Discovery 750HD (64 slices)
kVp	120
Display field of view (cm)	22
Pitch	1.375
Tube Currents (effective mAs)	25, 50, 100, 200
Rotation time (second)	0.7
Beam width (mm)	40 (64x0.625)
Slice thickness (mm)	1.25, 2.5, 5
Overlap (%)	0
Reconstruction algorithms	STANDARD, SOFT

The challenge of such a reliability study lies in the fact that reference values for RFs are generally quite difficult to obtain, especially for in vivo lesions, because of unknown tissue composition, as well as anatomic, physiologic, and even positional variations among different patients. In view of this point, we aimed to carry out a pilot study on RF reliability using the ACR CT phantom (American College of Radiology CT accreditation phantom) (23). The ACR CT phantom is a widely used CT QC phantom, and has a well-defined CT number for each object inside module 1.

In this study, we attained CT images of the phantom under 24 image acquisition settings using a GE Discovery 750HD scanner (GE Healthcare, Waukesha, WI). The reliability of 8 widely used RFs—*mean, std, skewness, kurtosis, GLCM* [*gray-level co-occurrence matrix* (24)]-*energy, GLCM-contrast, GLCM-correlation,* and *GLCM-homogeneity*—was investigated on the 24 sets of CT images.

METHODS

Scanning the ACR CT Phantom

A Gammex CT ACR 464 phantom was scanned on a GE Discovery 750HD scanner using a routine adult abdomen protocol at 4 different tube currents (25, 50, 100, 200 Effective mAs). The CT images were then reconstructed with 3 different slice thicknesses (1.25, 2.5, 5 mm) and 2 convolution kernels (STANDARD, SOFT), resulting in a total of $4 \times 3 \times 2 = 24$ sets of CT images. The CT scanning parameters used in this study are listed in Table 1.

Preparation of Image Region and ROIs for Extracting Real Feature Value

The ACR CT phantom is composed of 4 modules and primarily constructed from water-equivalent materials (23). Each module contains several components made of different materials. In our study, 2 circular objects from module 1, made of polyethylene and acrylic each, were selected to create image patterns for feature extraction. Polyethylene and acrylic are materials with CT numbers of −95 HU and 120 HU at a 120-kVp setting falling within the ranges of the abdominal CT window.

For each object, a 2-dimensional region of 45×45 mm containing the object was cropped from the CT image located at the center of module 1 along the axial direction. Within the cropped region, 100 regions of interest (ROIs) were randomly generated. The criteria to generate ROIs included the following:

(1) The center of the ROI should be located inside the object.
(2) ROI shall cover part of the object and part of the background outside the object, for the purpose of studying radiomic features on nonhomogenous patterns rather than only on homogenous patterns, such as that derived from cartridge phantoms filled with paper/rubber in the literature (19).
(3) The size of the ROI must range from 12×12 mm to 18×18 mm; the sizes of the cropped region and ROIs were empirically determined on the basis of the physical size of the object (a cylinder with diameter = 25 mm and depth = 4 cm as provided in the manual of ACR CT phantom). The process of preparing the cropped region and ROIs is illustrated in Figure 1.

Preparation of Computer-Generated Images for Extracting Reference Feature Value

A noise-free digital image series to simulate module 1 of the ACR CT phantom was generated for the extraction of reference feature values. The 2 selected objects (polyethylene and acrylic) were reproduced via an image-processing algorithm on the basis of designated parameters (eg, location, size, shape, and density in CT number) provided in the phantom manual (23). The rest of the computer-generated images were defined as water-equivalent background with a CT number = 0. The image region and ROIs for feature extraction from the computer-generated images were copied from those used in the scanned phantom images to guarantee that they were identical so that variations introduced by position misalignment and density difference could be minimized and the bias of real value to reference value would be purely because of the different image acquisition settings.

Extraction of Feature

In our study, 8 2D RFs were investigated, including 4 histogram-based features—*mean*, *std (standard deviation)*, *skewness*, and *kurtosis*—and four texture-based GLCM features (24), *GLCM-energy*, *GLCM-contrast*, *GLCM-correlation*, and *GLCM-homogeneity*. *Mean*, *std*, *skewness*, and *kurtosis* are first-order statistic features to characterize an histogram of image intensity. GLCM features are textural features characterizing the gray-tone spatial dependencies of an image, that is, quantifying the relationship between pixels within an ROI. Details of definitions of the 8 RFs are provided in the online Supplemental Material.

In the implementation, the 8 RFs were calculated on each ROI by using an in-house feature extraction algorithm programmed on the MATLAB 2016b platform (MathWorks, Natick, MA). Before feature calculation, images were interpolated into isotropic pixel spacing of 0.5×0.5 mm^2.

Reliability of Feature

In our study, feature reliability was defined as the degree of predicting reference feature value from real feature value. Reference feature value was the feature value extracted from noise-free computer-generated phantom images, while real feature value is the feature value extracted from CT images attained from the physical ACR phantom. High predictability means that

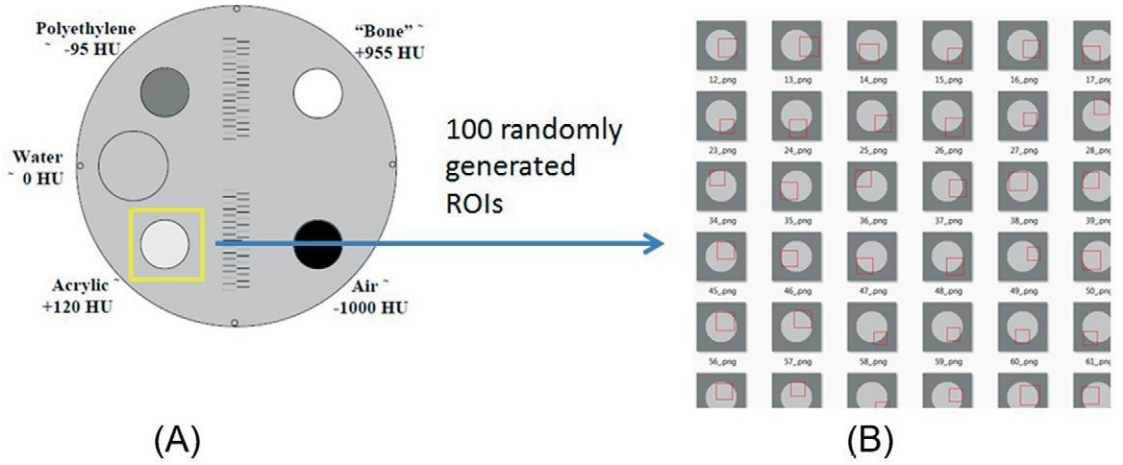


Figure 1. Example of an ACR CT phantom image being prepared for feature extraction. One region containing the object (yellow frame of 45 × 45 mm) is selected from the ACR CT phantom image (A). Samples of 100 randomly generated nonhomogenous ROIs of the object (B). Each ROI was a 2-dimensional square with random location and random size ranging from 12 × 12 mm to 18 × 18 mm.

a change in reference feature value can be correctly reflected by a proportional change in the real feature value. If an RF exhibited high predictability under a certain image acquisition setting, then the RF calculation was believed to be reliable.

Consequently, R^2, a statistical metric widely used to assess the proportion of variance in the dependent variable that is predictable from the independent variable, was adopted to quantify feature reliability. An R^2 value of 1 indicated that the reference feature value could be predicted by a real feature value to a degree of 100%, whereas an R^2 value of 0 indicated that there was no relation between reference feature value and real feature value. The R^2 equation can be defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (x_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (1)$$

Where x_i represents real feature value extracted from the ith ROI on the ACR CT phantom images, y_i represents reference feature value extracted from the corresponding ith ROI on the computer-generated phantom images, $\bar{y}$ represents the mean of reference feature values extracted from 200 ROIs, and n equals 200 (100 ROIs from each of the 2 ACR objects).

Figure 2 shows an example of how to use R^2 to assess feature reliability under certain image acquisition setting. The graphs (A) and (B) in Figure 2 present the skewness values, one of the histogram-based RFs, calculated from ROIs under the image acquisition settings of "convolution kernel = STANDARD, slice thickness = 1.25 mm, and Effective mAs = 200" and "convolution kernel = STANDARD, slice thickness = 1.25 mm, Effective mAs = 25," respectively. The feature data used to estimate R^2 value consisted of 200 pairs of skewness values, corresponding to the reference and real skewness values calculated on the 200 ROIs on the computer-generated and physical phantom images, respectively. As shown in Figure 2, high reliability ($R^2 = 0.9575$) indicated that reference skewness values approximated the real skewness values measured at high tube current, while low reliability ($R^2 = 0.4021$) indicated reference skewness values diverged from real skewness values measured at low tube current.

RESULTS

Figure 3 shows the reliability values for the 8 RFs under 24 image acquisition settings, combinations of 4 tube currents (25, 50, 100, 200 Effective mAs), 3 slice thicknesses (1.25, 2.5, 5 mm), and 2 convolution kernels (STANDARD and SOFT). Overall, we were able to observe that feature reliability decreased with a decrease in tube current, features were more reliable on 5-mm CT images than on 1.25- and 2.5-mm CT images, there was little difference in feature reliability between CT images of STANDARD and SOFT convolution kernels, and histogram-based RFs are more reliable than textural RFs.

We averaged the reliability values across individual image acquisition parameters to further investigate their influence (Table 2). For example, when investigating the influence of "200 Effective mAs," we averaged the feature reliability values of "STANDARD_ST125_EffmAs200," "STANDARD_ST250_EffmAs200," "STANDARD_ST500_EffmAs200," "SOFT_ST125_EffmAs200," "SOFT_ST250_EffmAs200," and "SOFT_ST500_EffmAs200" together as presented in Figure 3.

To facilitate the analysis, we empirically set $R^2 > 0.85$ as high reliability. As shown in Table 2, in the case of tube current, 100 Effective mAs could be regarded as a threshold to guide the application of RFs, that is, using tube current ≥100 Effective mAs resulted in more reliable RFs, especially the histogram-based RFs, while using tube current <100 Effective mAs produced only a few reliable RFs. For slice thickness, 5-mm CT images yielded more reliable RFs. For convolution kernel, the STANDARD and SOFT showed similar influence on feature re-

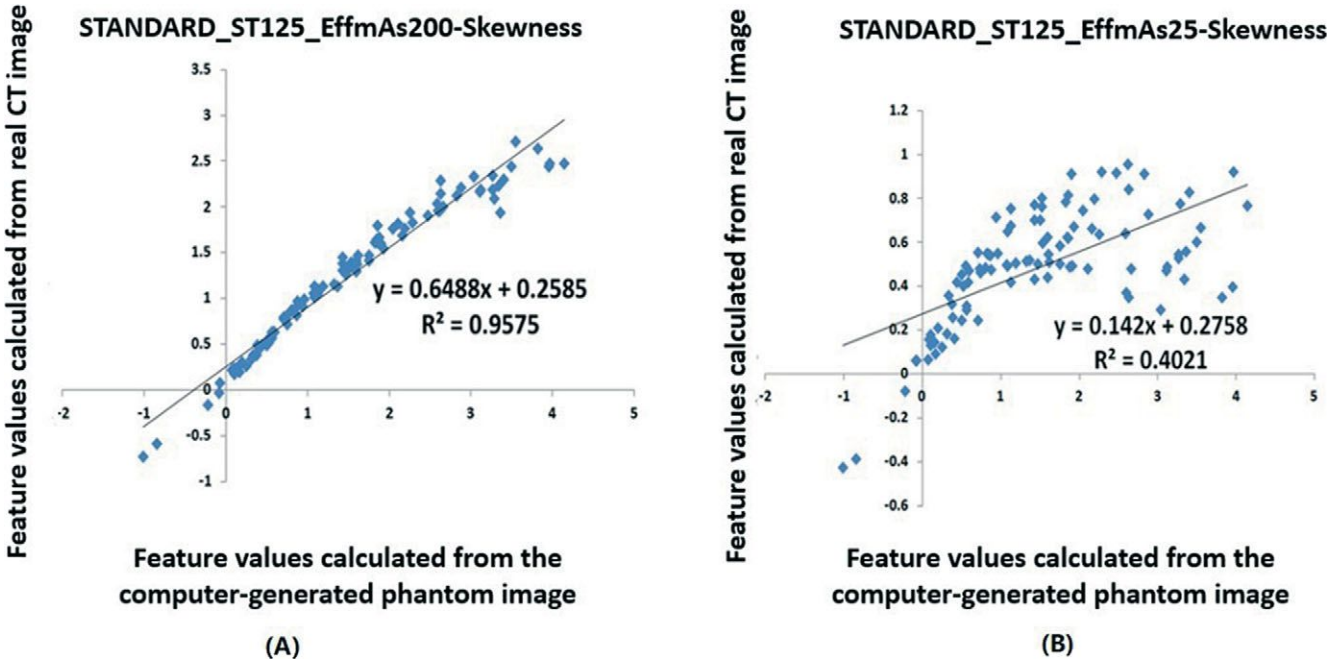


Figure 2. Reliability of the skewness feature at high tube current (A) and low tube current (B), respectively. Each point on the plot corresponds to the values calculated from one randomly generated ROI on the computer-generated (X-axis) and the physical phantom images (Y-axis), respectively. There are a total of 200 points on each plot corresponding to the 200 randomly generated ROIs from the two selected objects in the phantom.

liability. The feature *mean* and *std* showed extremely high reliability across all image acquisition settings.

We observed an obvious unusual trend that the average reliability of GLCM-energy at slice thickness of 1.25 mm was higher than that at slice thickness of 2.5 mm (see Table 2). As we turned to the details of reliability presented in Figure 3, we found that the unusual trend was caused by a great drop of reliability at tube current 25 Effective mAs, a very low dose condition for the slice thickness of 2.5 mm. Actually, based on our results, low tube current easily led to unusual trends for some RFs, for example, GLCM-homogeneity at 50 and 25 Effective mAs and GLCM-homogeneity at 1.25- and 2.5-mm slice thicknesses in Table 2.

Features	STANDARD_ ST125_Effm As200	STANDARD_ ST125_Effm As100	STANDARD_ ST125_Effm As50	STANDARD_ ST125_Effm As25	STANDARD_ ST250_Effm As200	STANDARD_ ST250_Effm As100	STANDARD_ ST250_Effm As50	STANDARD_ ST250_Effm As25	STANDARD_ ST500_Effm As200	STANDARD_ ST500_Effm As100	STANDARD_ ST500_Effm As50	STANDARD_ ST500_Effm As25
mean	0.998	0.999	0.997	0.993	0.996	0.998	0.998	0.996	0.996	0.997	0.999	0.996
std	0.989	0.985	0.965	0.953	0.990	0.987	0.975	0.966	0.990	0.989	0.986	0.985
skewness	0.958	0.812	0.511	0.402	0.973	0.929	0.749	0.454	0.984	0.972	0.907	0.780
kurtosis	0.931	0.824	0.619	0.500	0.969	0.933	0.800	0.565	0.972	0.968	0.927	0.813
glcm-energy	0.941	0.947	0.885	0.768	0.963	0.935	0.930	0.498	0.924	0.947	0.938	0.760
glcm-contrast	0.895	0.656	0.124	0.328	0.917	0.771	0.238	0.119	0.922	0.867	0.619	0.538
glcm-correlation	0.847	0.794	0.802	0.754	0.868	0.811	0.793	0.813	0.894	0.828	0.796	0.814
glcm-homogeneity	0.716	0.546	0.228	0.333	0.746	0.633	0.037	0.168	0.746	0.736	0.296	0.495
Features	**SOFT_ST125 _EffmAs200**	**SOFT_ST125 _EffmAs100**	**SOFT_ST125 _EffmAs50**	**SOFT_ST125 _EffmAs25**	**SOFT_ST250 _EffmAs200**	**SOFT_ST250 _EffmAs100**	**SOFT_ST250 _EffmAs50**	**SOFT_ST250 _EffmAs25**	**SOFT_ST500 _EffmAs200**	**SOFT_ST500 _EffmAs100**	**SOFT_ST500 _EffmAs50**	**SOFT_ST500 _EffmAs25**
mean	0.997	0.999	0.998	0.993	0.996	0.998	0.998	0.995	0.996	0.997	0.999	0.997
std	0.989	0.986	0.972	0.959	0.989	0.988	0.978	0.973	0.989	0.989	0.987	0.985
skewness	0.955	0.883	0.649	0.453	0.980	0.957	0.841	0.602	0.985	0.980	0.944	0.862
kurtosis	0.951	0.882	0.723	0.591	0.972	0.955	0.874	0.657	0.970	0.972	0.953	0.870
glcm-energy	0.926	0.930	0.835	0.786	0.959	0.937	0.830	0.504	0.925	0.952	0.901	0.733
glcm-contrast	0.905	0.748	0.208	0.365	0.923	0.830	0.402	0.205	0.922	0.880	0.745	0.669
glcm-correlation	0.871	0.799	0.813	0.750	0.888	0.825	0.801	0.808	0.910	0.846	0.806	0.814
glcm-homogeneity	0.746	0.507	0.082	0.275	0.817	0.736	0.048	0.165	0.775	0.788	0.309	0.592

Figure 3. Reliability of 8 radiomic features under 24 image acquisition settings. Top panel with pink title: reconstructed using Standard kernel; Bottom panel with yellow title: reconstructed using Soft Kernel. For example, the number of 0.998 in the top-left cell is the R^2 value of the feature *mean* calculated between the computer-generated image and CT scan image obtained at 200 Effective mAs and reconstructed using STANDARD kernel, 1.25 mm slice thickness.

Table 2. Average of Reliability Values Under Individual Image Acquisition Parameters

Features	Tube Current				Slice Thickness			Convolution Kernel		
	200 Effective mAs	100 Effective mAs	50 Effective mAs	25 Effective mAs	1.25 mm	2.5 mm	5.0 mm	STANDARD	SOFT	All
Mean	0.997	0.998	0.998	0.995	0.997	0.997	0.997	0.997	0.997	0.997
Std	0.989	0.987	0.977	0.970	0.975	0.981	0.987	0.980	0.982	0.981
Skewness	0.973	0.922	0.767	0.592	0.703	0.811	0.927	0.786	0.841	0.813
Kurtosis	0.961	0.922	0.816	0.666	0.753	0.841	0.931	0.818	0.864	0.841
GLCM-energy	0.940	0.942	0.887	0.675	0.877	0.819	0.885	0.870	0.852	0.861
GLCM-contrast	0.914	0.792	0.389	0.371	0.529	0.551	0.770	0.583	0.650	0.616
GLCM-correlation	0.880	0.817	0.802	0.792	0.804	0.826	0.839	0.818	0.828	0.823
GLCM-homogeneity	0.758	0.658	0.167	0.338	0.429	0.419	0.592	0.473	0.487	0.480

DISCUSSION

In this study, we introduced the concept of RF reliability and evaluated the RF reliability of 8 commonly used RFs under 24 different image acquisition settings. The 24 image acquisition settings involved 3 image acquisition parameters, for example, tube current, slice thickness, and convolution kernel, and covered a wide range of imaging protocols for abdominal CT imaging. Moreover, our study was based on heterogeneous ROIs, that is, ROIs containing both object and background, which is an advantage over previous studies using homogenous ROI phantoms, for example, paper/rubber-filled cartridges (19).

Overall, for the ACR CT phantom, tube current affected reliability the most, slice thickness the second, and convolution kernel the least. The small effect of convolution kernels was due to the similarity of the 2 "smooth" kernels used in this abdominal study. The histogram-based RFs showed much higher reliability than textural RFs.

For tube current, 200 Effective mAs represented high-dose CT imaging, while 25 Effective mAs represented low-dose CT imaging. It is quite intuitive that CT images derived from high-dose scanning would yield more reliable RFs as it produced higher quality images than low noise scanning. Therefore, to obtain high RF reliability, high-dose CT imaging is recommended, especially for those radiomic studies using textural RFs. When keeping all other imaging acquisition parameters unchanged, increasing the slice thickness from 1.25 mm to 5 mm can reduce image noise by 50%. It is reasonable to believe that thick-section CT imaging yielded more reliable RFs. However, thick-section CT imaging introduces larger partial volume effect than thin-section CT imaging. In clinical practice, partial volume effect is one of the main negative effects that lowered image resolution and thus blurred fine structures within/around lesions, for example, small vessels, boundary of tumor margin, etc. It will also affect some RFs extracted from thick-section CT images. Therefore, the selection of RFs and slice thickness should depend on the aim of the radiomic study. Because the 2 convolution kernels, STANDARD and SOFT, both belonged to smooth soft-tissue kernels which yielded low-noise image, their influence on RF reliability was similar. Also, our results showed that smooth soft-tissue kernels used by abdominal CT scans had little impact on RF reliability.

In this study, 2 categories of RFs, histogram-based and the textural, were investigated. Histogram-based RFs showed much higher reliability than textural RFs, especially the *mean* and *std*. It is actually one of the basic requirements for a CT scanner that *mean* should be reliable across different image acquisition settings. Our results showed this. For the *std*, its high reliability was somewhat due to the use of the polyethylene and acrylic objects to create image patterns, which possessed dozens of Hounsfield unit (HU) intensity different from the water-equivalent background. Nevertheless, according to this finding, it is quite reliable to apply *std* in charactering tumor lesions with dozens of HU difference from the background, such as liver metastasis of colorectal cancer [mean, 68 HU; range, 40–115 HU as reported in the CRYSTAL clinical trials (25, 26)] and gastrointestinal stromal tumors [mean, 72 HU; range, 46–156 HU as reported in the Choi criteria study (27)].

In contrast to histogram-based RFs, more attention should be paid to the use of textural RFs. Textural RFs are easily affected by tube current, which is an imaging parameter directly proportional to patient radiation dose. High tube current guarantees high reliability of textural RFs, but leads to high patient dose. Therefore, the use of textural RFs should depend on the aim of a study. For example, it is inadvisable to use textural RFs in a low-dose CT screening study (28), whereas it might be safe to use textural RFs in a CT-based radiation therapy study (29).

There were several limitations of our pilot study. First, the created image patterns were simple, involving only 2 materials for each pattern, polyethylene, and a water-equivalent background, or acrylic and a water-equivalent background. Second, only a small set of RFs from 2 feature categories were investigated. Third, only 1 CT scanner was used. To address these limitations, we propose future studies, including designing more sophisticate phantoms that mimic in vivo lesions with the help of 3D-printing technique (30), using a high-throughput analysis method to evaluate a large scale of RFs (20), and involving multiple scanners from multiple institutions to attain CT images under more image acquisition settings (19).

CONCLUSION

In this study, we explored the reliability of RFs on multiple CT image acquisition settings. To the best of our knowledge, this is the first study investigating RF reliability by comparing real feature values calculated from scanned phantom images and reference feature values computed from computer-generated phantom images. We found that CT image acquisition settings influenced RF reliability to varying degrees. Therefore, attention should be paid when using RFs for CT-based radiomic studies, especially textural RFs.

Supplemental Materials

Supplemental Material: http://dx.doi.org/10.18383/j.tom.2016.00005.sup.01

ACKNOWLEDGMENTS

This work was supported in part by Grant U01 CA140207 and U01 CA225431 from the National Cancer Institute (NCI). The content is solely the responsibility of the authors and does not necessarily represent the funding sources.

Disclosures: No disclosures to report.

Conflict of Interest: The authors have no conflict of interest to declare.

REFERENCES

1. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, Bussink J, Monshouwer R, Haibe-Kains B, Rietveld D, Hoebers F, Rietbergen MM, Leemans CR, Dekker A, Quackenbush J, Gillies RJ, Lambin P. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:4006.
2. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology. 2015;278:563–577.
3. Gevaert O, Xu J, Hoang CD, Leung AN, Xu Y, Quon A, Rubin DL, Napel S, Plevritis SK. Non-small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data–methods and preliminary results. Radiology. 2012;264:387–396.
4. Parmar C, Leijenaar RT, Grossmann P, Rios Velazquez E, Bussink J, Rietveld D, Rietbergen MM, Haibe-Kains B, Lambin P, Aerts HJ. Radiomic feature clusters and prognostic signatures specific for Lung and Head & Neck cancer. Sci Rep. 2015;5:11044.
5. Liu Y, Kim J, Balagurunathan Y, Li Q, Garcia AL, Stringfield O, Ye Z, Gillies RJ. Radiomic features are associated with EGFR mutation status in lung adenocarcinomas. Clin Lung Cancer. 2016;17:441–448.e6.
6. Zhou M, Scott J, Chaudhury B, Hall L, Goldgof D, Yeom KW, Iv M, Ou Y, Kalpathy-Cramer J, Napel S, Gillies R, Gevaert O, Gatenby R. Radiomics in brain tumor: image assessment, quantitative feature descriptors, and machine-learning approaches. AJNR Am J Neuroradiol. 2018;39:208–216.
7. Valdora F, Houssami N, Rossi F, Calabrese M, Tagliafico AS. Rapid review: radiomics and breast cancer. Breast Cancer Res Treat. 2018;169:217–229.
8. West DL, Kotrotsou A, Niekamp AS, Idris T, Giniebra Camejo D, Mazal NJ, Cardenas NJ, Goldberg JL, Colen RR. CT-based radiomic analysis of hepatocellular carcinoma patients to predict key genomic information [abstract]. JCO. 2017;35.15_suppl.e15623.
9. Jeong WK, Jamshidi N, Felker ER, Raman SS, Lu DS. Radiomics and radiogenomics of primary liver cancers. Clin Mol Hepatol. 2018. [Epub ahead of print].
10. Starmans MP, Miclea RL, van der Voort SR, Niessen WJ, Thomeer MG, Klein S, editors. Classification of malignant and benign liver tumors using a radiomics approach. Proc. SPIE. 2018; 105741D.
11. Yu H, Scalera J, Khalid M, Touret A-S, Bloch N, Li B, Qureshi MM, Soto JA, Anderson SW. Texture analysis as a radiomic marker for differentiating renal tumors. Abdom Radiol (NY). 2017;42:2470–2478.
12. van Rossum PS, Xu C, Fried DV, Goense L, Lin SH. The emerging field of radiomics in esophageal cancer: current evidence and future potential. Transl Cancer Res. 2016;5:410–23.
13. Cook GJ, Yip C, Siddique M, Goh V, Chicklore S, Roy A, Marsden P, Ahmad S, Landau D. Are pretreatment 18F-FDG PET tumor textural features in non-small cell lung cancer associated with response and survival after chemoradiotherapy? qj J Nucl Med. 2013;54:19–26.
14. Aerts HJ, Grossmann P, Tan Y, Oxnard GR, Rizvi N, Schwartz LH, Zhao B. Defining a radiomic response phenotype: a pilot study using targeted therapy in NSCLC. Sci Rep. 2016;6:33860.
15. Ganeshan B, Panayiotou E, Burnand K, Dizdarevic S, Miles K. Tumour heterogeneity in non-small cell lung carcinoma assessed by CT texture analysis: a potential marker of survival. Eur Radiol. 2012;22:796–802.
16. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, Forster K, Aerts HJ, Dekker A, Fenstermacher D, Goldgof DB, Hall LO, Lambin P, Balagurunathan Y, Gatenby RA, Gillies RJ. Radiomics: the process and the challenges. Magn Reson Imaging. 2012;30:1234–1248.
17. Zhao B, Tan Y, Tsai WY, Qi J, Xie C, Lu L, Schwartz LH. Reproducibility of radiomics for deciphering tumor phenotype with imaging. Sci Rep. 2016;6:23428.
18. Lu L, Ehmke RC, Schwartz LH, Zhao B. Assessing agreement between radiomic features computed for multiple CT imaging settings. PloS One. 2016;11:e0166550.
19. Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, Rodriguez-Rivera E, Dodge C, Jones AK, Court L. Measuring CT scanner variability of radiomics features. Invest Radiol. 2015;50:757–765.
20. Berenguer R, Pastor-Juan MdR, Canales-Vázquez J, Castro-García M, Villas MV, Legorburo FM, Sabater S. Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters. Radiology. 2018;288(2):407–415.
21. Li Q, Gavrielides MA, Sahiner B, Myers KJ, Zeng R, Petrick N. Statistical analysis of lung nodule volume measurements with CT in a large-scale phantom study. Med Phys. 2015;42:3932–3947.
22. Gavrielides MA, Berman BP, Supanich M, Schultz K, Li Q, Petrick N, Zeng R, Siegelman J. Quantitative assessment of nonsolid pulmonary nodule volume with computed tomography in a phantom study. Quant Imaging Med Surg. 2017;7:623–635.
23. McCollough CH, Bruesewitz MR, McNitt-Gray MF, Bush K, Ruckdeschel T, Payne JT, Brink JA, Zeman RK; American College of Radiology. The phantom portion of the American College of Radiology (ACR) computed tomography (CT) accreditation program: practical tips, artifact examples, and pitfalls to avoid. Med Phys. 2004;31:2423–2442.
24. Haralick RM, Shanmugam K. Textural features for image classification. IEEE Trans Syst Man Cybern. 1973:610–621.
25. Dercle L, Lu L, Lichtenstein P, Yang H, Wang D, Zhu J, et al. Impact of variability in portal venous phase acquisition timing in tumor density measurement and treatment response assessment: metastatic colorectal cancer as a paradigm. JCO Clin Cancer Inform. 2017;1:1–8.
26. Van Cutsem E, Köhne CH, Hitre E, Zaluski J, Chang Chien CR, Makhson A, D'Haens G, Pintér T, Lim R, Bodoky G, Roh JK, Folprecht G, Ruff P, Stroh C, Tejpar S, Schlichting M, Nippgen J, Rougier P. Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer. 2009;360:1408–1417.
27. Choi H, Charnsangavej C, Faria SC, Macapinlac HA, Burgess MA, Patel SR, Chen LL, Podoloff DA, Benjamin RS. Correlation of computed tomography and positron emission tomography in patients with metastatic gastrointestinal stromal tumor treated at a single institution with imatinib mesylate: proposal of new computed tomography response criteria. J Clin Oncol. 2007;25:1753–1759.
28. National Lung Screening Trial Research Team, Church TR, Black WC, Aberle DR, Berg CD, Clingan KL, Duan F, Fagerstrom RM, Gareen IF, Gierada DS, Jones GC, Mahon I, Marcus PM, Sicks JD, Jain A, Baum S. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011;365:395–409.
29. Huynh E, Coroller TP, Narayan V, Agrawal V, Hou Y, Romano J, Franco I, Mak RH, Aerts HJ. CT-based radiomic analysis of stereotactic body radiation therapy patients with lung cancer. Radiother Oncol. 2016;120:258–266.
30. Filippou V, Tsoumpas C. Recent advances on the development of phantoms using 3D printing for imaging with CT, MRI, PET, SPECT, and ultrasound. Med Phys. 2018;45:e740–e60.

Publish today in Tomography

Scope

Tomography publishes basic (technical and pre-clinical) and clinical scientific articles which involve the advancement of imaging technologies. Tomography encompasses studies that use single or multiple imaging modalities including for example CT, US, PET, SPECT, MR and hyperpolarization technologies, as well as optical modalities (i.e. bioluminescence, photoacoustic, endomicroscopy, fiber optic imaging and optical computed tomography) in basic sciences, engineering, preclinical and clinical medicine. Studies involving hardware and software advances along with chemical and molecular probe developments are also of high interest for publication in Tomography.

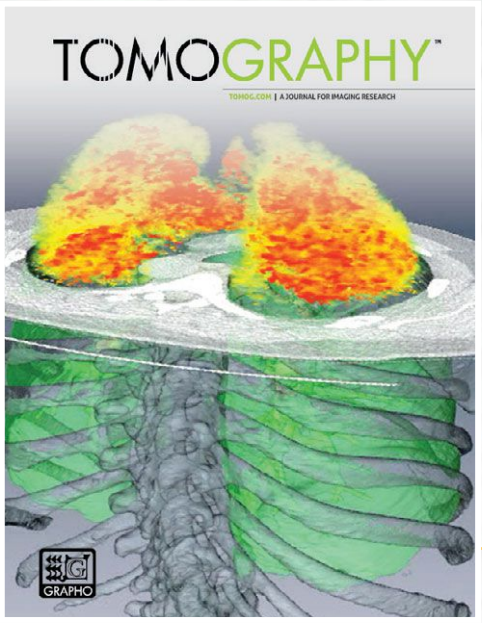
Submission

Submission of manuscripts can be accomplished through the web site www.Tomography.org using our online submission system.

Open Access Compliance

Funding Agencies and Article Repositories

Certain funders, including the NIH, members of the Research Councils UK (RCUK), and Wellcome Trust require deposit of the Accepted Version in a repository after an embargo period. Tomography allows all of its authors to deposit their articles on such repositories immediately upon publication.



Instructions for Authors

Research Articles

Original studies considered for publication will provide the field of imaging with significant advances in basic, preclinical, clinical, hardware and software using single or multi-modal imaging technologies. In addition, studies that use imaging technologies to advance basic and clinical medicine through interrogation of biological and pathological processes are also of interest.

Advances in Brief

Advances in Brief are timely high impact contributions on a topic important to imaging researchers. These Advances will undergo an accelerated review process. Papers submitted as Advances in Brief may combine the Results and Discussion sections.

Reviews

Tomography will publish Reviews which are timely and important to imaging researchers. Reviews should be written as concisely as possible. All Review articles will be subject to peer review. The parameters for this category are provide below but serve only as guidelines as longer reviews will be considered if the topic warrants additional space in the journal.

Perspectives

Tomography will publish Perspectives which are used to present new insights on an active or emerging area of imaging research or application of imaging technologies in the preclinical or clinical setting along with the author's personal viewpoint on the area including unresolved areas for future research along with key issues facing development and/or deployment.

Image Reports

Image Reports are discrete Reports (approximately 1,500 words) that provide significant advances and offer unique insights into the application of imaging technologies for the further advancement of preclinical or clinical diagnosis or treatment response. No specific limits are place on overall numbers of figures or tables but up to 4 would be reasonable.

Consensus Papers

Papers which provide a broad consensus on a timely topic related to imaging which is sponsored from investigators participating at a national meeting or meeting sponsored by a government funding agency. Authors are encouraged to contact the editorial office prior to submission.

TOMOGRAPHY®

AN INTERNATIONAL JOURNAL FOR IMAGING RESEARCH

gehealthcare.com

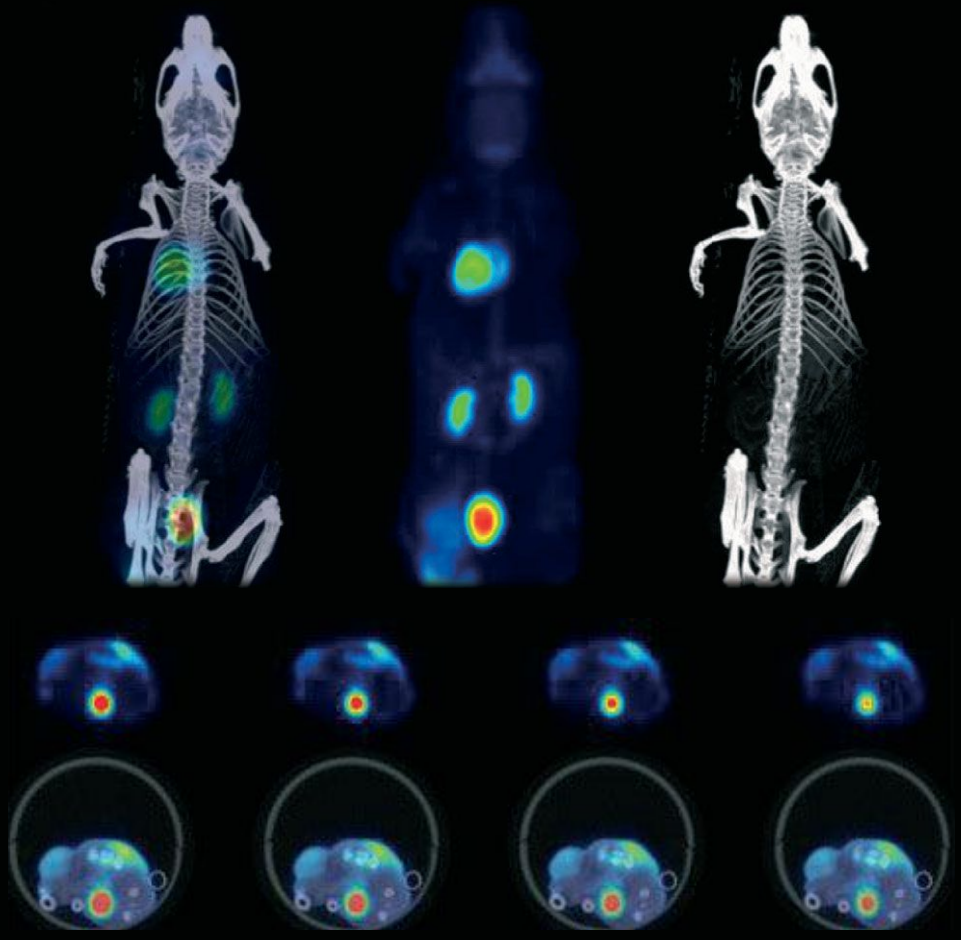
SIGNA™ Returns.

For more than 30 years, SIGNA has stood for quality, trust, and innovation. This year, SIGNA Returns with the introduction of a family of new products that bring you the trusted performance you've relied on for over three decades, plus powerful forward-thinking technologies to help enhance your clinical confidence.

Visit gehealthcare.com for more information.

BRUKER

Welcome to the World of Preclinical Imaging



Finest Detail in Every Organ

Accurate Quantification

Low Dose Fat Scan Times

Total Body FOV 80x200mm

- Introducing the PET/CT Si78 high performance PET and CT
- Total body PET for mice and rats
- Powered by ParaVision 360
- Simplified workflow
- Supports a wide range of applications

Discover more at: www.bruker.com/pci

Innovation with Integrity