CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Statistical assessment of treatment response in a cancer patient based on pre-therapy and post-therapy FDG-PET scans
Authors	Wolsztynski, Eric;O'Sullivan, Finbarr;O'Sullivan, Janet;Eary, Janet F.
Publication date	2016-12-18
Original Citation	Wolsztynski, E., O'Sullivan, F., O'Sullivan, J. and Eary, J. F. (2016) 'Statistical assessment of treatment response in a cancer patient based on pre-therapy and post-therapy FDG-PET scans', Statistics in Medicine, 36(7), pp. 1172-1200. doi:10.1002/sim.7198
Type of publication	Article (peer-reviewed)
Link to publisher's version	10.1002/sim.7198
Rights	© 2016 John Wiley & Sons, Ltd. This is the peer reviewed version of the following article: Wolsztynski, E., O'Sullivan, F., O'Sullivan, J. and Eary, J. F. (2016) 'Statistical assessment of treatment response in a cancer patient based on pre-therapy and post-therapy FDG-PET scans', Statistics in Medicine, 36(7), pp. 1172-1200, which has been published in final form at http://dx.doi.org/10.1002/sim.7198. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.
Download date	2026-11-06 10:17:26
Item downloaded from	https://hdl.handle.net/10468/3486

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Research Article

Statistics
in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Statistical assessment of treatment response in a cancer patient based on pre- and post-therapy FDG-PET scans

E. Wolsztynski*,[1], F. O'Sullivan[1], J. O'Sullivan[1], J. F. Eary[2]

This work arises from consideration of sarcoma patients in which fluorodeoxyglucose Positron Emission Tomography (FDG-PET) imaging pre- and post-chemotherapy is used to assess treatment response. Our focus is on methods for evaluation of the statistical uncertainty in the measured response for an individual patient. The Gamma distribution is often used to describe data with constant coefficient of variation, but it can be adapted to describe the pseudo-Poisson character of PET measurements. We propose co-registering the pre- and post- therapy images and modelling the approximately paired voxel-level data using the Gamma statistics. Expressions for the estimation of the treatment effect and its variability are provided. Simulation studies explore the performance in the context of testing for a treatment effect. The impact of mis-registration errors and how test power is affected by estimation of variability using simplified sampling assumptions, as might be produced by direct bootstrapping, is also clarified. The results illustrate a marked benefit in using a properly constructed paired approach. Remarkably the power of the paired analysis is maintained even if the pre- and post- image data are poorly registered. A theoretical explanation for this is indicated. The methodology is further illustrated in the context of a series of FDG-PET sarcoma patient studies. These data demonstrate the additional prognostic value of the proposed treatment effect test statistic. **Copyright © 2016 John Wiley & Sons, Ltd.**

Keywords: therapeutic effectiveness, percentage change in mean, paired analysis, patient-adaptive treatment, Gamma distribution.

1. Introduction

1.1. Methodological background

Positron emission tomography (PET) with ^{18}F-fluorodeoxyglucose (FDG) is a widely used clinical imaging diagnostic in the management of cancer [1, 2, 3, 4, 5, 6]. It is useful in particular for early assessment of therapeutic response, with a view to favor timely and individualized treatment planning [1, 2, 4]. This question is common to a range of cancers (carcinomas in particular), and also arises in the design of novel protocols for clinical trials. The PET-specific set of guidelines, PERCIST, proposed by Wahl *et al* in 2009 [2] attests of its accrued relevance to therapeutic assessment.

Much of the analysis of FDG-PET data is based on standardized uptake values (SUV), which are the image intensity values (PET measured local photon counts) used to characterise metabolic FDG activity within a tumor. SUVs are defined as the ratio of the PET-measured radioactivity (measured in kBq per ml of tissue) to the injected dose (kBq) per unit weight of the patient. These values are dimensionless under the assumption that 1ml of tissue weighs 1gm [7]. Sometimes lean body mass is used in place of the gross weight of the patient. Our analysis is invariant to such scaling. A number of PET image-based tumor biomarkers are obtained from the SUV: $\mathrm{SUV}_{\mathrm{mean}}$, $\mathrm{SUV}_{\mathrm{max}}$, $\mathrm{SUV}_{\mathrm{peak}}$, and volume-based quantitators such as metabolically active tumor volume (MATV) and total lesion glycolysis (TLG), which is the product of $\mathrm{SUV}_{\mathrm{mean}}$ with MATV [2]. $\mathrm{SUV}_{\mathrm{mean}}$ and TLG both implement averaging over a selected metabolically active volume. $\mathrm{SUV}_{\mathrm{max}}$ and

[1] *School of Mathematical Sciences, University College Cork, Cork, Ireland*
[2] *Department of Radiology, University of Alabama, Birmingham, USA*
* *Correspondence to: eric.w@ucc.ie*

 Copyright © 2016 John Wiley & Sons, Ltd.

SUV_{peak} are also widely considered for baseline diagnosis and staging. The latter are however sensitive to voxel-level variability in measurements. In their landmark review [2], the authors of PERCIST discuss over thirty different methods for therapeutic assessment, and conclude that SUV-based approaches emerge as the "most widely applied, generally correlating well with more complex approaches".

Early assessment of tumor response to treatment by PET involves the comparison of pre- and post-treatment imaging information. Most typically, SUV_{mean}, SUV_{peak} or TLG, depending on the disease and treatment center, are computed independently within predefined volumes of interest (VOI) taken from the pre- and post-therapy sets. These measures are then compared to form an assessment of patient response [1, 2, 8, 9, 10, 11, 12, 13]. Whatever the choice of index, an assessment of proportional change in quantitation (percentage-change) is usually preferred over a comparison of absolute change values.

Other quantitators, and in particular intratumoral heterogeneity, are also gaining interest, although they are not yet applied routinely in cancer care. Macro-level heterogeneity can be assessed from PET imaging data and a number of heterogeneity quantitation techniques have established relationship with patient outcome in a number of diseases [6, 14, 15, 16], which demonstrates that a more detailed exploitation of PET imaging data is relevant to modern medical and clinical practice. The literature now abounds with analytical image processing procedures for texture analysis [17, 18, 19, 20]. Notably, a number of these techniques consist in deriving statistical variables at the voxel level [14, 16, 19, 21].

1.2. Need to evaluate reliability of response assessment

Clinical utility of these image-derived biomarkers is validated by correlation with endpoints such as patient outcome or therapeutic response. These methodologies have an established prognostic utility in the diagnostic and treatment of many types of cancers. They are however often considered without an associated assessment of variability, or this assessment may be performed naively in an unpaired fashion (using e.g. bootstrapping), which we suggest is not appropriate in this context. Many studies illustrate this that may be found with a simple Pubmed search. In [22, 23] for example, the authors present thorough statistical analyses, but ones that exclude an individual assessment of variability in the change of SUV_{mean} for each patient (studies usually quote the variability of quantitation of change across subcohorts, which only serves to summarize the distribution of measures of change in SUV_{mean} for those cohorts, not for a given patient). This general trend is observed in spite of recurrent considerations on test/retest procedures for data reproducibility (see e.g. in [2]), which are indicative of the general concern that PET imaging data must be statistically reliable. At the patient level, there is a practical interest in having a mechanism to routinely assess the validity and significance of imaging biomarker values for treatment effects. This quantitation may be used clinically in its own right, as a novel variable. Such information would also provide further opportunities for patient-adaptive care, as knowledge of the degree of reliability of the assessment of therapeutic effectiveness could be useful for subsequent treatment path selection.

The contribution of this paper is twofold. First, the paper develops a statistical model that aligns with the use of percent-change in SUVmean (which is an important routine marker, and part of the PERCIST guideline), from which we derive a measure of associated standard error. We are not aware of similar statistical developments to support existing response quantitation methods elsewhere. The statistical framework has the pseudo-Poisson property and is based on data pairing, in a concept further developed in Section 1.3. The paper also demonstrates that data pairing (i.e. scan co-registration) is not unpractical and should therefore be considered over naive unpaired quantitation. Second, we consider using the derived quantitation of paired standard error for prognostic assessment, therefore proposing an outcome variable that has not yet been considered elsewhere, to the best of our knowledge.

1.3. Assessing variability in response quantitation

Beyond the possibility to derive approximate confidence intervals, an evaluation of the variability associated with this assessment would be useful in terms of further characterization of the metabolic information. In particular, the standard error associated with a sample mean uptake may help discriminate between homogeneous and heterogeneous uptake distributions with otherwise comparable mean values. In this view, and as in other models for treatment effect in medical statistics, pairing is advocated here in order to integrate the fact that the same biological volume of interest, and thus a common structure of PET pseudo-count rates, is imaged at different time points, when assessing variability. In a generic experimental setting, data pairing arises in situations where the statistical experiment consists in making two separate observations on each subject from a sample of n such subjects, such as "before-and-after" clinical trial studies. Pairing consists in analysing the (fixed) individual effects, i.e. the difference between the observations obtained for each individual [24, 25, 26]. We may still view the pre- and post-therapy errors as independent, but the common underlying structure proscribes the use of a naive variance estimator.

In terms of PET imaging data obtained at multiple time points from a single patient, if the pre- and post-treatment scans are properly aligned (i.e. co-registered), it is possible to make voxel-level comparisons between them (a voxel,

Copyright © 2016 John Wiley & Sons, Ltd.

or volume element, is the 3D equivalent of an image pixel). Methodologically, the resulting data configuration can be considered as a paired study in which the experimental units are image voxels. The obvious statistical interest in pairing the treatment data is thus to exploit more detailed levels of information available at the voxel-level, on metabolic variations happening at a same physiological location. Elaborate co-registration techniques [27, 28] exist that make this general pairing principle implementable in practice. Co-registration of PET/CT or PET/MR acquisitions onto the initial scan are often considered, as in [29]. The attenuation (CT) image routinely obtained with PET scanning may also be used for this purpose. A recent study by Guo et al [30] proposed a paired voxel-level analysis for brain tumor imaging data. This work draws on the incorporation of additional MR scan information which would not be routinely available in many clinical settings where FDG-PET is used to follow tumor response. In any case, although co-registration of PET imaging data cannot be perfect, we argue in what follows that there is reason to utilizing even mis-registered information in image-based assessment of therapeutic response. In the validation study described in Section 4, where the approach is applied to clinical sarcoma datasets, this co-registration step is in fact performed manually, using expert guidance. Appropriate assessment of quantitation variability may be used to refine the scoring of risk associated with a given level of therapeutic response.

We suggested above that a paired assessment of voxel-response variability, σ_P, may be interpretable. Assuming acceptable scanner co-registration, σ_P captures an element of spatial variability in the difference in activity pre- and post-therapy, since a more heterogeneous spatial distribution is likely to yield higher variability of the observed paired differences than would a more homogeneous pattern. Adjusting the proportional change in uptake means for spatially-relevant variability thus yields an interpretable test-like, normalized statistical variable. This assessment can complement that of proportional change as it assays the level of homogeneity of the magnitude of voxel changes throughout the volume. Unpaired assessment, on the other hand, does not have this potential. Potential use of the paired standard error quantitation with respect to intratumoral heterogeneity is further discussed in Section 5.

The first contribution of this paper is the derivation in Section 2 of closed-form expressions for paired and unpaired quantitation of metabolic change, which are derived from a model for the global change in population mean allowing for marginal variations at the voxel level, with a restriction to univariate treatment effect. We incorporate the fact that PET data have count-like variation [31, 32] in the analysis, using an approach based on a Gamma distribution with the pseudo-Poisson assumption. Under this model, maximum likelihood solutions yield simple closed-form solutions. As we will see, modelling the pre- and post-therapy uptake data using a Gamma model (with the pseudo-Poisson characteristic) actually corresponds to the analysis of the percentage change in population means, one of the predominant indexes considered in the field. This aligns with the current PERCIST guidelines [2] for PET-based therapeutic response assessment. It is worth noting that the other main indexes used in the community also implement a form of averaging (total lesion glycolysis uses averaging over the whole sample, and SUV_{peak} is the average of standardized uptake values in a neighborhood of the most intense voxel). Other techniques such as (paired or unpaired) t-tests [24, 25, 26, 33, 34], which analyse absolute differences in population means rather than proportional ones, and other parametric or semi-parametric alternatives [35, 36, 37, 38] may be considered too but they do not conform to standard practice.

In Section 3, numerical experiments are used to illustrate specific aspects related to the role of the approximated associated paired standard error, including its impact on test power. Both Sections 2 and 3 also provide a conceptual examination of the impact of mis-alignment of data during the pairing process, which is a realistic prospect – especially when analyzing spatial information defined on a multidimensional grid, such as “before and after” images or medical scans.

Validation of the principle of paired imaged-based therapeutic assessment, via multivariate survival analysis, is exposed in Section 4. We consider an application to the non-invasive assessment of therapeutic response in patients with sarcoma, a rare malignant disease with (typically) slow-growing solid masses that may or may not respond well to chemotherapy prior to resection, depending on tumor subtype. Investigative work was carried out on a cohort of 50 clinical sarcoma studies acquired by Positron Emission Tomography at the University of Washington School of Medicine. The paired, test-like variable mentioned above is included in a multivariate risk analysis in order to query its potential relationship with cancer patient outcome. This study exhibits a potential benefit in incorporating this statistical quantitation of response performed at the voxel level for both survival and disease progression.

2. Paired assessment of the treatment effect

In what follows we consider the problem of comparing two paired volumes of interest, namely Ω_0 and Ω_1, with voxel uptake data $\{(\mathrm{x}_j, Y_j)_i\}_{i=1}^N$, with $j = 0, 1$ labeling respectively the pre- and post-therapy datasets, $\mathrm{x}_j \in \mathbb{R}^3$ denoting voxel coordinates and Y_j the corresponding set of measured uptake levels (i.e. voxel intensities). In this context, by pairing we understand that the two volumes Ω_0 and Ω_1 are obtained by co-registration of the two imaging sets, before VOI boundaries

be drawn over the co-registered paired of images. These two regions therefore comprise of the same set of N voxels, i.e. voxel coordinates are such that $\mathrm{x}_{0i} = \mathrm{x}_{1i} = \mathrm{x}_i, i = 1, \ldots, N$.

2.1. *Log-linear modelling of treatment effect*

If the pre- and post-treatment data (Y_0, Y_1) were Gaussian random variables, then the standard paired t-test analysis can be applied provided the voxel values are independent conditional on (a potentially random but unobservable) voxel effect. However such assessment procedures are not robust to cases that strongly violate the Gaussian assumption (they also do not align with current medical methodologies for image-based therapeutic assessment). Here we consider modelling Y_{ji} via the pre- and post-treatment rates λ_{ji} at unit voxel x_i $(i = 1, \ldots, N; j = 0, 1)$ assuming the variances are proportional to the means, i.e.

$$\mathrm{Var}(Y_{ji}) = \phi \mathrm{E}(Y_{ji}) = \phi \lambda_{ji} \tag{1}$$

where $j \in \{0, 1\}$ denotes respectively pre- and post-treatment information, and where ϕ denotes a positive dispersion factor [39], $\phi < 1$ and $\phi > 1$ indicating under- and over-dispersion respectively. The effect of treatment may then be represented by an overall factor γ in the relationship between the pre- and post-treatment observations, i.e. $\lambda_{1i} = \gamma \lambda_{0i}$. For simplification, we can pose that $\lambda_{0i} = \lambda_i$ denotes unknown unit voxel effect at baseline. In this framework, therapeutic effect γ may be estimated assuming that pre- and post-therapy uptake variables both have the pseudo-Poisson property (1). To this end we found that a Gamma distribution has the potential to represent the variance of the uptake information adequately (a justification is provided in Appendix C). We model the two-sample comparison problem for $i = 1, \ldots, N$ with

$$Y_{0i} \mid \lambda_i \sim \mathcal{G}\left(\frac{\lambda_i}{\phi_0}, \phi_0\right), \qquad Y_{1i} \mid \lambda_i \sim \mathcal{G}\left(\frac{\gamma \lambda_i}{\phi_1}, \phi_1\right) \tag{2}$$

where the $\{\lambda_i\}_{i=1}^N$ describe individual voxel effects and γ represents the global change in uptake following therapy. Conditional on λ_i, observations Y_{0i} and Y_{1i} are independent. We may equally consider the problem of estimating β in $\log(\lambda_{1i}) = \beta + \log(\lambda_{0i}), i = 1, \ldots, N$. Here we write the pdf for the Gamma distribution $\mathcal{G}(\alpha, S)$ in terms of its shape α and scale S parameters (both strictly positive) as

$$f(y; \alpha, S) = \frac{1}{\Gamma(\alpha) S^\alpha} y^{\alpha - 1} \exp\left(-\frac{y}{S}\right)$$

($\Gamma(.)$ denoting the Gamma function). Under this parametrization, we have $E(Y) = \alpha S$ and $\mathrm{Var}(Y) = \alpha S^2$, which, in combination with (2), satisfies (1). The choice of the Gamma distribution to model the pre- and post-therapy uptake data (beyond the fact that it provides the pseudo-Poisson property) is further motivated in Appendix C, where we assess the adequacy of this model against standardized PET data designed specifically for the purpose of universal PET scanner calibration.

2.1.1. Interpretation of model scale. In (2) the parameters ϕ_j quantify the scale of the uptake data, and under (1) adjust the level of overdispersion. In terms of PET imaging data, for a given patient this scale is primarily governed by radiotracer dose per unit weight of patient, scanner technology, and physical characteristics of the patient including attenuation and motion. Of these dose is most important. Care is taken to ensure that the dose per unit mass of the patient is the same for pre- and post-therapy scans. Current guidelines advise against follow-up imaging being performed on a different scanner, see e.g. RECIST [40, p. 242 and 246] and PERCIST [2, p. 133S] guidelines. Moreover, there is usually an attempt in routine practice to achieve the same level of dose at different scanning time points, so that the two attenuation scans be comparable. It is therefore reasonable to assume that $\phi_0 \approx \phi_1$ given two uptake samples obtained at reasonably close scanning time points for a same patient and with the same PET scanner. It is also reasonable to consider that two patients would have two different ϕ characteristics. With this in mind we consider the slightly simpler modelling approach

$$Y_{0i} \mid \lambda_i \sim \mathcal{G}\left(\frac{\lambda_i}{\phi}, \phi\right), \qquad Y_{1i} \mid \lambda_i \sim \mathcal{G}\left(\frac{\gamma \lambda_i}{\phi}, \phi\right) \tag{3}$$

In particular this second model allows for paired estimation of the quantity ϕ for a given patient.

2.1.2. Interpretation of rates λ and relation to pairing. In (3) the $\{\lambda_i\}$ describe fixed voxel effects, i.e. they constitute the invariant underlying structure of glucose uptake in the tissue that is being imaged. Their invariance between two imaging time points implies that overall uptake changes are captured by γ. These quantities must thus be estimated along with γ. The pre- and post-therapy uptake samples Y_0 and Y_1 may be assumed independent conditionally on the structure λ, since the random errors in measurements result mainly from data acquisition and may be assumed to be independent at separate

Copyright © 2016 John Wiley & Sons, Ltd.

injection time points. As such, "pairing" is here understood as the co-registration of the two imaging sets so as to capture the same set of underlying $\{\lambda_i\}$. Note that the standard assumption underlying the usual paired t-test involves $Y_{1i} - Y_{0i}$ being i.i.d. (Gaussian) with common mean. In the present case, this is not true, as the mean of $Y_{1i} - Y_{0i}$ is dependent on the unknown voxel effect λ_i.

2.1.3. Assumption of spatial independence. In this paper we make the underlying assumption of spatial independence of the count rate data $\{\lambda_i\}_{i=1}^N$, and furthermore that of uptake data $\{Y_j\}_{i=1}^N, j = 0, 1$. In other words we assume here that the realizations within a sample Y_j are independent of each other. In this sense our derivations correspond to the case of an ideal PET scanning modality coupled with perfect reconstruction. Application to real PET imaging data requires the incorporation of spatial correlation κ in order to obtain a more accurate patient-specific interpretation of quantities of interest. This is discussed further in Section 2.5.

2.2. Closed-form expression for the therapeutic effect

The estimation problem defined by (1) and (3) allows the derivation of approximating expressions of maximum likelihood estimators for ϕ, λ and γ. We refer the reader to Appendix A for technical details on derivations of the following quantities; this section is a summarised presentation of the closed-form analytic expressions we obtained. Based on (3) and under the assumptions presented above, a general expression for the log-likelihood for the two-sample problem with $Y = (Y_0, Y_1)$ is easily obtained and joint maximum likelihood solutions for λ and γ take the form

$$\hat{\gamma} = \frac{\sum_{i=1}^N Y_{1i}}{\sum_{i=1}^N Y_{0i}} = \frac{\bar{Y}_1}{\bar{Y}_0} \tag{4}$$

where $\bar{Y}_j$ denotes the sample mean of the observations $\{Y_{ji}\}_{i=1}^N, j = 0, 1$. Expression (4) thus evaluates the percentage change in mean measurements before and after treatment. Note that this estimator is unaffected by pairing – the same quantity can be used to quantify treatment effect in an unpaired scheme. From (4) we also derive $\hat{\beta} = \log(\hat{\gamma})$. The marginal effects are estimated jointly with (4) for each voxel $i = 1, \ldots, N$, by

$$\hat{\lambda}_i = \frac{Y_{0i} + Y_{1i}}{1 + \hat{\gamma}} \tag{5}$$

We now propose two finite-sample evaluations of the maximum likelihood estimator of scale parameter ϕ, obtained from paired analysis for model (3). A first estimator is

$$\tilde{\phi} = \frac{1}{N\hat{\gamma}} \sum_{i=1}^N \frac{(\hat{\gamma} Y_{0i} - Y_{1i})^2}{Y_{0i} + Y_{1i}} \tag{6}$$

For high counts, ϕ may be approximated instead with

$$\hat{\phi} = \frac{1}{N} \sum_{i=1}^N \frac{(\hat{\gamma} Y_{0i} - Y_{1i})^2}{\hat{\gamma}^2 Y_{0i} + Y_{1i}} \tag{7}$$

Recall that (6) and (7) are two different paired assessments of the quantity ϕ; they become equivalent at higher counts, as demonstrated in Appendix B.1. If one had reasons to doubt that $\phi_0 \approx \phi_1 \approx \phi$, then ϕ_0 and ϕ_1 may be estimated independently from samples $\{Y_{0i}\}_{i=1}^N$ and $\{Y_{1i}\}_{i=1}^N$ respectively, e.g. using the pseudo-Poisson assumption (1) with sample approximations for the quantities $\phi_j = \frac{\mathrm{Var}(Y_{ji})}{E(Y_{ji})}$. This possibility is further explored in Appendix B.2. In the above expressions we also use the whole tumor volume for evaluation. Alternative estimates for ϕ may be obtained using a reliable subset of the specified volume of interest. This may be helpful for instance in situations where co-registration is found particularly difficult.

2.3. Assessment of the variability associated with $\hat{\beta}$

In Appendix A we also derive closed-form expressions for estimators of the variances associated with the estimator $\hat{\beta}$,

$$\hat{\beta} = \log(\hat{\gamma}) = \log(\bar{Y}_1) - \log(\bar{Y}_0) \tag{8}$$

By the Delta method,

$$\sigma^2_{\hat{\beta}} \approx \frac{\phi}{N\bar{\lambda}} \left(1 + \frac{1}{\hat{\gamma}}\right) \tag{9}$$

Copyright © 2016 John Wiley & Sons, Ltd.

where $\bar{\lambda}$ is the sample mean of the estimated marginal effects $\hat{\lambda}_i$ given by (5). This expression is obtained under the assumption that $\bar{\lambda}_1 = \gamma\bar{\lambda}_0$. It may be approximated using uptake samples and e.g. (7) by

$$\hat{\sigma}^2_{\hat{\beta}} \approx \frac{\hat{\phi}}{N}\left(\frac{1}{\bar{Y}_0} + \frac{1}{\bar{Y}_1}\right) \tag{10}$$

A naive estimator if variance is

$$\hat{\sigma}^2_{\hat{\beta},U} = \frac{\mathrm{Var}_N(Y_0)}{\bar{Y}_0^2 N} + \frac{\mathrm{Var}_N(Y_1)}{\bar{Y}_1^2 N} \tag{11}$$

where $\mathrm{Var}_N(Y_j)$ is the usual sample variance. The appendix shows that

$$\sigma^2_{\hat{\beta},U} \approx \frac{1}{\bar{\lambda}N}\left(\left(1+\frac{1}{\gamma}\right)\phi + \frac{2\xi(\lambda)}{\bar{\lambda}}\right) \tag{12}$$

where $\xi(\lambda) = (\sum_{i=1}^N(\lambda_i - \bar{\lambda})^2)/N$ is the variation in count rates λ_i. Constructions (8), (10) and (11) are clearly invariant to a change in scale of observations Y. Note that if $\xi(\lambda) = 0$ (λ_i constant) then $\sigma^2_{\hat{\beta},U}$ reduces to $\sigma^2_{\hat{\beta}}$, but otherwise this naive estimator will be positively biased.

Evaluation of the naive standard error may be performed empirically in a number of ways, depending on the choice of a definition for $\mathrm{Var}_N(.)$. For example the sample quantities $\mathrm{Var}_N(\bar{Y}_0)$ and $\mathrm{Var}_N(\bar{Y}_1)$ may be bootstrapped directly to evaluate $\mathrm{Var}(\hat{\beta}) = \mathrm{Var}(\log(\bar{Y}_0)) + \mathrm{Var}(\log(\bar{Y}_1))$ (but not surprisingly, this will produce the approximation (11)). Based on variance estimators for $\hat{\beta}$the variance of $\hat{\gamma}$ can be derived (using the Delta method). This gives

$$\hat{\sigma}^2_{\hat{\gamma}} = (\exp(\hat{\sigma}^2_{\hat{\beta}}) - 1))\exp(2\hat{\beta} + \hat{\sigma}^2_{\hat{\beta}}) \tag{13}$$

and

$$\hat{\sigma}^2_{\hat{\gamma},U} = (\exp(\hat{\sigma}^2_{\hat{\beta},U}) - 1))\exp(2\hat{\beta} + \hat{\sigma}^2_{\hat{\beta},U}) \tag{14}$$

2.4. *Quantities of interest*

2.4.1. Confidence intervals around $\hat{\gamma}$. For N large and under usual regularity conditions, (10) may be used in the Gaussian approximation (by the Central Limit Theorem) for the maximum likelihood estimator $\hat{\beta}$ with mean β,

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2_{\hat{\beta}}\right) \tag{15}$$

with $\sigma_{\hat{\beta}} = E(\hat{\sigma}_{\hat{\beta}})$. Quantities (13) and (14) are of particular interest as they allow to construct approximate confidence intervals for the measure of percentage change in the mean, defined as $\hat{\gamma} - 1 = (\bar{Y}_1 - \bar{Y}_0)/\bar{Y}_0$. Based on the Gaussian approximation (15), confidence bounds for this popular quantitator may be derived from those obtained for $\hat{\beta} = \log(\hat{\gamma})$ using (4) and (10) in $[b_L; b_U]$ where $b_L = \hat{\beta} - z_\alpha\hat{\sigma}_{\hat{\beta}}$, $b_U = \hat{\beta} + z_\alpha\hat{\sigma}_{\hat{\gamma}}$, and $z_\alpha = \Phi^{-1}(1-\alpha/2)$, Φ denotes the cumulative distribution function of the standard normal distribution, and $[b_L, b_U]$ is a $(1-\alpha)\times 100\%$ confidence interval for β. Then the exponential transform of these bounds $[g_L = \exp(b_L), g_U = \exp(b_U)]$ may be applied to the more interpretable value $\hat{\gamma}$. Note that in light of discussions in Section 2.1.3 and further in Section 2.5, confidence intervals would be biased unless spatial correlation was incorporated in the assessment.

2.4.2. Hypothesis tests for the percentage change in mean uptake. Qualitative assessment of therapeutic effectiveness may also be performed, based for example on the result of a two-sided test for metabolic change (with the null hypothesis of no change in mean, i.e. $\beta = \beta_0$), or that of a one-sided test for partial metabolic response (under the null hypothesis that the percentage change in mean is not lower than some clinical threshold of interest, i.e. $\beta \geq \beta_0$). We may construct a paired test based on the statistic

$$t_{QP} = \frac{\hat{\beta} - \beta_0}{\hat{\sigma}_{\hat{\beta}}} \tag{16}$$

for some null assumption on the value of β_0 (e.g. $H_0 : \beta_0 = 0$ for the two-sided test mentioned above), using (4) and (10) – see for instance [24] for details on the construction of paired tests of hypotheses. Based on the previous Normal approximation for $\hat{\beta}$, t_{QP} can be assumed to be t-distributed. A test based on $\hat{\gamma}$ may be constructed similarly. Note that t_{QP} and γ are not linearly related.

Copyright © 2016 John Wiley & Sons, Ltd.

2.4.3. Performance ratio for a paired analysis. The benefit conferred by data pairing in measuring the accuracy of the estimate $\hat{\beta}$ may be evaluated by

$$\frac{\hat{\sigma}_{\hat{\beta},U}}{\hat{\sigma}_{\hat{\beta}}} = \left(\frac{\bar{Y}_0^2 \mathrm{Var}(\bar{Y}_1) + \bar{Y}_1^2 \mathrm{Var}(\bar{Y}_0^2)}{\hat{\phi}(\bar{Y}_0^2 + \bar{Y}_1^2)} \right)^{\frac{1}{2}} \tag{17}$$

The ratio of $\sigma_{\hat{\beta},U}$ over $\sigma_{\hat{\beta}}$ boils down to

$$\frac{\sigma_{\hat{\beta},U}}{\sigma_{\hat{\beta}}} = \left(1 + \frac{\xi(\lambda)}{\phi \bar{\lambda}} \right)^{\frac{1}{2}} \tag{18}$$

This ratio does not depend on global response γ. The ratio

$$\frac{\hat{\sigma}_{\hat{\gamma},U}}{\hat{\sigma}_{\hat{\gamma}}} \tag{19}$$

can be used to evaluate the impact of a more careful assessment of variance on $\hat{\gamma}$ inference.

2.5. Spatial dependence of imaging uptake data

As stated earlier in Section 2.1.3, the above expressions are derived using the underlying assumption of spatial independence of uptake data $\{Y_j\}_{i=1}^N, j = 0, 1$, thus assuming ideal and perfectly reconstructed PET imaging data. Application to real PET imaging data requires the incorporation of spatial correlation κ in order to obtain a more accurate patient-specific interpretation of $\hat{\sigma}_{\hat{\beta}}$ (for example when using confidence intervals on the percentage-change in means). Note that by not including a correction for spatial correlation in assaying $\hat{\sigma}_{\hat{\beta}}$, we under-estimate assessment variability. We further discuss this parameter in concluding Section 5.

Excluding κ from the quantitation is however not critical with regards to the significance of quantity t_{QP}, since in that construction the standard error term is used merely for standardizing the measure of proportional change. Omitting the spatial correlation in this normalization does not affect the assessment of direction of this proportional change, and as such it does not impact the degree of alignment of the risk covariate with patient outcome information. Inclusion of κ is thus not mandatory in terms of defining a quantitation that correlates with patient survival or disease progression. Therefore one could consider t_{QP} rather than $\hat{\sigma}_{\hat{\beta}}$ for assessment under the assumption of spatial independence. Performance ratios (17) and (19) also remain relevant since the correction for within-sample spatial correlation is identical in the paired and unpaired cases.

2.6. Characterisation of the impact of mis-alignment

The process of pairing two sets of observations implies that pairs of measurements be made according to a common basis – in the case of PET imaging data, the two imaged sets of observations are aligned (or co-registered) according to a common grid of spatial coordinates (e.g. of a same voxel in three-dimensional scans acquired before and after therapy). This process is often complex and likely to introduce sub-grid approximations (e.g. interpolation), which introduces errors in the paired data model (see e.g. [41]). In the case of co-registration of two scans of a cancer patient before and after therapy, a number of factors including patient position, tumor evolution, and impact of treatment on physiology, render the notion of "optimal co-registration" difficult to define, if not devoid of foundation. Nonlinear discrepancies resulting from mis-alignment are not easily summarized in a model. Mis-alignment however typically occurs at a magnitude small enough to allow for pairing to remain a plausible approach.

Let us consider the case where the baseline dataset Y_0 is aligned onto Y_1, and where we compute $\hat{\beta}$ based on mis-registered data Y_0^Δ. Recalling our earlier assumption that $\bar{\lambda}_1 \approx \gamma \bar{\lambda}_0$, and assuming that mis-registration preserves most of the content present in the volume of interest with Y_0^Δ, where superscript Δ denotes mis-registration, then it is realistic to accept that $\bar{\lambda}_1 \approx \gamma \bar{\lambda}_0^\Delta$. With the latter approximation, and given (10), the variance of the estimator of therapeutic effect derived from mis-registered data (say, $\hat{\sigma}^2_{\hat{\beta},\Delta}$) may be considered as follows:

$$\sigma^2_{\hat{\beta},\Delta} = \phi^\Delta \left(\frac{1}{N\bar{\lambda}_0^\Delta} + \frac{1}{N\bar{\lambda}_1} \right) \approx r_\Delta \sigma^2_{\hat{\beta}} \tag{20}$$

with $r_\Delta = \frac{\phi^\Delta}{\phi}$. From (20), the effect of mis-registering the pre- and post-treatment data is a scaling of the standard error $\sigma_{\hat{\beta}}$ (our measure of assessment accuracy) achieved in the case of "ideal" registration. As before, for γ-based quantitation the standard error in the presence of mis-registration error becomes $\hat{\sigma}^2_{\hat{\gamma},\Delta} = (\exp(\hat{\sigma}^2_{\hat{\beta},\Delta}) - 1)\exp(2\hat{\beta} + \hat{\sigma}^2_{\hat{\beta},\Delta})$. In the following section we illustrate the impact of registration error and demonstrate how a paired analysis of therapeutic assessment standard error based on mis-registered information may still be preferable over an unpaired analysis.

Copyright © 2016 John Wiley & Sons, Ltd.

3. Illustrative experiments

We now provide an illustration of principle of the proposed paired assessment methodology for treatment effect. With the following implementation we do not simulate response data using established PET simulators, which would require deploying a computationally significant process, see e.g. [42]. Validation of our approach is presented in Section 4 with an application to real PET imaging data. However, the data simulation model used here is justified in Appendix C where the produced datasets are shown to compare well with real output PET imaging information.

Below we use a simplified framework on simulated 2D images to demonstrate the effect of pairing imaging data over an unpaired analysis, including in the presence of mis-registration. A first experiment was conducted using an overdispersed model (Section 3.1) to illustrate how pairing improves the evaluation of accuracy (Section 3.2). Then we evaluated the impact of mis-registration in two different experimental frameworks (Section 3.3).

3.1. Experimental framework

The 16×16 pre- and post-treatment imaging data (i.e. $N = 256$) were simulated according to (3). The marginal voxel rates λ_i at voxels $\mathrm{x}_i = (x_1, x_2)_i$ were elliptically distributed within the image, using $\lambda(x_1, x_2) = c_0 + c_1 \exp(-2(x_1, x_2)^T \Sigma (x_1, x_2))$ with $c_0 = 1$, $c_1 = 12$, for the 2×2 covariance matrix $\Sigma = [1, -\rho, -\rho, 1]_{2\times 2}$ with $\rho = 0.5$, and $x_k \in [-a_k, a_k]$ (for $k = 1, 2$), here with $a_1 = 1.5$ and $a_2 = 1.2$. A Gamma distribution with scale $\phi = 1.2$ was used to generate independent realizations of pre- and post-therapy uptake values as

$$Y_{0i} \sim \mathcal{G}(\tau\lambda_i/\phi, \phi); \qquad Y_{1i} \sim \mathcal{G}(\gamma\tau\lambda_i/\phi, \phi) \tag{21}$$

where γ controls the overall change in distribution shape $\tau\lambda_i/\phi$. In this model τ plays a role comparable to the normalization of PET counts by dose concentration (usually carried out to form SUV data in the output PET image), in that for instance, $E(Y_{0i}) = \tau\lambda_i$ and $Var(Y_{0i}) = \phi\tau\lambda_i$, which is consistent with pseudo-Poisson property (1). Figure 1 shows an example of (e.g. baseline) data simulated with Gamma model (21), for varying τ values (i.e. for varying tumor definitions). This figure indicates, in reference to the PET response data example of Figure 2, that the distribution of such simulated uptake data has reasonable alignment with a true PET scenario; further justification for this simulation model is provided in Appendix C. In the experiments described hereafter, we set $\tau = 1$, which yields images with lower signal-to-noise ratio (which would typically reduce the difference $\hat{\sigma}_{\hat{\beta},P} - \hat{\sigma}_{\hat{\beta},U}$, as discussed further), and also allows us to control image contrast with c_1.

As (21) models a linear overall change in voxel rates as in (3), with $E(Y_{1i}) = \gamma E(Y_{0i})$, this simulation framework implements features of our methodology and serves to motivate our approach. We stress here that (3) is merely a proposal representation of dominant changes in metabolic information at successive imaging time points (local adjustments being represented by individual effects λ_i). Its aim is not to capture complex underlying features of a tumor process, such as its biology; however it reflects well what is currently done in practice where overall metabolic changes are summarized by overall (linear) changes in a choice of standard indexes among those described in Section 1.1.

Voxel-level analysis of real PET information is depicted in Figure 2, where the pre- and post-therapy scans are aligned and a common volume of interest Ω (e.g. ellipsoidal) is specified simultaneously for the same set of voxels, based on the largest of the pre- and post-therapy tumor objects. Alignment is performed based on expert human assessment and consists in rigid transformations (3D rotations and shifts) of the post-therapy image to overlay the pre-therapy image.

In the framework described above, the spatial structure of the simulated data is convenient in that co-registration is easily controlled with the ellipsoidal structure. It is worth noting that some real PET datasets present themselves with a similar pattern; it is the case e.g. in homogeneous sarcomas and also in lung cancers. The experiments were repeated $N_S = 5,000$ times to assess the paired assessment approach (with a focus on its statistical power and magnitude of the associated standard error).

3.2. Effect of alternative variance estimators

We compared the powers of the paired and unpaired tests t_{QP} based on (16), derived from the estimates of interest and at the 5% significance level. Figure 3 illustrates the benefit of pairing on power of one-sided tests based on $\hat{\beta}$ of the null hypothesis $H_0 : \beta > \beta_0 = \log(0.7)$, evaluated for a varying response $\gamma \in (0.40, \ldots, 0.90)$ in (21). The solid black line indicates the theoretical power of the test statistic $\mathrm{Power}(t_{QP} \mid \beta) = \mathrm{Prob}\,(t_{QP} < 1.96 \mid \beta)$, using the standard Normal distribution function, and an approximate theoretic expression for the standard error

$$\sigma^2_\beta = \frac{\phi}{\bar{\lambda}_0 N}\left(1 + \frac{1}{\gamma}\right)$$

Copyright © 2016 John Wiley & Sons, Ltd.

derived in the case of ideally paired images, as in (9), but for the true response value γ. The dashed line indicates the unpaired counterpart obtained using approximate theoretic standard error (12) evaluated using the true response, i.e.

$$\sigma^2_{\beta,U} = \frac{1}{\bar{\lambda}_0 N}\left(\phi + \frac{\xi(\lambda_0)}{\bar{\lambda}_0}\right)\left(1 + \frac{1}{\gamma}\right)$$

recalling that $\xi(\lambda_0) = (\sum_{i=1}^N (\lambda_{0,i} - \bar{\lambda}_0)^2)/N$ is the energy associated with the sample image of pre-therapy raw counts $\lambda_0 = \{\lambda_0\}_{i=1}^N$. This figure clearly depicts the theoretical advantage of not using the standard variance estimators in this case. The empirical power of the paired test based on $\hat{\sigma}^2_{\hat{\beta}}$, evaluated as the observed proportion of Monte-Carlo values of the test statistics that were found to be statistically significant at the 5% significance level, aligns with its theoretic counterpart when using the true value ϕ (red, 'x'), thus demonstrating that the sample approximation (10) is adequate. Similarly, the approximating expression for the unpaired standard error is also shown to align with its theoretical counterpart (blue, broken line with fat dots). The power derived from $\hat{\sigma}^2_{\hat{\beta}}$ but using the paired estimate (7) for ϕ, rather than the true value ϕ, has reasonable alignment with the empirical power for the paired test using the exact ϕ; the bias induced by the use of $\hat{\phi}$ decreases as dose level τ increases in (21). The other broken curve corresponds to a mis-registered case and is described further in Section 3.3.

It is worth noticing that the magnitude of the gain obtained from pairing as quantified by ratio (19) does not depend on the true underlying therapeutic response, but rather on the image statistics (which are partly driven by tumor features). For this reason, we describe here another experiment to illustrate the impact of pairing with respect to image contrast. We considered the ratio of standard errors (19) as a function of range of voxel intensities (which may be characterized by c_1 in model (21)), a ratio greater than 1 indicating a benefit in pairing. Figure 4 shows the variation of the mean ratio of standard errors, over $N_S = 5,000$ Monte-Carlo simulations of the null scenario where $\gamma = 0.7$, w.r.t. varying $c_1 \in \{1.5, 3.5, 5.5, \ldots, 17.5\}$, for the ideal paired case (solid line) as well as for various typical mis-registration errors described further. This range of contrast values would correspond to imaging data going from extremely noisy to extremely well defined; the case $c_1 = 12$ of data with an acceptable signal-to-noise ratio is indicated by a vertical dashed line. In this experiment the benefit of pairing is systematic for all contrasts, even under mis-registration. This example also shows that the overstatement of the variance estimate for $\hat{\beta}$ increases with the dynamic range of the data, and confirms one's intuition that the gain yielded by pairing is greater for better-defined imaged objects of interest.

3.3. *Effect of mis-registration*

3.3.1. Simulated experiment. In the experimental setup described in Section 3.1, we simulated mis-registration as a spherically symmetric shift Δ, with $\Delta \sim N(0, \sigma^2_\Delta)$ and where σ_Δ therefore controlled the level of typical mismatch error committed, in terms of number of voxels, for data generated according to experimental model (21). For practical reasons the randomized registration error Δ was capped at 75% of the grid length. Figure 3 (left panel) illustrates the empirical powers for the tests based on $\hat{\gamma}$ obtained for the unpaired and ideal paired cases (i.e. (16) using $\hat{\sigma}^2_{\hat{\beta},U}$ and $\hat{\sigma}_{\hat{\beta}}$ respectively), as well as for a mis-registered scenario ((16) using $\hat{\sigma}_{\hat{\beta},\Delta}$) with $\sigma_\Delta = 4$. Note that here, $\sigma_\Delta = 4$ corresponds to a case of a typical mis-registration error of one quarter of the image. In this example the plots illustrate how the unpaired test generally remains less powerful than the paired ones carried out on mis-registered data. Similar comparisons were obtained for stronger mis-registration errors (e.g. $\sigma_\Delta = 6$) in this illustrative example. Note that such comparisons depend on both the magnitude of Δ and on sample size.

The right panel of Figure 3 provides further illustration of the advantage of pairing. It exhibits two additional (Monte Carlo) curves. The higher of these two new curves (red, fat stars) shows the power of an empirical paired test based on (10) under perfect coregistration but carried out on (Y_0, Y_1^S) where Y_1^S denotes a random "scrambling" of image Y_1, i.e. Y_1^S holds the same content as does Y_1 but this content has been spatially randomly rearranged. The other addition (blue, hollow diamonds) shows the power of the empirical unpaired test based on (11). On this figure the power of the unpaired construction remains lower than that of the "scrambled" version. This extreme example illustrates how pairing may still provide a more useful assessment of standard error than would an unpaired analysis, given the same image content, even in situations where co-registration appears very difficult.

Figures 4 and 5 further illustrate how sub-optimal registration may still be preferable to unpaired assessment, in terms of performance ratio $\frac{\hat{\sigma}_{\hat{\beta},U}}{\hat{\sigma}_{\hat{\beta},\Delta}}$ defined by (18) in the null scenario $\gamma = 0.7$, and for a relatively realistic signal-to-noise level ($\tau = 1$). In Figure 4 we used $\sigma_\Delta \in \{1, 3, 5, 7\}$. The mean ratio remains consistently above 1 for all σ_Δ values. We verified that this distinction increases as τ increases. For the experiment summarized in Figure 5, the typical mis-registration error σ_Δ was varied between 1 and 10 voxel units. The ratio distributions remain significantly over 1, even for extreme cases with a mis-registration scale up to $\sigma_\Delta = 10$. The results here suggest that a paired approach can remain more useful than an unpaired one even in the case of a large mis-registration error. One may note that the performance ratio does not depend on the true value of the response γ used for the simulation.

Copyright © 2016 John Wiley & Sons, Ltd.

3.3.2. Practical experiment. A practical experiment carried out using a dataset of 50 GE-Advance clinical sarcoma studies acquired at University of Washington in Seattle further illustrates the influence of mis-registration on the proposed assessment. Sarcomas are a particular type of malignant tumors that tend to be diagnosed late and may present complex spatial metabolic distributions. Table 1 summarizes some characteristics for this dataset, and Figure 2 illustrates one of these sarcoma studies.

In this experiment, the pre-therapy scans were used as baseline information, and also shifted (using shifts $\Delta \in \{-2, ..., 2\}^3 \subset \mathbb{Z}^3$ successively) to create a mis-registered post-treatment image. Ellipsoidal regions of interest were drawn by hand for the post-treatment images of the primary tumor site, once paired with the pre-therapy ones. In the ideal case $\Delta = (0, 0, 0)$, the estimated response is thus $\gamma = 1$ by construction. The merit of this experiment is that it allows testing the proposed analysis on real PET data, with realistic noise characteristics, and uptake distributions that may depart from experimental model (3).

Figure 6 illustrates the distributions for $\hat{\gamma}$, $\hat{\sigma}_{\hat{\gamma},\Delta}$ and $\hat{\sigma}_{\hat{\gamma},U}$ obtained in all 125 scenarios. The estimated response (ideally $\hat{\gamma} = 1$) over this range of Δ was not biased (leftmost panel). Some rather large errors in therapeutic response quantitation were occasionally obtained for the more extreme registration mismatches; however we found that the mean and median $\hat{\gamma}$ was 1 (s.e. 0.0481). Further summary statistics of $|\hat{\gamma}|$ for this experiment yielded percentiles $P_{0.05} = 0.9439, P_{0.25} = 0.9841, P_{0.50} = 1.0000, P_{0.75} = 1.0160, P_{0.95} = 1.0594$, so this experimental can be considered useful and exploitable. The large errors in $\hat{\gamma}$ result from poor estimation of the ϕ parameter in a few mis-registered scenarios. Finite-sample evaluation of the quantitation standard error (rightmost panel) yielded a performance ratio $\hat{\sigma}_{\hat{\gamma},U}/\hat{\sigma}_{\hat{\gamma},\Delta}$ consistently above 1 (values for this ratio range between 1.047 and 7.128), with medians at 0.0061 (s.e. 0.0136) and 0.0125 (s.e. 0.0266) respectively for $\hat{\sigma}_{\hat{\gamma},\Delta}$ and $\hat{\sigma}_{\hat{\gamma},U}$. These coherent results, obtained from real PET imaging data samples, confirm the illustrations of principle of previous numerical experiments presented up to Section 3.3.1, which were run from proposed simulation model (3).

We also carried out a similar experiment on more severe scenarios. For example, using $\Delta = (3, 3, 1)$, $\Delta = (5, 5, 2)$ and $\Delta = (7, 7, 3)$ yielded results shown in Table 2 that confirmed the benefit of pairing, on average. The performance ratio $\hat{\sigma}_{\hat{\gamma},U}/\hat{\sigma}_{\hat{\gamma},\Delta}$ was above 1 in 84% of cases for this experiment. This complementary experiment also suggests that the bias is only mildly affected by mis-registration (it is in fact rather affected by mis-segmentation of the tumor volume), whereas the assessment standard error will increase with larger errors, as suggested already by the decrease in power in Figure 3. These PET scans achieve a definition of about 4.25 mm per voxel (and about 4.5 mm on the longitudinal axis); mis-registration of five voxels roughly yields a 21.2 mm mismatch, which is considered quite large.

4. Application to therapeutic assessment in sarcoma: prognosis

In this section we consider the problem of evaluating the response to neoadjuvant chemotherapy for the treatment of human sarcomas. With the pseudo-Poisson framework we can adapt the analysis to account for the empirical structure of reconstructed PET data, whose variance is proportional to the mean [31, 43], and analyze the change in the FDG biodistribution at the primary site pre- versus post-therapy, based on the paired tracer uptake information. We applied this methodology to the aforementioned cohort of 50 FDG-PET studies of patients with sarcoma (17 deaths and 22 disease progression events), for which the primary tumor site was imaged at baseline and mid-chemotherapy, under the same treatment plan. A multivariate prognostic model applicable at time of first patient follow-up was used to validate the proposed approach. All data and associated patient outcome information were acquired at the University of Washington School of Medicine.

4.1. Methodology

4.1.1. Data pairing. To allow for paired analysis of ϕ, it is necessary to first co-register the pre- and post-therapy VOIs so as to align the two sets of uptake information at the voxel level. Co-registration of FDG-PET images posterior to their acquisition may be best performed by exploiting the contextual information of a corresponding MRI or CT scan (see e.g. [44]), which would provide detailed physiological landmarks. Here however, we consider that only PET information is available, which provides but crude anatomical features. The pre- and post-therapy scans were co-registered by visual assessment using 3D translations and rotations. Note that rotations of the scans in the 3D plane induce an interpolation of the voxel coordinates, which can be dealt with e.g. by re-binning the latter onto a regular grid.

Following this procedure, tumor delineation can be specified, which may be basic (e.g. ellipsoidal or cylindrical) or elaborate (cf. [14]). The co-registered VOIs are then merged into one region of N voxels such that the output data array contains its 3D coordinates x_i, $i = 1, \ldots, N$, and its corresponding pair of uptake values $Y_i = (Y_{0i}, Y_{1i})$, respectively the pre- and post-therapy voxel-level uptake data within the common VOI. In this study the ellipsoidal VOIs were drawn based on the largest of the pre- and post-therapy tumor objects and so as to retain a minimum of background voxels from

Copyright © 2016 John Wiley & Sons, Ltd.

the corresponding scan. An example dataset is illustrated in Figure 2. VOIs were drawn using the open-source imaging software AMIDE [45].

4.1.2. VOI analysis. Conservative thresholding at SUV>1 was applied to the input VOIs to compute classical comparative measures of percentage change (denoted by prefix '%') in SUV_{mean}, SUV_{max} and SUV_{peak}. In parallel to that, the input VOIs were paired (see Section 4.1.1) before double-thresholding at SUV > 1 was applied, in order to remove those voxels with low SUV in both sets of images. This allowed deriving measures for the standard paired and unpaired t-tests, and for the therapeutic response $\hat{\gamma} - 1$ using the exponential of (8), and associated paired (13) and unpaired (14) estimated standard errors, denoted respectively $\hat{\sigma}_{\hat{\gamma},P}$ and $\hat{\sigma}_{\hat{\gamma},U}$ hereafter. SUV_{peak} was computed as the average uptake in the 27-voxel cubic neighborhood of the hottest voxel. The other variables included for analysis were gender, age at baseline, tumor grade (low or high) and tumor site (trunk or limb). Analyses were performed using R [46].

4.2. Measured accuracy

We assessed the effect of pairing in terms of the ratio $\hat{\sigma}_{\hat{\gamma},U}/\hat{\sigma}_{\hat{\gamma},P}$ of standard errors measured without and with pairing. The first histogram of Figure 7 indicates that a finer assessment of accuracy was obtained in 98% of cases thanks to pairing the pre- and post-therapy information (without attempting to refine the initial co-registration). The only case for which $\hat{\sigma}_{\hat{\gamma},U} < \hat{\sigma}_{\hat{\gamma},P}$ yielded $\hat{\sigma}_{\hat{\gamma},U}/\hat{\sigma}_{\hat{\gamma},P} = 0.967$ (with $\hat{\sigma}_{\hat{\gamma},U} \approx 0.0062$ and $\hat{\sigma}_{\hat{\gamma},P} \approx 0.0063$) was a very large sarcoma in the leg (ca. 23,700 voxels before thresholding), with drastic reduction in SUV levels (with SUV_{max} declining from 15.7 to 6.7 units). As the tumor volume increased between the two imaging time points, and the spatial uptake pattern was heterogeneous (in particular with a large inactive core), our conservative thresholding approach may have left a large amount of undesired background. A better suited thresholding strategy would help obtain a refined assessment of paired standard error over the unpaired reference, and change the performance ratio in favor of the paired analysis. However as the two quoted values suggest, the assessment of standard error is already very small and therefore useful for further qualitative analysis in either case. The second histogram is concerned with the magnitude of the measured standard errors for the 50 studies. This figure shows that the therapeutic assessment approach we considered appears to be quite reliable for a majority of cases. For all but one case the estimated standard error was found to be lower than 5%. This scale for standard error is relevant to qualitative decision-making processes applied routinely that use assessments based on observed changes in metabolic activity. Note that the one case above the 5% mark had $\hat{\sigma}_{\hat{\gamma},P} = 0.09$ and corresponds to a relatively small volume (339 voxels before thresholding).

4.3. Univariate analysis

Since information on the histopathologic response at primary site was not available for this study, we benchmarked the quantitators of interest independently against patient outcome, in terms of overall survival (OS) and progression-free survival (PFS). To do so we evaluated the area under the curve (AUC) of their empirical, unsmoothed receiving operator characteristics (ROC) curves using the R package pROC, and the AUCs were computed using trapezoids [46, 47]. Note that this approach does not constitute a direct assessment of the performance of response quantitation, as the measures were computed at the primary site only, and patient outcome is likely to be driven by other co-factors also. Table 3 provides the AUCs computed for the main pseudo-markers under comparison. For this experiment, the baseline markers had a weaker connection with outcome than did comparison-based measures – all baseline measures yielded AUCs below 61. The paired assessment $\hat{\gamma}$ achieved the highest AUC, which is in line with the interest it receives in routine practice and motivates further assessment of its standard error. Comparable results were found when using smoothed ROC curves.

We also compared the various markers in terms of univariate prognostic value, via a univariate Cox regression [46, 48]. The results are also shown in Table 3, in terms of likelihood ratio, which quantifies the effect of adding the variable to the baseline survival scenario, and significance of the variable considered, for both OS and PFS. Conclusions drawn from this experiment align strongly with those yielded by AUC comparison. The paired assessment $\hat{\gamma} - 1$ turned out to be significant for both OS and PFS ($p < 0.01$), and also yielded the highest likelihood ratios. The paired tests were not found to be significant for this dataset. All percentage change measures were found to be significant ($p < 0.05$). Tumor site was the only other significant univariate prognostic variable for PFS.

4.4. Multivariate analysis

The prognostic utility for OS and PFS of the 14 covariates considered in this study ($\hat{\sigma}_{\hat{\gamma},U}$ was dropped at this stage) was analyzed via multivariate Cox survival analysis [46, 48]. All continuous covariates were standardized before analysis, i.e. initial measurements $\{x_c\}_{i=1,\ldots,50}$, where c labels the covariate, e.g. $\text{SUV}_{\text{max},0}$, were transformed into $\tilde{x}_c = (x_c - \bar{x}_c)/\sigma_{x_c}$, where $\tilde{x}_c$ and σ_{x_c} denote mean and standard deviation of the values of covariate x_c. We first performed model comparison

Copyright © 2016 John Wiley & Sons, Ltd.

[49] over the 50 datasets on the basis of both Akaike's Information Criterion (AIC) minimization and likelihood cross-validation (LCV), so as to identify subsets of covariates deemed most valuable for prognosis. We found that the AIC-optimal subset of covariates was (age, site, $\hat{\gamma}-1$, $\text{SUV}_{\text{mean0}}$, t_{QP}, $\hat{\sigma}_{\hat{\gamma},P}$). In parallel, selecting the subset (age, site, $\hat{\gamma}-1$, $\text{SUV}_{\text{mean0}}$, $\hat{\sigma}_{\hat{\gamma},P}$) optimized LCV. For both AIC and LCV, a number of models may have been selected as their scores were comparable, within reason, to those achieved by the optimal set of covariates; all included $\hat{\gamma}$ and $\hat{\sigma}_{\hat{\gamma},P}$. This model analysis highlights the relevance of including a measure based on $\hat{\sigma}_{\hat{\beta}}$ in a multivariate description of patient risk for this study, whether directly as a standalone variable or indirectly e.g. in the form of a test-like statistic such as t_{QP}.

Furthermore, our multivariate Cox analyses found proportional change quantitation $\hat{\gamma}-1$ to be statistically significant (it is the case in particular for subsets of covariates optimizing LCV and AIC, which yielded statistical significance of $\hat{\gamma}-1$ ($p<0.001$ for both OS and PFS, and for both survival models). This further motivates our proposal to include an assessment of its variability to help enhance treatment planning and other medical decision-making aspects.

Hereafter we will focus on a risk model that includes t_{QP}, in order to focus on its contribution to the description of risk of death and of disease progression. Our choice to consider the test-like statistic and exclude using the assessed standard error as a standalone covariate from here on is further motivated by our discussion in Section 2.5.

A multivariate Cox analysis obtained using covariates age, site, $\text{SUV}_{\text{mean0}}$, $\hat{\gamma}-1$ and the paired test statistic t_{QP} is shown in Table 4. All covariates in the model were found to be significant for at least one of OS or PFS; t_{QP} was found to be statistically significant at the 5% level in both. The last row indicates the value of the likelihood ratio (LR) when comparing the likelihoods associated with the multivariate models with and without t_{QP}, along with the corresponding p-value of the likelihood ratio test using a χ^2-distribution with 1 degree of freedom. The likelihood ratios both clearly indicated that the paired test statistic was a significant addition to the prognostic model. The model also yields high concordance ($C=0.79$ and $C=0.75$ respectively for OS and PFS).

Figure 8 shows Kaplan-Meier curves (a nonparametric estimate of hazard rates) for high- and low-risk groups as identified by the median hazard within the cohort of studies [48]. It illustrates the benefit of incorporating t_{QP} (solid blue curves) in the second survival model as it yields a clearer distinction between low- (higher curves) and high-risk groups (lower curves). The separation between survival curve estimates (i.e. between low- and high-risk groups) became statistically significant, in terms of a log-rank (Mantel-Haenszel) test [50], when including t_{QP} in the multivariate model.

These survival analyses, including survival model selection, suggest that the proposed paired assessment is relevant for the analysis of prognosis. Further analysis of the impact of delineation could be the focus of another application, with the inclusion of automatic segmentation procedures such as those of [14, 51]. Robustness of measures based on the SUV_{max} and SUV_{peak} to the choice of delineation procedure would likely play a role in this context. Note that the assessment considered here is conducted at the primary site only. A surrogate marker for metastasis information, such as surgical margin information or tumor heterogeneity, could be incorporated in the validation model in order to mark the presence of secondary sites. In this respect, we believe that the proposed assessment of estimation accuracy (using the derived standard errors) will provide an indication of spatial heterogeneity.

5. Discussion

Positron emission tomography using ^{18}F-fluorodeoxyglucose (FDG-PET) is a predominant imaging diagnostic tool for the treatment of cancer and of other conditions, due to its unique ability to describe metabolic activity, which is difficult to recover from other imaging modalities. It is now gaining particular attention for early assessment of therapeutic response for improved individualized treatment planning in oncology. In this view, the PET-specific set of guidelines, PERCIST, proposed by Wahl *et al* in 2009 [2], reinforces this emerging importance of the modality in assessing therapy effectiveness. In this context, standard practice consists in assaying the proportional change in a summary of metabolic tracer uptake activity observed pre- and post-therapy; typical such statistical summaries include the mean and peak uptake measured within the volume of interest.

Although both quantitations consist in measuring an average within a predefined volume, an assessment of the corresponding standard error is seemingly systematically omitted in practice. In this work we model pre- and post-therapy data with Gamma distributions that feature the pseudo-Poisson characteristic. We calibrated this model against clinical sarcoma datasets from the University of Washington, and we also carried out a suitability analysis using standard calibration phantom data provided by the American College of Radiology Imaging Network. Under this framework we first obtain that the maximum likelihood estimator for global change in uptake distribution actually is the percentage change in mean uptake, a solution that aligns with standard medical and clinical practice (relying on SUV_{mean} quantitation). We explore the feasibility of pairing the pre- and post-therapy uptake information (in the sense of image co-registration) in order to adequately evaluate the standard error associated with this quantitation. We illustrate both analytically and numerically how pairing yields a benefit in the assessment of the standard error. This has obvious implications in

Copyright © 2016 John Wiley & Sons, Ltd.

terms of the performance of statistical tests and confidence estimators derived from this framework, over an unpaired approach. Whether expert-guided or automatized, scan co-registration will induce a form of misalignment due to complex nonlinear deformations (body position and motion, changes in organ shapes, tumor shape, etc.). Since pairing requires data alignment, and this may in turn introduce errors, the impact of data mis-alignment on the theoretical benefit of pairing was evaluated in concept. It was found that mis-alignment should not necessarily prevent the use of paired observations, and this was also verified with clinical data.

We applied this statistical framework to non-invasive, early therapeutic response assessment in patients with sarcoma, a rare form of cancer, from baseline and follow-up FDG-PET scans. Paired analysis of the standard error associated with the percentage change in mean PET probe uptake incorporates the relationship between pre- and post-therapy PET imaging information at a voxel level. The advantages of a paired analysis in deriving a measure of accuracy of therapeutic assessment, compared to a traditional unpaired procedure, were highlighted. Our experiments also confirmed the clinical, prognostic utility of the proposed evaluation of treatment effect, for both patient survival and disease progression, in a multivariate prognostic model. In this setup the proposed quantitation was found to be a significant addition to the model and a significant prognostic indicator for both survival and disease progression.

We found that clinical PET imaging data tend to be well described with the Gamma model used here, both for static and dynamic PET datasets. However other statistical frameworks may be used that implement the assumption of pseudo-Poisson uptake characteristics. A common alternative is to model treatment effects with a pseudo-Poisson Generalized Linear Model approach, where standard implementation relies on a Newton-Raphson procedure, which we considered in previous work. Under that approach we derived closed-form expressions for the corresponding treatment effect estimator and for its associated standard error, which also readily provided an optimization-free solution to the estimation problem along with an evaluation of associated variability. Remarkably, the analytical form of the GLM-based estimator is also an evaluation of the percentage change in mean value of the observations, as is the case when modelling response with the Gamma model proposed in this paper. Approximate expressions for estimation variance were also derived analytically but proved to be less efficient on real PET data (in terms of separation from unpaired quantitation) than those derived under the present Gamma model.

We suggested in Section 1.3 that the paired standard error quantitation could be explored in the context of measuring intratumoral (macro level) heterogeneity, since due to pairing it evaluates the magnitude of variations in intensity of matched voxels within a given volume of interest. This could be either as a univariate approach or as part of a multivariate formulation of heterogeneity. Tumor heterogeneity is a complex process that requires taking into consideration a variety of physiological, biological and metabolic factors. How one should define intratumoral heterogeneity from non-invasive imaging data is not clear. Typical texture-based (co-adjacence,) or distribution-based quantitations (entropy, uniformity, kurtosis, co-occurrence moments,) could yield as many different assessments for the process, and this is likely to be disease-dependent also.

In some settings, the proposed approximation for the standard error could be further calibrated to account for spatial correlation among image voxels. In the application to PET data considered here, the standard errors for $\hat{\beta}$ and $\hat{\gamma}$ could be normalized to correct for spatial correlation induced by the scanner acquisition process. Typically, this value would be expected to be approximately $\kappa = 0.9$ ([43, 52]) – this value may be provided by the manufacturer – and the appropriate correction factor should be $(1+\kappa^2)/(1-\kappa^2)$. The present analysis does not implement this. It is certainly worthy of further investigation as incorporating this parameter could yield to improving the power of tests based on t_{QP}. Future work will explore this.

As a possible extension of the present contribution, analytical treatment of a model extended to multivariate responses (described by multiple additive components in a loglinear representation) would be of interest to a variety of applications.

Paired data (treatment versus control, pre- versus post-, etc.) are commonly encountered in a wide variety of medical and non-medical contexts. While there are many situations where measurements (perhaps after transformation) might follow the classical assumptions of the paired t-test, there are certainly situations where this might not be well justified. For example, count data (and scaled counts) are fundamental to the various proteomic and genomic investigations that are so widely used in medical research, and the development of novel methodologies for analysis of paired studies in such settings is of on-going interest in the field–see, for example, Pham et al [53]. There is potential to adapt our approach for evaluation in this context. Future work will explore this.

Funding

This work was supported in part by the National Institutes of Health (USA) under CA-65537 and CA-42045 and by Science Foundation Ireland under 11/PI/1027. *Conflict of Interest*: None.

Copyright © 2016 John Wiley & Sons, Ltd.

References

1. Eary J, Krohn K. Positron emission tomography: imaging tumor response. *European Journal of Nuclear Medicine and Molecular Imaging* 2000; **27**(12):1737–1739.
2. 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**(5):122S–150S.
3. O'Sullivan F, Muzi M, Mankoff D, Eary J, Spence A, Krohn K. Voxel-level mapping of tracer kinetics in PET studies: a statistical approach emphasising tissue life tables. *Ann. Appl. Stat.* 2014; **8**(2):1065–1094.
4. de Geus-Oei L, Oyen W. Predictive and prognostic value of FDG-PET. *Cancer Imaging* 2008; **110**(8):70–80.
5. Doot R, McDonald E, Mankoff D. Role of PET quantitation in the monitoring of cancer response to treatment: review of approaches and human clinical trials. *Clinical and Translational Imaging* 2014; **2**(4):295–303.
6. Eary J, O'Sullivan F, O'Sullivan J, Conrad E. Spatial heterogeneity in sarcoma 18F-FDG uptake as a predictor of patient outcome. *J Nucl Med* 2008; **49**(12):1973–1979.
7. Kinahan P, Fletcher J. PET/CT standardized uptake values (SUVs) in clinical practice and assessing response to therapy. *Seminars in ultrasound, CT, and MR* 2010; **31**(6):496–505.
8. Larson S, Erdi Y, Akhurst T, Mazumdar M, Macapinlac H, Finn R, Casilla C, Fazzari M, Srivastava N, Yeung H, *et al.*. Tumor treatment response based on visual and quantitative changes in global tumor glycolysis using PET-FDG imaging. the visual response score and the change in total lesion glycolysis. *Clin Positron Imaging* 1999; **2**(3):159–171.
9. Schutze S, Rubin B, Vernon C, Hawkins D, Bruckner J, Conrad E, Eary J. Use of positron emission tomography in localized extremity soft tissue sarcoma treated with neoadjuvant chemotherapy. *Cancer* 2005; **103**(2):339–348.
10. Tap WD, Dry SM, Elashoff D, Chow K, Evilevitch V, Eckardt JJ, Phelps ME, Weber WA, Eilber FC. FDG-PET/CT imaging predicts histopathologic treatment responses after the initial cycle of neoadjuvant chemotherapy in high-grade soft-tissue sarcomas. *Clin Cancer Res* 2009; **15**(8):2856–2863.
11. Denecke T, Hundsdörfer P, Misch D, Steffen IG, Schnberger S, Furth C, Plotkin M, Ruf J, Hautzel H, Stver B, *et al.*. Assessment of histological response of paediatric bone sarcomas using FDG-PET in comparison to morphological volume measurement and standardized MRI parameters. *Eur J Nucl Med Mol Imaging* 2010; **37**(10):1842–1853.
12. Quak E, van de Luijtgaarden A, de Geus-Oei LF, van der Graaf W, Oyen W. Clinical applications of positron emission tomography in sarcoma management. *Expert Review of Anticancer Therapy* 2011; **11**(2):195–204.
13. Vanderhoek M, Perlman S, Jeraj R. Impact of the definition of peak standardized uptake value on quantification of treatment response. *J Nucl Med* 2012; **53**(1):4–11.
14. O'Sullivan F, Wolsztynski E, O'Sullivan J, Richards T, Conrad E, Eary J. A statistical modelling approach to the analysis of spatial patterns of FDG-PET uptake in human sarcoma. *IEEE Trans. Med. Imag.* 2011; **30**(12):2059–2071.
15. Carlino M, Saunders C, Haydu L, Menzies A, Curtis CJM, Lebowitz P, Kefford R, Long G. (18)F-labelled fluorodeoxyglucose-positron emission tomography (FDG-PET) heterogeneity of response is prognostic in dabrafenib treated BRAF mutant metastatic melanoma. *Eur J Cancer* 2013; **49**(2):395–402.
16. Soussan M, Orlhac F, Boubaya M, Zelek L, Ziol M, Eder V, Buvat I. Relationship between tumor heterogeneity measured on FDG-PET/CT and pathological prognostic factors in invasive breast cancer. *PLoS ONE* 2014; **9**(4):881–901.
17. Hatt M, le Rest CC, van Baardwijk A, Lambin P, Pradier O, Visvikis D. Impact of tumor size and tracer uptake heterogeneity in 18F-FDG PET and CT Non Small Cell Lung Cancer tumor delineation. *J. Nucl. Med.* 2011; **52**(11):1690–7.
18. Davnall F, Yip C, Ljungqvist G, Selmi M, Ng F, Sanghera B, Ganeshan B, Miles K, Cook G, Goh V. Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice? *Insights Imaging* 2012; **3**(6):573–589.
19. Tixier F, Hatt M, Valla C, Fleyry V, Lamour C, Ezzouhri S, Ingrand P, Perdrisot R, Visvikis D, Rest CL. Visual versus quantitative assessment of intratumor 18f-fdg pet uptake heterogeneity: Prognostic value in non-small cell lung cancer. *J Nucl Med* 2014; **55**(8):1235–1241.
20. Hatt M, Majdoub M, Vallieres M, Tixier F, Rest CL, Groheux D, Hindie E, Martineau A, Pradier O, Hustinx R, *et al.*. 18f-FDG PET uptake characterization through texture analysis: Investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort. *J Nucl Med* 2015; **56**(1):38–44.
21. Rahmim A, Coughlin J, Gonzalez M, Endres C, Zhou Y, Wong D, Wahl R, Sossi V, Pomper M. Novel parametric PET image quantification using texture and shape analysis. *Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2012 IEEE*, 2012; 2227–2230.
22. Byun B, Kim S, Lim S, Lim I, Kong C, Song W, Cho W, Jeon D, Lee S, Koh J, *et al.*. Prediction of response to neoadjuvant chemotherapy in osteosarcoma using dual-phase (18)F-FDG PET/CT. *Eur Radiol* 2015; **25**(7):2015–2024.
23. Hussien A, Furth C, Schönberger S, Hundsdoerfer P, Steffen I, Amthauer H, Müller H, Hautzel H. FDG-PET response prediction in pediatric Hodgkins lymphoma: Impact of metabolically defined tumor volumes and individualized SUV measurements on the positive predictive value. *Cancers (Basel)* 2015; **7**(1):287–304.
24. Rice JA. *Mathematical Statistics and Data Analysis*. Duxbury Press, 2001.
25. Yang L, Tsiatis AA. Efficiency study of estimators for a treatment effect in a pretest-posttest trial. *American Statistician* 2001; **55**:314–321.
26. Gibbons JD, Chakraborti S. *Nonparametric Statistical Inference, Fourth Edition, Revised and Expanded*. Marcel Dekker, New-York, 2003.
27. Neemuchwala H, Hero A, Carson P. Image registration using entropic graph-matching criteria. *Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers, Asimolar, California, 5 pages*, 2002.
28. Neemuchwala H, Hero A, Zabuawala S, Carson P. Image registration methods in high dimensional space. *International Journal of Imaging Systems and Technology,* 2007; **16**(5):130–145.
29. Calais J, Dubray B, Nkhali L, Thureau S, Lemarignier C, Modzelewski R, Gardin I, Fiore FD, Michel P, Vera P. High FDG uptake areas on pre-radiotherapy PET/CT identify preferential sites of local relapse after chemoradiotherapy for localy advanced oesophageal cancer. *Eur J Nucl Med Mol Imaging* 2015; **42**(6):858–867.
30. Guo M, Yap J, van den Abbeele A, Lin N, Schwartzman A. Voxelwise single-subject analysis of imaging metabolic response to therapy in neuro-oncology. *Stat.* 2014; **3**:172–186.
31. Carson R, Yan Y, Daube-Witherspoon M, Freedman N, Bacharach S, Herscovitch P. An approximation formula for the variance of PET region-of-interest values. *IEEE Trans Med Imaging* 1993; **12**(2):240–250.
32. Maitra R, O'Sullivan F. Variability assessment in PET and related generalized deconvolution models. *J Amer Stat Assoc* 1998; **93**(444):1340–1355.
33. Bonetti M, Gelber RD. Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics* 2004; **5**(3):465–481.

Copyright © 2016 John Wiley & Sons, Ltd.

34. de Valpine P, Bitter HM, Brown MPS, Heller J. A simulationapproximation approach to sample size planning for high-dimensional classification studies. *Biostatistics* 2009; **10**(3):424–435.
35. Hougaard P, Lee MLT, Whitmore GA. Analysis of overdispersed count data by mixtures of poisson variables and poisson processes. *Biometrics* 1997; **53**(4):1225–1238.
36. Leon S, Tsiatis AA, Davidian M. Semiparametric estimation of treatment effect in a pretest-posttest study. *Biometrics* 2003; **59**(4):1046–1055.
37. Huang CY, Qin J, Follmann DA. Empirical likelihood-based estimation of the treatment effect in a pretestposttest study. *J Am Stat Assoc* 2008; **103**(483):1270–1280.
38. Lee JH, Han G, Fulp WJ, Giuliano AR. Analysis of overdispersed count data: application to the human papillomavirus infection in men (him) study. *Epidemiology and Infection* 6 2012; **140**:1087–1094.
39. McCullagh P, Nelder J. *Generalized Linear Models, 2nd ed.* Boca Raton: Chapman and Hall/CRC, 1989.
40. Eisenhauer E, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *European Journal of Cancer* 2009; **45**(2):228–247.
41. Lopiano KK, Young LJ, Gotway CA. Estimated generalized least squares in spatially misaligned regression models with Berkson error. *Biostatistics* 2013; **14**(4):737–751, doi:10.1093/biostatistics/kxt011.
42. Jan S, Benoit D, Becheva E, Carlier T, Cassol F, Descourt P, Frisson T, Grevillot L, Guigues L, Maigne L, *et al.*. GATE V6: a major enhancement of the GATE simulation platform enabling modelling of CT and radiotherapy. *Phys. Med. Biol.* 2011; **56**:881–901.
43. O'Sullivan F. Locally constrained mixture representation of dynamic imaging data from PET and MR studies. *Biostatistics* April 2006; **7**(2):318–338.
44. Grosu AL, Piert M, Molls M. Experience of PET for target localisation in radiation oncology. *British Journal of Radiology, Supplement* 2005; **28**(1):18–32.
45. Loening A, Gambhir S. AMIDE: A free software tool for multimodality medical image analysis. *Molecular Imaging* 2003; **2**(3):131–137.
46. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2010. URL http://www.R-project.org/, ISBN 3-900051-07-0.
47. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011; **12**(1):77.
48. Therneau T, Grambsch P. *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New-York, 2000.
49. Burnham K, Anderson D. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd ed.* Springer-Verlag, New-York, 2002.
50. Harrongton D, Fleming T. A class of rank test procedures for censored survival data. *Biometrika* 1982; **69**:553–566.
51. Hatt M, le Rest CC, Albarghach NM, Pradier O, Visvikis D. PET functional volume delineation: a robustness and repeatability study. *European Journal of Nuclear Medicine and Molecular Imaging* 2011; **38**(4):663–672.
52. O'Sullivan F, Roy S, O'Sullivan J, Vernon C, Eary J. Incorporation of tumour shape into an assessment of spatial heterogeneity for human sarcomas imaged with FDG-PET. *Biostatistics* 2005; **6**(2):293–301.
53. Pham T, Jimenez C. An accurate paired sample test for count data. *Bioinformatics* 2012; **28**(18):i596–i602.

A. Derivations of closed-form solutions

Hereafter we provide details on the derivations of quantities of interest under the assumption of Gamma-distributed uptake information. In particular we show that the solution to the therapeutic response quantitation problem in this model actually is the percentage-change in mean uptake. We also provide finite-sample analytic and approximating expressions for paired and unpaired standard error terms.

Assuming that pre- and post-therapy uptake values (resp. y_{0i} and y_{1i}) have the pseudo-Poisson property (1) and posing that $\lambda_{1i} = \gamma\lambda_{0i}$, we model the two-sample comparison problem for $i = 1, \dots, N$ with

$$y_{0i} \sim \mathcal{G}\left(\frac{\lambda_i}{\phi}, \phi\right), \qquad y_{1i} \sim \mathcal{G}\left(\frac{\gamma\lambda_i}{\phi}, \phi\right) \tag{22}$$

where the $\{\lambda_i\}_{i=1}^N$ describe fixed individual effects and γ represents the global change in uptake following therapy. Here we consider the pdf for the Gamma distribution $\mathcal{G}(\alpha, S)$ in terms of its shape α and scale S parameters (both strictly positive) as

$$f(y; \alpha, S) = \frac{1}{\Gamma(\alpha)S^\alpha} y^{\alpha-1} \exp\left(-\frac{y}{S}\right) \tag{23}$$

($\Gamma(.)$ denoting the Gamma function). Under this parametrization, we have

$$E(y) = \alpha S, \qquad \mathrm{Var}(y) = \alpha S^2 \tag{24}$$

which, in combination with (22), satisfies (1). As discussed in Section 2 we assume the pre- and post-therapy uptake information to be uncorrelated. Under this model, the log-likelihood expression for the two-sample problem is given by

$$l(Y; \gamma, \lambda, \phi) = -\sum_{i=1}^{N}\left[\frac{Y_{0i}+Y_{1i}}{\phi} - \frac{\lambda_i}{\phi}\log\left(\frac{Y_{0i}}{\phi}\right) - \gamma\frac{\lambda_i}{\phi}\log\left(\frac{Y_{1i}}{\phi}\right) + \log\left(\Gamma\left(\frac{\lambda_i}{\phi}\right)\right) + \log\left(\Gamma\left(\frac{\gamma\lambda_i}{\phi}\right)\right)\right] \tag{25}$$

Copyright © 2016 John Wiley & Sons, Ltd.

A.1. Paired assessment of treatment effect

First we derive maximum likelihood solutions for λ and γ in the joint estimation problem stated by (23). The score equation for a particular λ_i satisfies

$$\log\left(\frac{Y_{0i}}{\phi}\right)+\gamma\log\left(\frac{Y_{1i}}{\phi}\right)=\psi\left(\frac{\hat{\lambda}_i}{\phi}\right)+\gamma\psi\left(\frac{\gamma\hat{\lambda}_i}{\phi}\right) \tag{26}$$

where $\psi(.)$ denotes the digamma function. First-order expansions of the $\log(.)$ and $\psi(.)$ terms give the following useful approximations:

$$\begin{aligned}
\log(Y_{0i}) &\approx \log(\lambda_i)+\frac{Y_{0i}-\lambda_i}{\lambda_i}\\
\log(Y_{1i}) &\approx \log(\gamma\lambda_i)+\frac{Y_{1i}-\gamma\lambda_i}{\gamma\lambda_i}\\
\psi\left(\hat{\lambda}_i\phi\right) &\approx \psi\left(\lambda_i\phi\right)+\frac{1}{\phi}\psi_1\left(\frac{\lambda_i}{\phi}\right)(\hat{\lambda}_i-\lambda_i)\\
\psi\left(\gamma\hat{\lambda}_i\phi\right) &\approx \psi\left(\gamma\lambda_i\phi\right)+\frac{\gamma}{\phi}\psi_1\left(\frac{\gamma\lambda_i}{\phi}\right)(\hat{\lambda}_i-\lambda_i)
\end{aligned} \tag{27}$$

Using the above in (26), and furthermore that for x large, $\psi_1(x)\approx 1/x$ and $\psi(x)\approx\log(x)$ (in our context it is reasonable to assume that λ_i/ϕ be large enough, cf. Appendix C), we obtain

$$\frac{Y_{0i}-\lambda_i}{\lambda_i}+\frac{Y_{1i}-\gamma\lambda_i}{\lambda_i}\approx(1+\gamma)\frac{\hat{\lambda}_i-\lambda_i}{\lambda_i}$$

and

$$\hat{\lambda}_i\approx\frac{Y_{0i}+Y_{1i}}{1+\gamma} \tag{28}$$

The score equation for γ can be analyzed in a similar fashion, yielding

$$\sum_{i=1}^{N}\left[\frac{\lambda_i}{\phi}\left(\log\left(\frac{\gamma\lambda_i}{\phi}\right)+\frac{Y_{1i}-\gamma\lambda_i}{\gamma\lambda_i}\right)\right]=\sum_{i=1}^{N}\left[\frac{\lambda_i}{\phi}\left(\psi\left(\frac{\gamma\lambda_i}{\phi}\right)+\frac{\lambda_i}{\phi}\psi_1\left(\frac{\gamma\lambda_i}{\phi}\right)(\hat{\gamma}-\gamma)\right)\right] \tag{29}$$

from which it follows that $\hat{\gamma}\approx\frac{\sum_{i=1}^{N}Y_{1i}}{\sum_{i=1}^{N}\lambda_i}$. Combining this result with (28) we obtain the approximate MLE

$$\hat{\gamma}=\frac{\bar{Y}_1}{\bar{Y}_0} \tag{30}$$

where $\bar{Y}_j$ denotes the sample mean of the observations $\{Y_{ji}\}_{i=1}^{N}, j=0,1$. It is worth noticing that $\hat{\gamma}-1$ quantifies the percentage-change in uptake means, i.e. that the MLE for γ under model (22) coincides with the widely used therapeutic assessment $\%\mathrm{SUV}_{\mathrm{mean}}$.

A.2. Estimation of ϕ

For the estimation of ϕ it is convenient to rewrite the log-likelihood in terms of $\theta=1/\phi$, which yields the following score equation for θ:

$$\sum_{i=1}^{N}\left[(Y_{0i}+Y_{1i})-\lambda_i-\gamma\lambda_i-\lambda_i\log(\theta Y_{0i})-\gamma\lambda_i\log(\theta Y_{1i})+\lambda_i\psi(\theta\lambda_i)+\gamma\lambda_i\psi(\gamma\theta\lambda_i)\right]=0 \tag{31}$$

For $x>1$, $\psi(x)\approx\log(x)-\frac{1}{2x}$, so $\lambda_i\psi(\theta\lambda_i)+\gamma\lambda_i\psi(\gamma\theta\lambda_i)\approx\lambda_i\log(\theta\lambda_i)+\gamma\lambda_i\log(\gamma\theta\lambda_i)-\frac{1}{\theta}$. This gives the approximation

$$\frac{1}{\theta}\approx\frac{1}{N}\sum_{i=1}^{N}\left[(Y_{0i}+Y_{1i})-\lambda_i(1+\gamma)-\lambda_i\log\left(\frac{Y_{0i}}{\lambda_i}\right)-\gamma\lambda_i\log\left(\frac{Y_{1i}}{\gamma\lambda_i}\right)\right]$$

Copyright © 2016 John Wiley & Sons, Ltd.

Now, using the following second-order Taylor expansions in the above:

$$\begin{aligned}\log(Y_{0i}) - \log(\lambda_i) &\approx \frac{Y_{0i}-\lambda_i}{\lambda_i} - \frac{(Y_{0i}-\lambda_i)^2}{2\lambda_i^2} \\ \log(Y_{1i}) - \log(\gamma\lambda_i) &\approx \frac{Y_{1i}-\gamma\lambda_i}{\gamma\lambda_i} - \frac{(Y_{1i}-\gamma\lambda_i)^2}{2\gamma^2\lambda_i^2}\end{aligned} \tag{32}$$

we obtain the following approximate MLE for ϕ:

$$\frac{1}{\hat{\theta}} = \hat{\phi} = \frac{1}{2N}\sum_{i=1}^{N}\left(\frac{(Y_{0i}-\lambda_i)^2}{\lambda_i} + \frac{(Y_{1i}-\gamma\lambda_i)^2}{\gamma\lambda_i}\right) \tag{33}$$

Using (28) and (30) in the above, and adjusting the biased maximum likelihood estimator obtained initially for N degrees of freedom, we obtain

$$\tilde{\phi} = \frac{1}{N\hat{\gamma}}\sum_{i=1}^{N}\frac{(\hat{\gamma}Y_{0i}-Y_{1i})^2}{Y_{0i}+Y_{1i}} \tag{34}$$

At this point an alternative to (34) may also be derived. Considering the terms $U_i = \gamma Y_{0i} - Y_{1i}$ which arise form the MLE in (34), and assuming more loosely that $\bar{\lambda}_1 = \gamma\bar{\lambda}_0$ but not necessarily that $\lambda_{1i} = \gamma\lambda_{0i}$, we have

$$\mathrm{Var}(U_i) = \mathrm{Var}(\gamma Y_{0i} - Y_{1i}) = \gamma^2\mathrm{Var}(Y_{0i}) + \mathrm{Var}(Y_{1i}) = \phi(\gamma^2\lambda_{0i} + \lambda_{1i})$$

Since $E(U_i) = \gamma\bar{\lambda}_0 - \bar{\lambda}_1 = 0$, we also have that $\mathrm{Var}(U_i) = E(U_i^2) = E((\gamma Y_{0i} - Y_{1i})^2)$, which combined to the above expression yields a direct estimator for ϕ as

$$\hat{\phi} = \frac{E\left((\gamma Y_{0i} - Y_{1i})^2\right)}{\gamma^2\lambda_{0i} + \lambda_{1i}}$$

This analytic expression may be approximated with

$$\hat{\phi} = \frac{1}{N}\sum_{i=1}^{N}\frac{(\hat{\gamma}Y_{0i}-Y_{1i})^2}{\hat{\gamma}^2 Y_{0i} + Y_{1i}} \tag{35}$$

A.3. The mean and variance of $\hat{\beta}$

We can now derive closed-form expressions for estimated variances associated with the maximum likelihood estimator (30) for γ. At this point it is convenient to consider the log-transform of this estimator, i.e.

$$\hat{\beta} = \log(\hat{\gamma}) = \log(\bar{Y}_1) - \log(\bar{Y}_0) \tag{36}$$

with

$$E(\hat{\beta}) \approx \log(\bar{\lambda}_1) - \log(\bar{\lambda}_0) = \beta$$

By the Delta method,

$$\mathrm{Var}(\hat{\beta}) = \sigma^2_{\hat{\beta}} = \mathrm{Var}(\log(\bar{Y}_0)) + \mathrm{Var}(\log(\bar{Y}_1)) \approx \frac{\mathrm{Var}(\bar{Y}_0)}{\bar{\lambda}_0^2} + \frac{\mathrm{Var}(\bar{Y}_1)}{\bar{\lambda}_1^2} = \frac{\phi_0}{N\bar{\lambda}_0} + \frac{\phi_1}{N\bar{\lambda}_1}$$

Assuming as discussed in Section 2 that $\phi_0 \approx \phi_1 \approx \phi$, and furthermore that $\bar{\lambda}_1 = \gamma\bar{\lambda}_0$, we obtain the following approximate analytic expression for the paired standard error of $\hat{\beta}$:

$$\sigma^2_{\hat{\beta}} = \frac{\phi}{N\bar{\lambda}}\left(1 + \frac{1}{\gamma}\right) \tag{37}$$

This expression may be estimated using uptake samples and e.g. (35) by

$$\hat{\sigma}^2_{\hat{\beta}} = \frac{\hat{\phi}}{N}\left(\frac{1}{\bar{Y}_0} + \frac{1}{\bar{Y}_1}\right) \tag{38}$$

Copyright © 2016 John Wiley & Sons, Ltd.

A.4. Naive estimation of variability

If $\mathrm{Var}(\bar{Y}_j)$ is estimated by the usual sample variance we get

$$\sigma^2_{\hat{\beta},U} \approx \left(\frac{E(SD_N(Y_0)^2)}{N\bar{\lambda}_0^2} + \frac{E(SD_N(Y_1)^2)}{N\bar{\lambda}_1^2} \right) \tag{39}$$

subscript U denoting the usual approach, where

$$\begin{aligned}
SD^2_N(Y_j) &= \frac{1}{N}\sum_{i=1}^{N}(Y_{ji}-\bar{Y}_j)^2 = \frac{1}{N}\sum_{i=1}^{N}Y_{ji}^2 - \bar{Y}_j^2 \\
E\left(SD^2_N(Y_j)\right) &= \frac{1}{N}\sum_{i=1}^{N}\left(\mathrm{Var}(Y_{ji}) + E(Y_{ji})^2\right) - \mathrm{Var}(\bar{Y}_j) - E(\bar{Y}_j)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\lambda_{ji}\phi + \lambda_{ji}^2\right) - \frac{1}{N^2}\sum_{i=1}^{N}\lambda_{ji}\phi - \left(\frac{1}{N}\sum_{i=1}^{N}\lambda_{ji}\right)^2 \\
&= \bar{\lambda}_j\left(\phi\left(1-\frac{1}{N}\right)\right) + \frac{1}{N}\sum_{i=1}^{N}\left(\lambda_{ji}-\bar{\lambda}_j\right)^2 \\
&\approx \bar{\lambda}_j\phi + \xi(\lambda_j)
\end{aligned}$$

where

$$\xi(\lambda_j) = \frac{1}{N}\sum_{i=1}^{N}(\lambda_{ji}-\bar{\lambda}_j)^2, \qquad \bar{\lambda}_j = \frac{1}{N}\sum_{i=1}^{N}\lambda_{ji}$$

If $\lambda_{1i} = \gamma\lambda_{0i}$ then

$$E\left(SD^2_N(\lambda_1)\right) \approx \gamma\bar{\lambda}_0\phi + \gamma^2\xi(\lambda_0)$$

so

$$\sigma^2_{\hat{\beta},U} \approx \frac{\bar{\lambda}_0\phi + \xi(\lambda_0)}{N\bar{\lambda}_0^2} + \frac{\gamma\bar{\lambda}_0\phi + \gamma^2\xi(\lambda_0)}{N\bar{\lambda}_0^2\gamma^2} = \frac{1}{N\bar{\lambda}_0}\left(\frac{2\xi(\lambda_0)}{\bar{\lambda}_0} + \left(1+\frac{1}{\gamma}\right)\phi\right)$$

B. Numerical discussion of hypotheses

B.1. Case of varying scale parameter ϕ across treatment time points

In Section 2.1 we discussed our modelling of treatment effect using a pseudo-Poisson Gamma model, and proposed formulation (3) by which pre- and post-therapy data had a common scale parameter ϕ. This assumption was based on considerations on the routine practice of therapeutic assessment using PET imaging. Here we consider a numerical analysis to evaluate the estimation of model scale ϕ in a more elaborate situation in which ϕ_0 and ϕ_1 are function of dose τ in that $\phi = a/\tau$, where a is a scale factor and τ is dose per unit volume. (We do not discuss the eventuality of the use of different scanners throughout therapy.) This alternative model allows for the two model scales to differ as per (2), which stated for $i = 1, \ldots, N$ that

$$\begin{aligned}
y_{0i} \mid \lambda_i &\sim \mathcal{G}(\lambda_i/\phi_0, \phi_0) \\
y_{1i} \mid \lambda_i &\sim \mathcal{G}(\gamma\lambda_i/\phi_1, \phi_1)
\end{aligned}$$

With this model, normalisation of raw count data by injected dose τ yields dimensionless standardized uptake values (SUVs). Recall that the proposed estimators of therapeutic effect and associated standard error (constructions (8), (10) and (11)) are scale-invariant, i.e. the value of the injected dose τ has no bearing on the effectiveness of the estimation. Only $\hat{\phi}$ is affected by a change in τ, in the same way as Y would be. We implemented this model in simulations using $\phi_0 = a/\tau_0$ and $\phi_1 = a/\tau_1$, where τ_0 and τ_1 are known, a, γ and λ are unknown. We present here a specific example where we set $\tau_0 = 2\tau_1$, and with $\gamma = .25, a = 0.5$ and λ ranging quadratically between 0 and 1. A range of sample sizes N and doses were explored, with $N \in \{500, 1000, 2000, 4000, 10000\}$ and $\tau \in \{10^1, 10^2, 10^3, 10^4, 10^5, 10^6\}$. Note that the

Copyright © 2016 John Wiley & Sons, Ltd.

value $\tau = 10^4$ would be typical of FDG-PET data (with tissue activity measured at around 0.1 kBq/ml), lower values of τ yielding less defined uptake information. To estimate a we used $\hat{a} = \tau\hat{\phi}$, where $\hat{\phi}$ is derived using (5) in an adaptation of (7) to the case of different τ values before and after treatment, to obtain

$$\hat{a} = \frac{1}{N}\sum_{i=1}^{N} \frac{(y_{0i}\hat{\gamma} - y_{1i})^2}{\hat{\lambda}_i\hat{\gamma}(\hat{\gamma}/\tau_0 + 1/\tau_1)}.$$

For each (N, τ) experimental setting we carried out 1,000 Monte Carlo repetitions of the experiment, the results of which are presented in Figure 9. These results indicate that estimation of a is consistent with increasing dose (τ) and ROI size (N). For fixed N, estimation accuracy decreases in a quadratic fashion with τ. Estimate variability also diminishes with $1/N$. This behaviour replicates the performance of the usual estimator of variability in the standard two-sample paired test. There the usual estimator is the sample variance of the observed differences and its MSE behaves as σ^4/N, diminishing with σ^2 decreasing and N increasing. Note that in the Gamma model $1/\tau$ plays the role of σ^2.

B.2. Comparison of paired estimators for scale ϕ with respect to size and dose

We now evaluate the two estimators of ϕ in a paired model given by (6) and (7) in Section 2.2. In that section we suggested that (7) was more appropriate in a "high count" scenario (i.e. for higher doses τ). To do this we considered the Gamma model used in Appendix B.1, with $\phi_0 = \phi_1 = a/\tau$, τ known, γ, λ_i and a unknown. A Monte Carlo simulation was carried out for varying sample sizes (using N respectively in $\{500, 1000, 2000, 4000\}$), each with varying $\tau \in \{10^1, 10^2, 10^3, 10^4, 10^5, 10^6\}$, and for 1,000 repetitions of the experiment for each value (N, τ) of the experimental setting. The results are shown in Figure 10.

In this figure the top row of boxplots indicates distributions of a-estimates obtained for the various experimental settings, which demonstrate that a is recovered properly in this framework with $\phi_0 = \phi_1$. The bottom row of boxplots illustrates how with increasing τ, the estimates (6) and (7) coincide, regardless of sample size N. The figure also demonstrates how estimation of ϕ (standardised for dose) improves with increasing dose and ROI size.

C. Data simulation

A Gamma model was fitted to validation phantom data from the American College of Radiology Imaging Network (ACRIN) Core Laboratory. Views of this dataset are provided in Figure 11. The ACRIN PET and PET/CT Scanner Qualification is available at

http://www.acrin.org/CORELABS/PETCORELABORATORY/PETQUALIFICATION/tabid/485/Default.aspx

(last accessed December 14, 2015). These results motivate the simulation model used in Section 3, describing uptake y at voxel x as

$$y(x_i) \sim \mathcal{G}(\tau\lambda_i/\phi, \phi) \tag{40}$$

with $E(y) = \tau\lambda$, $\mathrm{Var}(y) = \tau\lambda\phi$, and where τ acts as a dose-normalizing constant and ϕ is a fixed value (the ACRIN data typically suggests $\phi = 1.2$). The count-rate dependent value of τ may vary with axial slice and time in dynamic PET data. In this model voxel rates λ_i are counts per volume, and may be understood as a conversion of the tracer concentration in tissue, i.e. $\lambda_i = C_{T,i}(t)K_t$ that we evaluate e.g. at the final time point t_{max} if using the summed PET image. Here $C_{T,i}(t)$ denotes the concentration at voxel i, in kBq per cc (typically 10^3 counts per second, i.e. within the rang 0.01 tp 1 kBq per cc) and conversion factor $K_t = \Delta_{V,i}\Delta_{F,t}\delta(t)$ is used to to convert tissue activity per unit volume into counts information, where $\Delta_{V,i} \approx (0.4)^3$ denotes volume of voxel i, $\Delta_{F,t}$ denotes duration in seconds of frame t and $\delta(t) = \exp(-\log(2)\frac{t}{hlife})$, dimensionless, is the decay (using e.g. $hlife = 109.7$ mns for FDG), i.e. the fraction of remaining PET tracer radioactivity at time t. The proposed model was fitted to both dynamic and static (i.e. summed) phantom data with satisfaction. For static imaging data, $C_{T,i}(t)$ and K_t become constant with respect to time and the only variation is spatial, in local tracer concentration. In this setting, the dose-adjusting factor τ controls the overall level of activity (the higher the dose, the greater the activity, and the smaller the estimation error for γ).

The Gamma model was chosen for its ability to model pseudo-Poisson characteristics as observed in PET data [31, 32], in particular with a proportional mean–variance relationship, and its asymptotic equivalence with normality as the uptake rate increases. The above model was fitted via maximum likelihood and Figure 12 illustrates the output of this analysis. Validation study of the Gamma model for simulated illustrative experiments, based on the ACRIN qualification data shown in Figure 11. Since imaged tracer activity is proportional to scanning duration, it is relevant to analyse model fitting with respect to this parameter. PET-data distributions were therefore obtained from four different imaging frames of varying duration (10, 30, 60 and 120 seconds) for a slice picked in the centre of the scanner. The histograms shown in Figure 12 indicate that the distribution of reconstructed PET values within the phantom slice tends to be more skewed at lower frame

Copyright © 2016 John Wiley & Sons, Ltd.

durations and more normal at larger ones, as expected. Gamma model (40) allows for this pattern, and its calibration must reflect this principle. For each frame duration the Gamma density model fitted to the data is indicated by the red line. Kolmogorov-Smirnov tests consistently led to failing to reject the null hypothesis of a difference between the data distribution and the Gamma model. All our findings in this analysis are compliant with the assumption that the above Gamma model adequately describes real PET data.

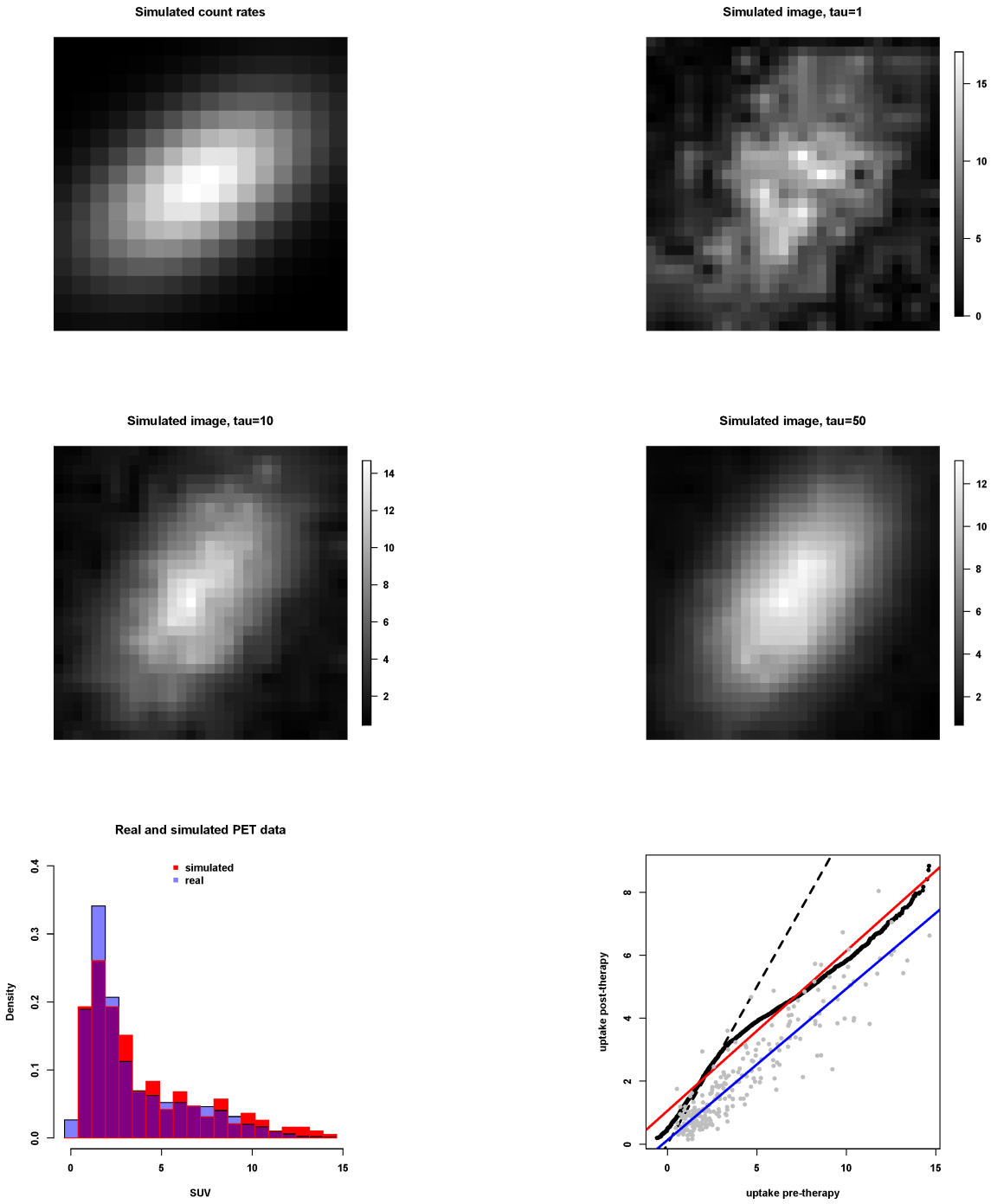


Figure 1. Top left: 2D mapping of the noiseless image response $\{\lambda_i\}_{i=1}^N$ used in (21). Top right: example of simulated data using $\{\lambda_i\}_{i=1}^N$ in (21) using $\gamma = 1, \tau = 1$ (interpolated to the half-pixel). Middle row: same simulated data image as top-right, but using $\tau = 10$ (left) or $\tau = 50$ (right) to illustrate how τ enhances image contrast. Bottom left: comparison of the histogram of the same simulated image with the true baseline PET dataset shown in Figure 2. Bottom right: illustration of relationship of true pre- vs post-therapy PET data (black dots, which form a curve) and simulated sets (gray dots). The dashed line indicates the unit ($y = x$) line.

Copyright © 2016 John Wiley & Sons, Ltd.

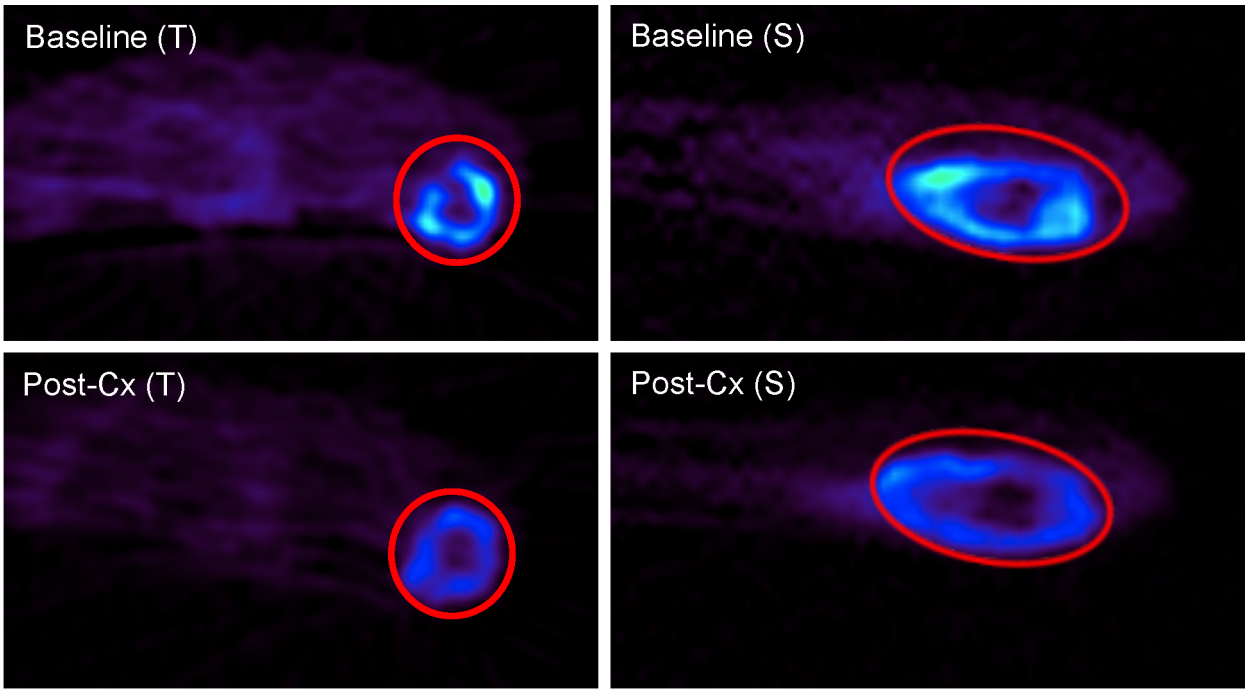


Figure 2. 17 y.o. male with grade 3 osteosarcoma with metastasis at baseline, in the humerus, imaged at baseline (top, $SUV_{max,0}$ = 14.6) and mid-chemotherapy (bottom, $SUV_{max,1}$ = 8.8). The views consist in sagittal (S) and transverse (T) "slices". The line corresponds to the ellipsoidal input ROI used for analysis.

Table 1. Characteristics of the dataset of 50 sarcoma studies, which consists of 27 female and 23 male patients, 20 bone and 30 soft-tissue sarcoma tumors, 28 grade 3 tumors and 22 tumours less than grade 3 (1 grade 1 and 21 grade 2 tumors), 38 tumors located in a limb and 12 located in the trunc, 17 deaths, 22 progressions.

	min	median	max
N	269	5802	50425
$SUV_{max,0}$	2.0	7.4	31.8
$SUV_{max,1}$	1.4	4.0	19.1
age	10	30	65

Table 2. Mean and standard error of estimates for extreme mis-registrations, using $\Delta = (3,3,1)$, $\Delta = (5,5,2)$ and $\Delta = (7,7,3)$, for the experiment of Figure 6, against the reference case $\Delta = (0,0,0)$.

	(0,0,0)	(3,3,1)	(5,5,2)	(7,7,3)
mean (s.e.) for $\hat{\gamma}$	1.0000 (0.0000)	1.0034 (0.0948)	0.9842 (0.1535)	0.9767 (0.2561)
mean (s.e.) for $\hat{\sigma}_{\hat{\gamma},P}$	0.0000 (0.0000)	0.0180 (0.0195)	0.0216 (0.0206)	0.0287 (0.0247)
mean (s.e.) for $\hat{\sigma}_{\hat{\gamma},U}$	0.0266 (0.0173)	0.0232 (0.0258)	0.0241 (0.0238)	0.0306 (0.0260)

Copyright © 2016 John Wiley & Sons, Ltd.

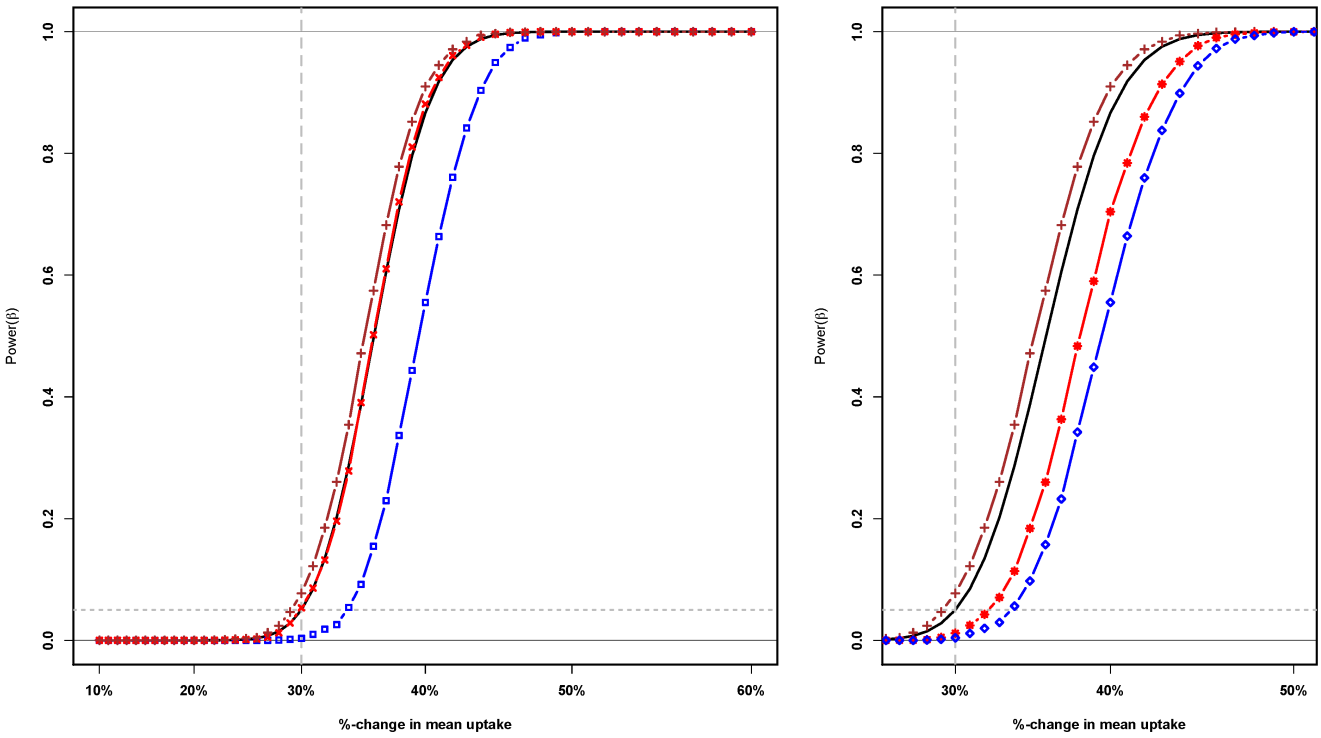


Figure 3. Illustrative experiment of power of one-sided tests t_{QP} (16), testing null hypothesis $H_0 : \beta > \beta_0 = \log(0.70)$ versus $H_A : \beta \leq \beta_0$. Similar tests may be used for example for the medical assessment of partial response of the disease following treatment (assuming e.g. that metabolic response corresponds to a decrease of 30% or more in mean activity). The output was obtained from $N_S = 5,000$ Monte Carlo simulations.
The solid black curve is the theoretic power of the paired test using the theoretic value $\sigma_{\hat{\beta}}$ (9) under the Normal distribution assumption. The power of the same paired test obtained using the Monte-Carlo average of analytic paired standard errors (10) calculated using the estimated response $\hat{\gamma}$ and the true values for $\phi_0 = \phi_1$ under ideal co-registration (red, '$\times$') is almost identical to it, which indicates adequacy of the proposed analytic expression for the standard error.
The *empirical* power for the test t_{QP} derived using (9) for perfectly paired data (brown, '+') is close to this theoretic power. The slight bias is due to the estimates for ϕ_0 (in this example $\phi_0 = \phi_1$). The empirical power for t_{QP} obtained from mis-registered data with $\sigma_\Delta = 4$ (green, 'o') also remains close to the theoretic power, which illustrates our proposal that perfect registration of the two imaging datasets is not critical in evaluating the standard error.
The blue curve marked with squares (left plot) gives the empirical power of the Monte-Carlo average naive test, using theoretic approaximation (12). The blue curve marked with diamonds (right plot) gives the empirical power of the Monte-Carlo average naive test, using the sample-based evaluation (11).
The grey horizontal line marks the 5% significance threshold. The grey dashed vertical line indicates β_0. A similar plot is obtained for empirical powers of a test based on γ rather than β.

Copyright © 2016 John Wiley & Sons, Ltd.

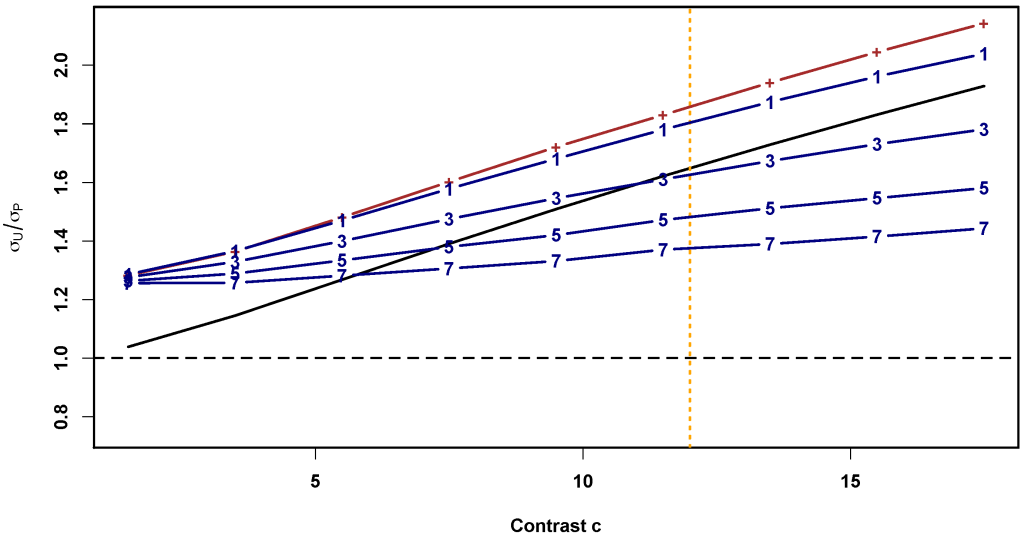


Figure 4. Linear variation of the ratio $\hat{\sigma}_{\hat{\gamma},U}/\hat{\sigma}_{\hat{\gamma},\Delta}$ against magnitude c_1. Here the quantities in the ratio are Monte Carlo means obtained for $N_S = 5,000$ repetitions using true response $\gamma = 0.7$. The solid line shows the ideally paired case $\sigma_\Delta = 0$, using exact analytic expressions (9) and (12) for $\hat{\sigma}_{\hat{\gamma},U}$ and $\hat{\sigma}_{\hat{\gamma},\Delta}$ respectively. The highest curve (brown, '+') is obtained using the corresponding sample-estimated standard errors (10) and (11) in the ratio, under perfect registration. The other, navy broken lines, from top to bottom, correspond to $\sigma_\Delta = 1, 3, 5$ and 7. In this experiment the ratios systematically indicate a benefit in pairing. Here we used $\tau = 1$; an increase in τ would result in an increasing ratio $\hat{\sigma}_{\hat{\gamma},U}/\hat{\sigma}_{\hat{\gamma},\Delta}$, which would magnify the effect of contrast illustrated here. The vertical, dashed orange line indicates the contrast used in the other experiments illustrated in this section.

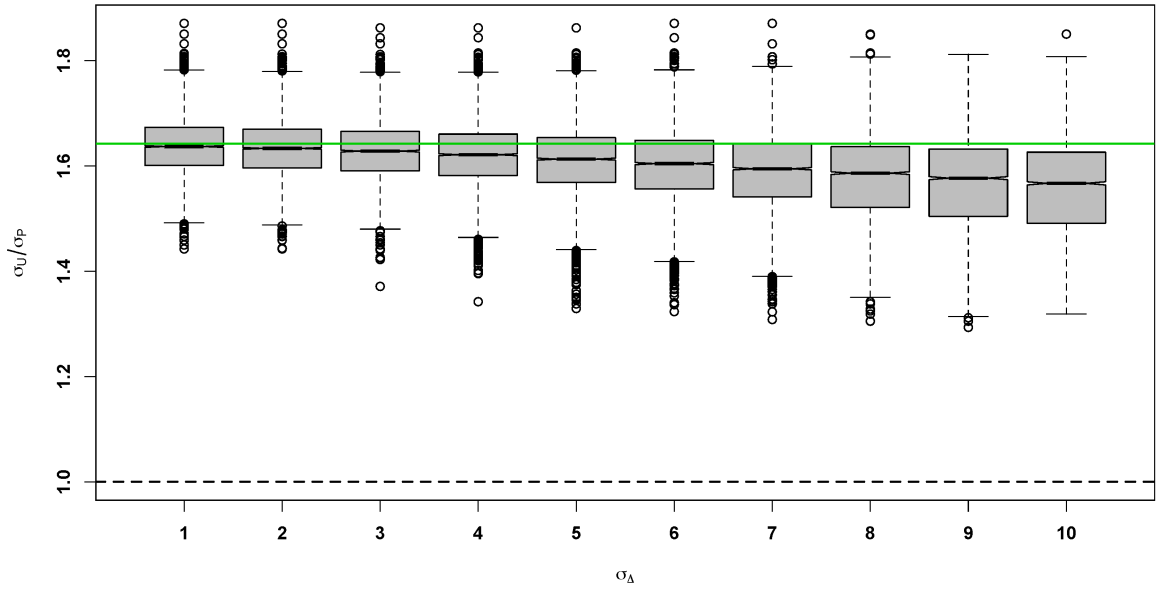


Figure 5. Distributions of the ratios $\hat{\sigma}_{\hat{\gamma},U}/\hat{\sigma}_{\hat{\gamma},\Delta}$ with respect to the typical mis-registration scale σ_Δ (in voxels), obtained for $N_S = 5,000$ Monte Carlo simulations using true response $\gamma = 0.7$. Here $\tau = 1$ was used. In this experiment pairing remains beneficial even for large mis-registration scales (the boxed volumes of interest are 16×16, taken out of simulated images of size 50×50). The straight line indicates the theoretic performance ratio (using the true ϕ value) in the case of perfect registration.

Copyright © 2016 John Wiley & Sons, Ltd.

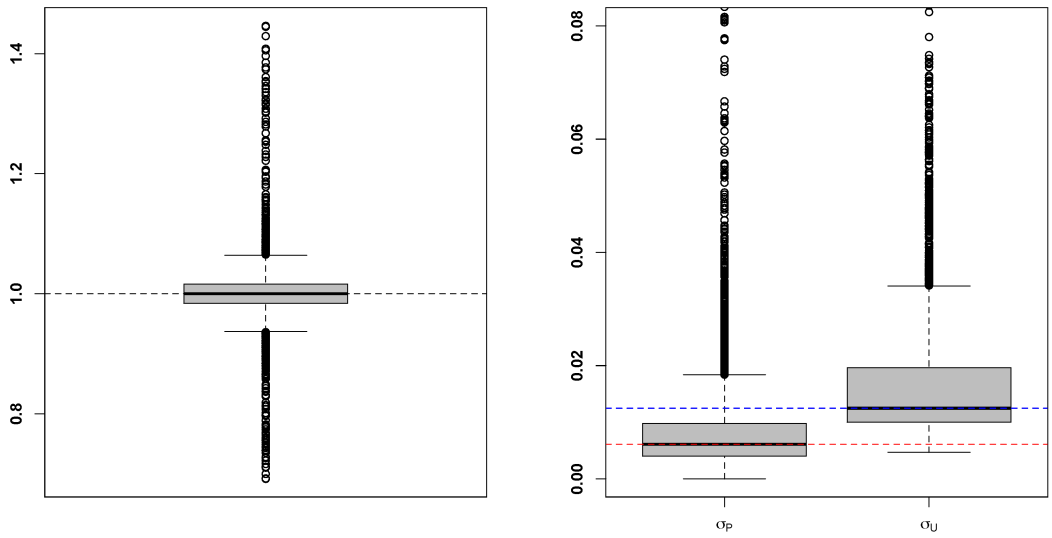


Figure 6. Results of synthetic experiment described in Section 3.3.2 involving mis-registered pre-therapy 3D sarcoma PET scans, where mis-registration is introduced artificially by mis-registering a PET scan onto itself and assessing response (hence ideally, $\hat{\gamma} = 1$). For each scan, the response analysis is carried out for perfectly registered sets as well as for all mis-registered scenarios in the 3 directions (i.e. x, y and z) vary using shifts $\Delta \in \{-2, \dots, 2\}^3 \subset \mathbb{Z}^3$ successively. The merit of this experiment is that it allows testing the proposed analysis on real PET data, with realistic noise characteristics, and uptake distributions that may depart from experimental model (3). Left: distribution of $\hat{\gamma}$ estimates for all 50 studies and all 125 scenarios, exhibiting reliability of the quantitation of therapeutic effectiveness for most cases. Right: corresponding distributions of $\hat{\sigma}_{\hat{\gamma},P}$ (leftmost boxplot) and $\hat{\sigma}_{\hat{\gamma},U}$ (rightmost boxplot) for all 125 scenarios and all patient datasets. This panel is truncated above to magnify comparison between distributions of paired and unpaired quantitations; in this experiment we found that the 99^{th} quantiles for these were respectively 0.0552 and 0.0959 (with resp. $\max(\hat{\sigma}_{\hat{\gamma},P}) = 0.3311$ and $\max(\hat{\sigma}_{\hat{\gamma},U}) = 1.0260$).

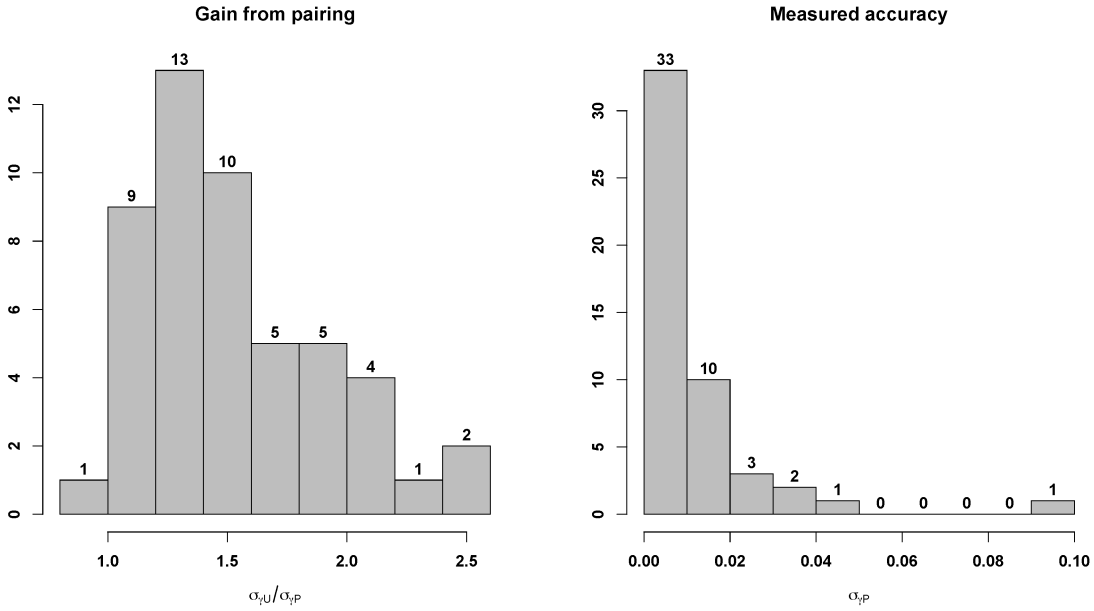


Figure 7. Left: distribution of observed ratios $\hat{\sigma}_{\hat{\gamma},U}/\hat{\sigma}_{\hat{\gamma},P}$ for the 50 sarcoma studies; pairing was helpful in evaluating the assessment standard error for all but one study (i.e. 98% of the dataset; we observed $\hat{\sigma}_{\hat{\gamma},U}/\hat{\sigma}_{\hat{\gamma},P} = 0.967$ for the one case yielding a smaller standard error for the unpaired assessment–see Section 4.2). Right: distribution of values of the paired standard errors for our dataset; the order of magnitude of the measured quantity suggested it has practical utility for our dataset (with $\hat{\sigma}_{\hat{\gamma},P} < 2\%$ for 86% of studies, and $\hat{\sigma}_{\hat{\gamma},P} > 4\%$ for 4% of cases). A correlation of 0.986 was observed between the collections of values for $\hat{\sigma}_{\hat{\gamma},U}$ and $\hat{\sigma}_{\hat{\gamma},P}$, confirming that they are qualitatively comparable.

Copyright © 2016 John Wiley & Sons, Ltd.

Table 3. Area Under the ROC Curve (AUC), Value of the Likelihood Ratio (LR) and associated p-value obtained by univariate Cox survival analysis for OS and PFS. Variables are ordered by decreasing level of LR for OS. Bold p-values indicate significance at the 5% level. Other values in bold indicate best performance.

	OS			PFS		
Variable	AUC	LR	p-value	AUC	LR	p-value
$\%\text{SUV}_{\text{mean}}(=\hat{\gamma}-1)$	**76.8**	**4.450**	**0.005**	**74.0**	**4.994**	**0.002**
$\%\text{SUV}_{\text{max}}$	75.1	3.047	**0.011**	71.5	2.745	**0.016**
$\%\text{SUV}_{\text{peak}}$	75.9	2.902	**0.014**	70.6	2.400	**0.024**
t_{QP}	67.9	1.580	0.091	63.3	1.574	0.089
$t_{QP,u}$	69.3	1.435	0.118	63.8	1.483	0.108
site	58.6	1.205	0.103	61.0	2.334	**0.022**
t-test	64.2	0.980	0.172	60.7	0.851	0.203
$\text{SUV}_{\text{max}0}$	58.2	0.304	0.408	51.2	0.044	0.762
age	53.8	0.299	0.443	50.3	0.006	0.912
$\hat{\sigma}_{\hat{\gamma},P}$	62.0	0.180	0.594	63.5	0.441	0.428
$\text{SUV}_{\text{peak}0}$	57.4	0.135	0.588	51.5	0.003	0.939
$\hat{\sigma}_{\hat{\gamma},U}$	58.5	0.103	0.672	62.7	0.329	0.471
$\text{SUV}_{\text{mean}0}$	56.7	0.091	0.677	62.0	0.534	0.324
grade	52.3	0.012	0.875	55.4	0.127	0.613
gender	50.8	0.006	0.911	58.6	0.512	0.320

Table 4. Multivariate Cox regression results for OS and PFS for the model selected on the basis of minimized AIC. The symbols $^+$, *, ** and *** correspond to p-values below 0.1, 0.05, 0.01 and 0.001 respectively. C indicates concordance for the model, i.e. the proportion of correct predictions from the model for this dataset.

	OS ($C=0.79$)		PFS ($C=0.75$)	
Variable	Hazard Ratio	95% C.I.	Hazard Ratio	95% C.I.
age	0.42	[0.21;0.83]*	0.81	[0.46;1.42]
site	0.25	[0.08;0.83]*	0.18	[0.05;0.57]**
$\hat{\gamma}-1$	14.18	[3.34;60.20]***	6.23	[2.18;17.82]***
t_{QP}	0.17	[0.05;0.59]**	0.32	[0.13;0.78]*
$t_{QP}\times$ age	2.06	[1.04;4.05]*	1.51	[0.91;2.51]
Likelihood Ratio test	$LR=8.30$, p-value $=0.0157$*		$LR=5.79$, p-value $=0.0553^+$	

Copyright © 2016 John Wiley & Sons, Ltd.

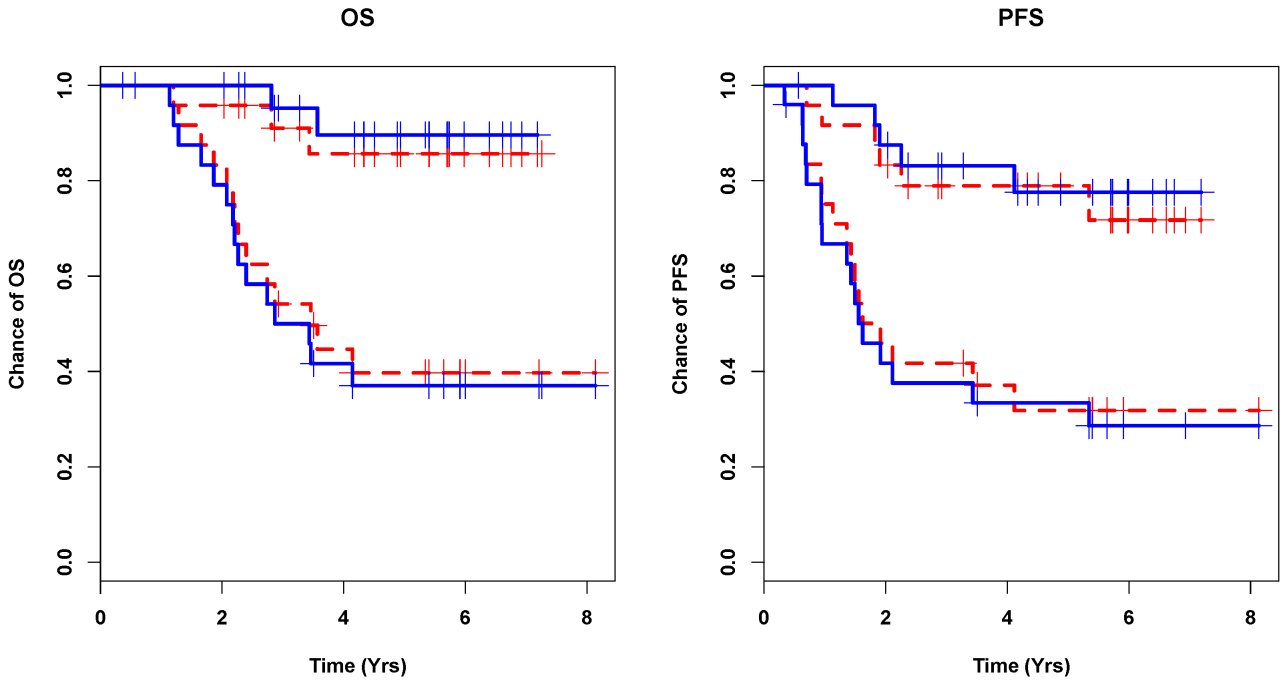


Figure 8. Kaplan-Meier curves for OS (left) and PFS (right) for the model used in Table 4 with covariates age, site, $\hat{\gamma} - 1$, t_{QP} and $t_{QP} \times$age. In each plot the solid blue curves correspond to the model that includes t_{QP}, and the dashed red curves, to the model without this quantitation. Separation between low-risk (top curves) and high-risk (lower curves) patients obtained using the full model is found to be statistically significant at the 0.5% level ($p = 0.0018$ for OS, $p = 0.0021$ for PFS). Note that the same separation between high- and low-groups using the reduced model (i.e. without t_{QP} and $t_{QP} \times$age) is also significant, at a higher significance level ($p = 0.0198$ for OS and $p = 0.0194$ for PFS).

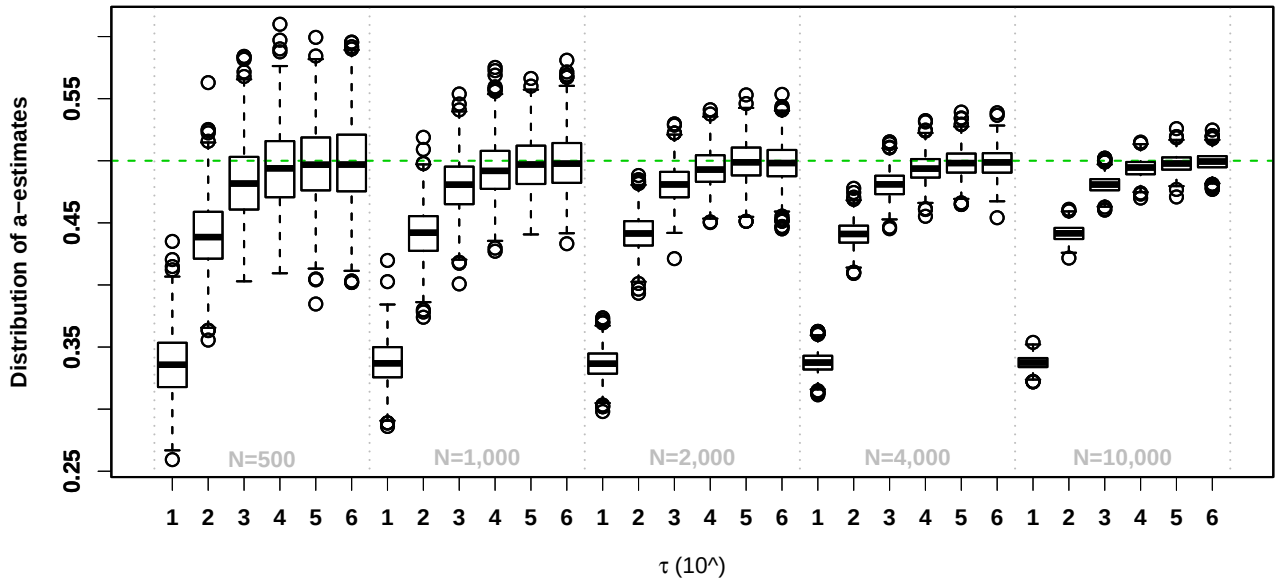


Figure 9. Distributions of estimates of a in a dose-controlled experiment, where $\phi_0 = a/\tau_0$ and $\phi_1 = a/\tau_1$ for known doses τ_0 and τ_1. Here we set $\tau_0 = 2\tau_1$. This Monte Carlo simulation demonstrates that accurate estimation of model scale is achieved in a model where different scales are allowed the pre- and post-therapy uptake samples. The horizontal green broken line indicates the true value of a. The five groups of boxplots correspond to different sample sizes, with respectively $N \in \{500, 1000, 2000, 4000, 10000\}$.

Copyright © 2016 John Wiley & Sons, Ltd.

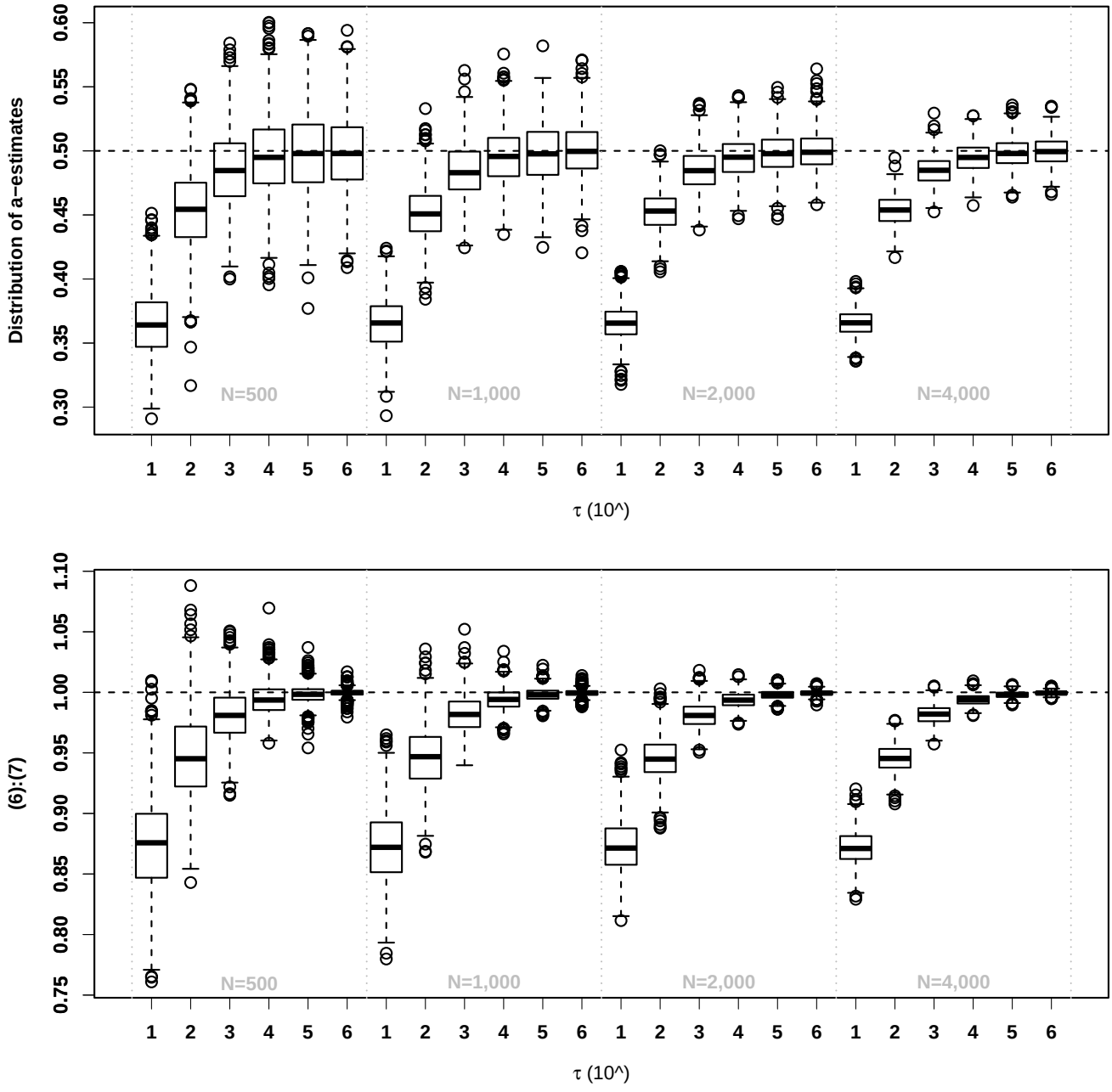


Figure 10. Top: distributions of estimates of $a = 0.5$ (horizontal dashed line) in a simulation framework similar to that of Figure 9 but where $\phi_0 = \phi_1$, i.e. $\tau_0 = \tau_1$. In each sample size grouping $N \in \{500, 1000, 2000, 4000\}$, Monte Carlo distributions are obtained for varying value of $\tau \in \{10^1, 10^2, 10^3, 10^4, 10^5, 10^6\}$. Bottom: ratio of estimates of ϕ obtained from paired constructions (6) and (7). A ratio of 1 (dashed line) indicates equivalence between the two estimators. These results suggest that estimator (7) tends to yield greater values than (6) especially at low doses, but that the two estimators become equivalent at higher doses, irrespective of sample size N. The case $\tau = 10^4$ would correspond to a typical dose used in routine practice.

Copyright © 2016 John Wiley & Sons, Ltd.

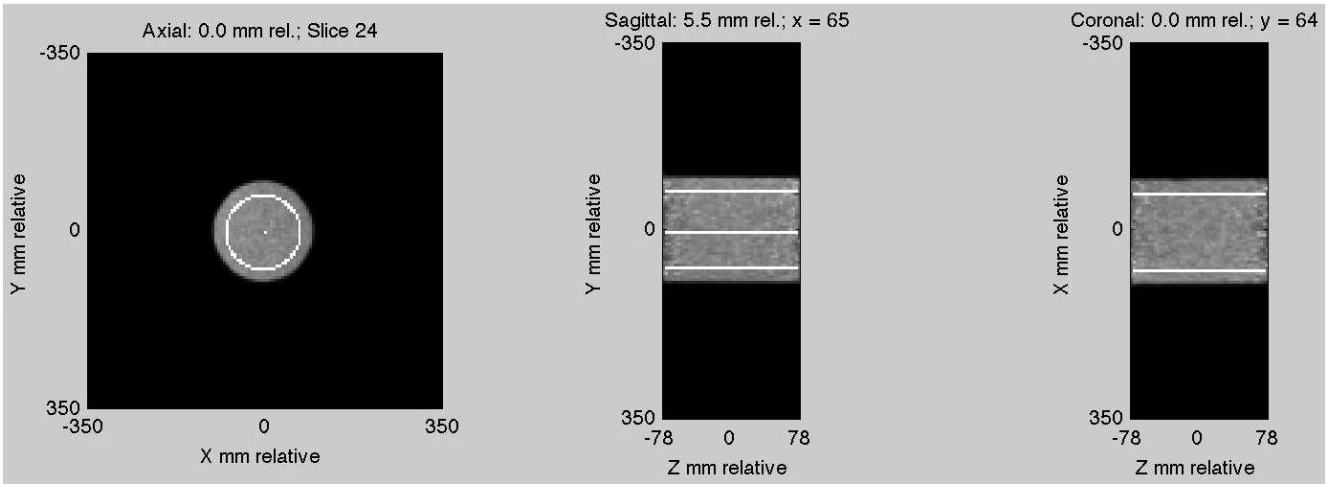


Figure 11. View of the ACRIN phantom dataset used for the validation study of the Gamma model, presented in Appendix C.

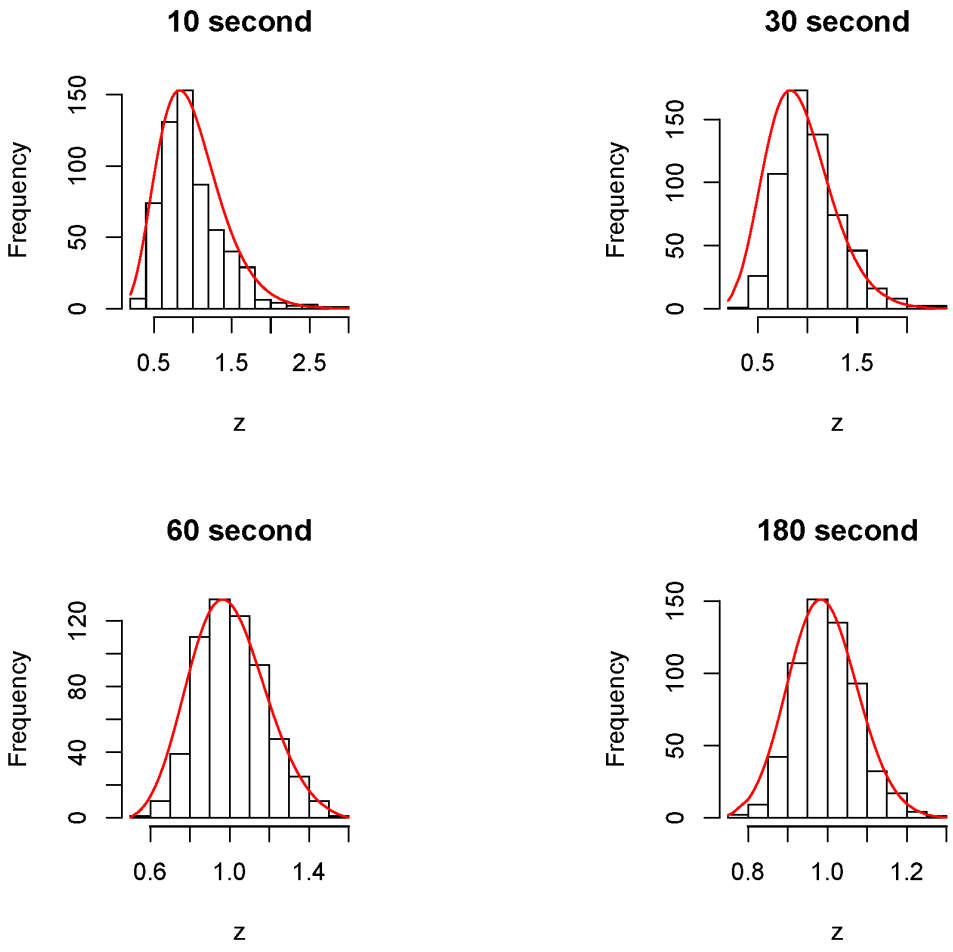


Figure 12. Validation study of the Gamma model for simulated illustrative experiments, based on the ACRIN qualification data whown in Figure 11. The histograms are obtained from four different frames of varying duration (activity is proportional to scan duration) for a slice in the centre of the scanner. The histograms show the distribution of reconstructed PET values within the phantom slice. The Gamma density model fit to the data is indicated by the red line. In each case the Kolmogorov-Smirnov test finds no significant difference between the data and the Gamma model. This is typical.

Copyright © 2016 John Wiley & Sons, Ltd.