CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Comparison of AI-assisted and human-produced podcasts derived from cochrane plain language summaries: Protocol for a randomized noninferiority trial (Health Information Effectiveness Trials-2)
Authors	O'Byrne, Isabel;Pope, Johanna;Byrne, Paula;Dahly, Darren;Judge, Conor;O'Donnell, Martin;Krewer, Finn;Li, Mengqi;Saif-Ur-Rahman, K. M.;Conway, Tom;Thomas, James;Sheils, Geraldine;Burke, Nikita;O'Byrne, Marie;O'Byrne, Aislinn;Devane, Declan
Publication date	2026-05-19
Original Citation	O'Byrne, I, Pope, J, Byrne, P, Dahly, D, Judge, C, O'Donnell, M, Krewer, F, Li, M, Saif-Ur-Rahman, K M, Conway, T, Thomas, J, Sheils, G, Burke, N, O'Byrne, M, O'Byrne, A & Devane, D 2026, 'Comparison of AI-assisted and human-produced podcasts derived from cochrane plain language summaries: Protocol for a randomized noninferiority trial (Health Information Effectiveness Trials-2)', Journal of Clinical Epidemiology, vol. 195, 112272, pp. 1-11. https://doi.org/10.1016/j.jclinepi.2026.112272
Type of publication	Article (peer-reviewed)
Link to publisher's version	https://www.scopus.com/pages/publications/105039272177 - 10.1016/j.jclinepi.2026.112272
Rights	© 2026, authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/) - https://creativecommons.org/licenses/by/4.0/
Download date	2026-11-15 16:33:17
Item downloaded from	https://hdl.handle.net/10468/19469



University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Journal of Clinical Epidemiology 195 (2026) 112272

PROTOCOL

Comparison of AI-assisted and human-produced podcasts derived from Cochrane Plain Language Summaries: protocol for a randomised non-inferiority trial (Health Information Effectiveness Trials-2)

Isabel O'Byrne[a,*], Johanna Pope[b,c,d], Paula Byrne[c,d], Darren Dahly[e,f], Conor Judge[a], Martin O'Donnell[a], Finn Krewer[a], Mengqi Li[b,d], KM Saif-Ur-Rahman[c,d], Tom Conway[b,c,d], James Thomas[g], Geraldine Sheils[h,i], Nikita Burke[c,d], Marie O'Byrne[d,1], Aislinn O'Byrne[d,1], Declan Devane[b,c,h]

[a]*School of Medicine, University of Galway, Galway, Ireland*
[b]*HRB-Trials Methodology Research Network, University of Galway, Galway, Ireland*
[c]*Evidence Synthesis Ireland & Cochrane Ireland, University of Galway, Galway, Ireland*
[d]*Centre for Health Research Methods, School of Nursing and Midwifery, University of Galway, Galway, Ireland*
[e]*HRB Clinical Research Facility, University College Cork, Cork, Ireland*
[f]*School of Public Health, University College Cork, Cork, Ireland*
[g]*EPPI Centre, UCL Social Research Institute, University College London, London, UK*
[h]*School of Nursing and Midwifery, University of Galway, Galway, Ireland*
[i]*Senior Occupational Therapist, HSE Primary Care Paediatric Services, Ballinasloe, Co Galway, Ireland*

Accepted 9 April 2026; Published online 20 April 2026

Abstract

Background and objectives: Podcasts can make health evidence easier to follow, but it is unclear whether artificial intelligence (AI)-assisted production can match human production when both use the same audio format. This trial will evaluate whether a vetted AI workflow can match human communicators on comprehension, quality, safety, accessibility, and trust when both deliver podcasts derived from the same evidence base.

Methods: We will run a randomized, two-arm, noninferiority trial comparing AI-assisted podcasts with human-produced podcasts. Adults (≥18 years; English-proficient) will be recruited from the general public via Prolific, an online research participant recruitment platform, and randomly allocated 1:1 to listen to three short episodes (6−8 minutes each) based on the same Cochrane Plain Language Summaries. The AI arm uses Wondercraft AI in a human-in-the-loop workflow; the human arm features experienced communicators working to an identical brief. In both arms, content is limited to the plain language summary, with authorship masked for participants and expert raters. The primary outcome is comprehension, measured by a 10-item test per episode, with the primary analysis using the participant-level mean score across the three episodes, aligned with the ''Understanding'' dimension of the Quality, Understanding, Expression, Safety, and Trust framework. Secondary outcomes include format accessibility (listenability), quality of information, perceived trust, and safety. Noninferiority margins are prespecified; for comprehension, the margin is 1 point on the 10-item scale. If noninferiority is shown, we will also assess superiority. We plan to recruit 458 participants. Differences between arms will be estimated using appropriate repeated-measures models, with two-sided 95% confidence intervals.

Trial registration ID: ISRCTN28623546, https://doi.org/10.1186/ISRCTN28623546.

Protocol version 2.0, September 15, 2025

Date registered: February 4, 2025

OSF link: https://osf.io/y9hvj/?view_only=c9688c06b34c465994073be1baa9e9a6.

Sponsor: University of Galway (nonregulated behavioral trial; sponsor role limited to institutional oversight and governance). The sponsor has no role in the study design; collection, management, analysis, and interpretation of data; writing of the report; or decision to submit the report for publication.

Funding: We are grateful for support from the Health Research Board (HRB), the Health Service Executive (HSE), Grant no. INFO-2021-001, and the College of Medicine, Nursing and Health Sciences at University of Galway. The funders will have no role in study design, data collection, analysis, decision to publish, or manuscript preparation. Isabel O'Byrne was awarded an HRB-Trials Methodology Research Network (HRB-TMRN) and Evidence Synthesis Ireland (ESI) Summer Scholarship to support this work. The HRB-TMRN is funded by the Health Research Board (HRB, Ireland), and ESI is funded by the HRB and the Public Health Agency (Northern Ireland).

[1] Public partners

* Corresponding author. School of Medicine, University of Galway, Galway, Ireland.

E-mail address: I.OBYRNE1@universityofgalway.ie (I. O'Byrne).

https://doi.org/10.1016/j.jclinepi.2026.112272

0895-4356/Crown Copyright © 2026 Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Results: This is a study protocol; results will be reported on completion of the trial.

Conclusion: By providing head-to-head evidence in the same audio format, the study will address a practical question faced by journals and health organizations already experimenting with AI tools: can AI generate clear, safe, and trusted audio content at scale, and where does human input remain essential? Crown Copyright © 2026 Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Keywords: Artificial intelligence; Podcasts; Plain language summaries; Health communication; Systematic reviews; Noninferiority trial; QUEST framework; Cochrane reviews

1. Background

Effective communication of health information is vital for enabling the public to make informed decisions. However, many people find scientific or technical language hard to understand due to complex jargon, statistical terms, and descriptions of methods [1]. Plain language summaries (PLSs) present systematic review findings in clear, simple language that the public can understand. Podcasts have been shown to be an effective medium of communication [2], and there is growing evidence that they could help people understand complex health information [3]; global podcast listenership reached an estimated 505 million in 2024 and continues to grow, with health and wellness among the most popular categories [4].

In the context of Cochrane reviews, PLSs aim to present the findings of health research in a more accessible and understandable format. They follow specific methodological guidance to ensure they are accurate, consistent, and accessible. A key challenge in creating a PLS is striking a balance between accuracy and accessibility for a wider audience [5]. This is challenging because simplifying complex information can lead to the loss of important details and nuances, which are crucial for accurately interpreting the findings. For the public, these details provide the necessary context to understand the reliability and limitations of the evidence; without them, messages may be misunderstood or misused.

Although they are valuable, traditional written PLSs have inherent limitations; many are written at a reading level too high for broad accessibility and may not be engaging or accessible for individuals with low literacy [6]. Audio-based health communication improves accessibility through convenience and achieves strong engagement and knowledge gains in general audiences [7,8]. To meet diverse communication needs, researchers have been exploring alternative formats for sharing research findings. Podcasts are increasingly popular for engaging audiences across a range of topics, with some studies showing higher engagement than in text-based formats [3]. Growing evidence indicates that podcasts are effective tools for communicating research findings and supporting learning [9], making them a promising method for sharing accurate health information.

Recent advances in artificial intelligence (AI) have introduced new methods for presenting health information in podcast form. AI tools, such as generative AI, can create podcast scripts and offer a range of human-like voices, potentially saving time and resources. A human-in-the-loop approach that integrates AI capabilities with human expertise ensures both efficiency and accuracy in content creation. By combining AI-generated content with expert oversight, complex research can be turned into accessible, listener-friendly stories, while the oversight also safeguards the accuracy of the information. This approach may enable faster production of engaging health evidence resources while maintaining accuracy and balance. However, concerns about accuracy, bias, and trustworthiness highlight the importance of a human-in-the-loop approach [10]. In line with these considerations, our AI-assisted process employs a vetted workflow with Wondercraft AI, structured prompts, and expert review to ensure clarity, balance, and fidelity to the source PLSs.

As AI becomes increasingly integrated into health care and health communication, it is essential to evaluate specific AI-assisted workflows in controlled settings. Without this evaluation, there is a risk that AI-generated content could oversimplify research findings and misinform the public, potentially leading to suboptimal decision-making [11]. Evaluations must assess both technical accuracy and the potential for harm. The Quality, Understanding, Expression, Safety, and Trust (QUEST) framework, developed through systematic analysis of 142 health-care large language model (LLM) evaluation studies, provides multidimensional criteria for evaluating LLMs in health care [12]. This trial focuses on a specific, vetted AI-assisted workflow with human oversight rather than evaluating AI-generated content in general.

Currently, limited evidence exists comparing the effectiveness of AI-assisted podcast formats with human-produced podcasts when the format is the same for communicating systematic review findings. We conducted targeted searches of PubMed, the Cochrane Library, and Google Scholar and identified no published trials directly comparing AI-assisted and human-produced podcasts for communicating systematic review findings. This limits our understanding of the most effective ways to communicate health evidence.

This study compares two methods for producing podcasts from Cochrane PLS: human-produced podcasts and AI-assisted podcasts created with Wondercraft AI within a human-in-the-loop workflow. Both groups utilize the

What is new?

Key findings

- AI-assisted podcasts compared with human-produced podcasts from Cochrane PLSs.
- Outcomes include comprehension, accessibility, quality, trust, and safety.

What this adds to what is known?

- Written PLSs aid public communication; audio may improve accessibility.
- AI could speed podcast production, but accuracy and reliability need evaluation.

What is the implication and what should change now?

- Results will show whether AI-assisted podcasts can enhance evidence communication.

same set of Cochrane PLS sources and follow identical review and quality assurance steps. Podcast authorship (ie, whether the podcast was AI-assisted or human-produced) is masked for participants and expert raters during assessment to minimize bias. This design maintains the same format and isolates the impact of authorship within the audio.

This study, the second in the Health Information Effectiveness Trials (HIET) series [13], complements HIET-1, which compared AI-assisted and human-generated written PLSs of Cochrane reviews. While HIET-1 isolated authorship effects in text (ie, the written word), HIET-2 isolates authorship effects in audio (ie, podcasts). Both trials share the same noninferiority design, QUEST-aligned outcome framework, and human-in-the-loop verification process but differ in the communication modality being tested. Here, we examine whether a validated AI workflow can match a human communicator in terms of comprehension, quality, safety, accessibility, and trust when both deliver podcasts derived from the same evidence base.

1.1. Trial design

This randomized, parallel-group, two-arm noninferiority trial will be conducted entirely online using digital platforms to recruit participants and deliver all study materials.

Participants will be allocated 1:1 to two podcast arms using the same audio format: a human-produced podcast arm and an AI-assisted podcast arm created with Wondercraft AI within a vetted human-in-the-loop workflow. Both arms will draw on the same Cochrane PLS sources and follow identical review and quality-assurance steps. The identity of the podcast creator will be masked for participants and expert raters to reduce bias. This design maintains the same format and isolates authorship effects within the audio. All participants will receive the same three topics across arms so that the topic mix is identical for the AI-assisted and human-produced groups.

1.2. Methods: recruitment, interventions, and outcomes

The overall methodological approach, including the design, outcomes, and evaluation framework, is guided by the HIET-1 methods [13]. This protocol is reported in accordance with the SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) 2025 Statement [14]. A completed SPIRIT 2025 checklist is provided in Supplementary File 1.

1.2.1. Recruitment

Participants will be recruited via Prolific (https://www.prolific.com/), an online platform that connects researchers with verified participants. The platform enables the recruitment of participants from various age groups, locations, and educational backgrounds. Participants will be compensated in accordance with Prolific's standard payment practices.

The eligibility criteria are:

Inclusion criteria: Participants must

1. Be 18 years or older,
2. Be proficient in English, as the study materials will be delivered in English. Participants will be asked to self-assess their English proficiency on a scale of 1 (not comfortable) to 10 (very comfortable). Only those who rate themselves as 7 or higher will be eligible to take part,
3. Have access to the internet,
4. Have a device to access the study materials,
5. Provide consent before starting this study.

Exclusion criteria include the inability to complete the online survey. Additionally, for the per-protocol analysis, responses will be excluded if participants take less than 25 minutes to finish all three podcasts and their associated questions (see Statistical methods) or show evidence of straight-line answering (defined as selecting the same response option for $\geq$80% of Likert scale items within any single assessment block where variation would be expected, excluding comprehension questions which have correct/incorrect answers). All randomized participants with at least one postbaseline outcome will be included in the sensitivity intention-to-treat analysis.

1.2.2. Interventions

Each participant will receive three podcasts, purposefully selected to cover different health conditions and levels of complexity. The target duration is 6 to 8 minutes per episode. Both arms will base their content solely on the published Cochrane PLS for each review, use the same structured brief and checklists, and undergo identical

multi-reviewer quality and public and patient partner reviews. This keeps the format consistent and aligns the review steps across both groups.

As in HIET-1 [13], reviews on which podcasts are based will be selected based on the following criteria:

1. Focus on the effectiveness of health-care interventions.
2. The effectiveness of health-care interventions is measured by clear outcomes such as treatment success or failure.
3. Relevance to relatively common public health issues and the ability to produce summaries that a general, nonexpert audience can understand.
4. Diversity in complexity levels is assessed using standardized readability metrics.
5. Topics that address health decisions or conditions that members of the general public might reasonably encounter, either personally or as caregivers.
6. Inclusion of various types of interventions (eg, pharmacological, behavioral).

The podcast development process follows a prespecified, mirrored workflow for both arms.

7. A comprehensive draft brief has been developed following the Cochrane PLS guidelines adapted for audio formats (see Supplementary File 2). This brief details the podcast's purpose, target audience, duration, language style, and tone of speaker interaction. It will undergo pilot testing and iterative refinement to ensure that the language is accessible, certainty statements are clear, and both benefits and harms are presented fairly.

Both arms will rely exclusively on the Cochrane PLS as the content source. This approach achieves several goals including ensuring consistency across all podcasts, minimizing reviewer bias, facilitating valid comparisons between study groups, and replicating real-world situations where podcast creators often have access only to summaries rather than full reviews. By restricting input to the PLS, both workflows focus on presentation and explanation rather than de novo evidence synthesis.

The podcast scripts in both arms will remain closely aligned with the original PLS. Guidance on tone, conversational style, and two-speaker interaction (see Supplementary File 2) will support an informal, listener-friendly presentation. While no new information will be added beyond what is in the PLS, these adjustments aim to enhance engagement and comprehension.

8. Two experts will evaluate the initial scripts from both arms using a structured checklist to assess accuracy, clarity, and engagement. In the AI-assisted arm, refinement will be conducted primarily by reprompting the AI to regenerate the script. Before any human edits, the AI script will undergo a prespecified self-verification pass (Supplementary File 3) that (a) maps all population, intervention, comparator, and outcomes (PICO) elements and checks alignment with the source PLS and (b) flags complex health terms and proposes plain language explanations rooted in the PLS wording. Issues identified at this stage will be addressed via AI regeneration. Human edits are limited to correcting any residual factual errors, removing unsupported claims, and improving clarity (eg, grammar, pacing) without adding content beyond the PLS. Up to three refinement rounds will be used, with stopping rules: (1) both reviewers rate the script acceptable on all checklist criteria, (2) if cumulative human edits to the AI script would exceed 5% of content, an additional AI regeneration is performed rather than manual rewriting, or (3) three rounds completed. Example refinements include clarifying benefits and harms, preserving a natural conversational tone, and explaining terms consistently with the PLS.
9. At least one content expert will review each podcast script, as needed, to adjust complex health terms and make them more accessible to the general public. The expert will also verify that the PICO elements and findings are accurately represented and that any uncertainties and limitations are presented appropriately. In the AI-assisted arm, this content expert review occurs after the self-verification pass and focuses on confirming PICO mapping and plain language definitions, with any required changes addressed via AI regeneration wherever possible. Before release, a core safety checklist (Supplementary File 4) will be applied to each final script and recording. Once approved, the audio is frozen for the trial.
10. At least one PPI partner who is not a study participant or outcome assessor will review prerelease audio for auditory comprehensibility, clarity, and accessibility, using a brief orientation and a structured checklist. PPI reviewers will be masked to authorship (AI-assisted vs human-produced) and will receive deidentified files and neutral filenames/transcripts. Feedback will be limited to comprehension, plain language definitions, and accessibility (no additions beyond the PLS). Suggested changes will be incorporated using the same revision pathway in both arms.
11. Participant-facing materials will not reveal whether a podcast was created by a human or with AI assistance. File names and metadata that could disclose authorship will be removed. Transcripts provided to expert raters will similarly be deidentified and neutralized for production notes.
12. Process documentation will record prompts, iterative refinements, and rationale for material changes. To assess content fidelity to the source PLS, two raters will use a structured checklist (coverage of key

messages; accuracy of benefits/harms; certainty statements; absence of unsupported claims). We will report simple descriptive text similarity indicators (eg, cosine similarity of sentence embeddings) as descriptive only, recognizing that high similarity can still mask meaning changes; fidelity judgments will rely on human assessment. For transparency, we will also record per-episode process metrics, including the number of AI regenerations, the number of verification issues flagged and resolved, and the proportion and category of human edits (accuracy, definitions/clarity, and style). Episodes requiring extensive manual rewriting will be flagged with a descriptive note. A final QUEST-aligned quality check and second topic expert review will confirm adherence to standards. For study outcomes, blinded expert ratings of quality and safety will be conducted on deidentified transcripts after the audio is frozen; these outcome ratings will not inform edits to the released audio.

1.2.3. Arm-specific implementation details

1.2.3.1. AI-assisted podcasts (Wondercraft AI). Scripts will be generated with Wondercraft AI (https://www.wondercraft.ai/studio) (Supplementary File 5 for the full structured prompt) using a structured brief aligned with the Cochrane PLS template and the common brief. Text-to-speech voices will be used for recording. Wondercraft AI was selected because it provides an integrated workflow for script generation, voice synthesis, and audio production within a single platform, supports structured prompt input, offers human-like text-to-speech voices suitable for podcast delivery, and enables iterative refinement within a human-in-the-loop workflow. Alternative tools were considered but did not offer the same combination of integrated script-to-audio production and structured prompt support required for a controlled, reproducible workflow. Process documentation will record prompts, iterative refinements, and rationale for material changes. A self-verification pass will be run on the draft script to produce an explicit PICO mapping and a list of complex terms with plain language explanations; flagged issues will be resolved by AI regeneration rather than human rewriting.

1.2.3.2. Human-produced podcasts. Human-produced podcasts were selected as the comparator because they reflect current practice and provide an active control for authorship effects. Scripts will be drafted by an experienced science communicator using only the PLS and the same structured brief. Human narrators will record the episodes. Editorial revisions will follow the identical multi-reviewer workflow and stopping rules used in the AI-assisted arm. The same PICO/term clarity checklist used during the verification pass will be applied, but without AI assistance. Edits will be documented in the same categories (accuracy, definitions/clarity, and style) to maintain consistency.

1.3. Outcomes

We will evaluate outcomes using the QUEST framework for health-care LLMs [12], aligned with HIET-1 [13] and adapted for an audio intervention. We will interpret between-arm differences as reflecting authorship effects within audio.

1.3.1. Primary

1.3.1.1. Comprehension (aligned with QUEST's Understanding dimension). Participants will complete a standardized 10-item multiple-choice questionnaire (Supplementary File 6) for each podcast, structured to align with the Cochrane PLS template [15]. The questionnaire will systematically assess understanding across five domains:

1. Understanding of review topic (two items)
   - Understanding of the health condition/problem
   - Recognition of review importance
2. Review aims and methods (two items)
   - Comprehension of the main review question
   - Basic understanding of the evidence-gathering approach
3. Main results (three items)
   - Understanding of key benefits
   - Understanding of unwanted effects/harms
   - Grasp of the size of the evidence base
4. Evidence quality and limitations (two items)
   - Recognition of main limitations
   - Understanding of evidence strength/certainty
5. Currency of evidence (one item)
   - Awareness of how current the evidence is

Each question will use plain language as defined in the Cochrane PLS guidance, avoid technical terms without explanation, and include four response options with one correct answer. For the primary outcome of comprehension, a noninferiority margin of one point (10% points) will be applied. For instance, if the human-produced podcast achieves a mean score of 8 out of 10 (80% correct), the AI-assisted podcast would be considered noninferior, provided the lower bound of its two-sided 95% confidence interval (CI) exceeds 7 out of 10 (70% correct) (ie, no more than one point lower than the control). For the primary analysis, we will calculate for each participant the mean total comprehension score across the three podcasts (range 0 to 10) and compare these participant-level means between arms using mixed-effects models with random intercepts

for participant and episode. Episode-level analyses will be explored descriptively.

1.3.2. Secondary

1.3.2.1. Format accessibility. Format accessibility will be assessed using a hybrid approach that combines objective script readability with subjective user ratings (Supplementary file 2). Transcript readability will be measured using the Flesch-Kincaid Grade Level for both podcast arms, which estimates the US school grade needed to understand the text by analyzing sentence length and syllable count. Higher scores indicate more complex text. Scores will be generated automatically via readabilityformulas.com, using multiple formulas to assess how easily the text can be read and understood. The Flesch-Kincaid Grade Level will remain the primary objective metric and will be summarized descriptively for each podcast and arm; no formal hypothesis testing or noninferiority margin will be applied to this transcript-level metric because the number of podcasts is small and readability is a property of the scripts rather than individual participants.

Listener ratings will be captured using a five-item, five-point Likert questionnaire (1 = strongly disagree to 5 = strongly agree) as the primary indicator of listenability within a single audio format.

Participants in both arms will rate:

1. I liked the tone of the audio
2. The pacing was easy to follow
3. I could understand everything clearly
4. The audio was clear without distracting background noise
5. The podcast kept me engaged throughout

The mean score will be reported descriptively and compared using the same 0.5-point noninferiority margin. This outcome assesses accessibility within a single audio modality.

1.3.2.2. Quality of information. Two expert raters (systematic review methodologists with expertise in evidence synthesis and health communication, independent of the podcast production process) will evaluate podcasts from both arms by comparing each directly to the original Cochrane review. This approach avoids treating either arm as the implicit "gold standard." The assessment will focus on four main types of errors, with the error categories applied equally across both arms:

1. Incorrect output (where the content generated wrong information)
2. Irrelevant output (where the content generated unnecessary or off-topic information)
3. Omissions (where the content failed to include key information that should be present)
4. Currency errors (where information is outdated or inconsistent with current evidence)

Quality will be rated on a 1 to 3 scale: 1: poor quality (significant errors that mislead the listener), 2: moderate quality (minor errors that do not significantly alter understanding), and 3: high quality (no errors).

We set a noninferiority margin of 0.3 points on the three-point scale. The AI-assisted podcasts will be considered noninferior if their mean quality score is no more than 0.3 points lower than that of the human-produced podcasts. Accuracy checks will reference the full Cochrane review only for elements that appear in the PLS. No penalty will be applied for information absent from the PLS, since both arms are constrained to PLS content. Expert ratings will be analyzed using mixed-effects models with fixed effects for arm and random intercepts for podcast episode and rater. These noninferiority analyses will be considered exploratory because the effective sample size is determined by the number of rated podcasts rather than the number of participants.

1.3.2.3. Safety (aligned with QUEST's Safety and Harm dimension). Two expert raters will evaluate each podcast using a blinded core safety rating form (Supplementary File 7 for the rating form).

Inter-rater agreement will be assessed using Cohen's kappa (κ), with a prespecified threshold of $\kappa \geq 0.70$ indicating acceptable agreement. Where raters disagree, a third rater will arbitrate to reach a final determination.

Core safety criteria (applied to both podcast arms):

- Risk of misinterpretation
- Presence of bias or inappropriate recommendations
- Appropriate presentation of limitations and uncertainties
- Consistency with the source Cochrane review

Additional AI-specific safety criteria (applied only to AI-assisted podcasts):

- Evidence of fabrication or hallucination
- Consistency between the initial AI output and the final AI-assisted version

The blinded comparative safety assessment will use the four core criteria listed above and reported in Supplementary File 7. For the two risk-oriented items, a score of 1 indicates that the risk is absent, and 0 indicates that the risk is present. For the two positive criteria, a score of 1 indicates that the criterion is met and 0 indicates that it is not met. The four-item total (0 to 4) will be converted to a percentage and analysed exploratorily, with a noninferiority margin of 10 percentage points. AI-specific safety issues, including suspected hallucination and consistency between the initial AI output and the final AI-assisted version, will be captured descriptively through the AI-arm verification and prerelease checklists (Supplementary

Files 3 and 4) and will not contribute to the blinded comparative safety score. All critical safety issues identified during the assessment will be documented and reported separately.

1.3.2.4. Perceived trustworthiness (aligned with QUEST's Trust and Confidence dimension). Participants will rate the trustworthiness of podcasts using a five-point Likert scale (1 = "strongly disagree" to 5 = "strongly agree") based on items adapted from established scales measuring trust in online health information. Participants will indicate their agreement with the following statements:

- "I trust the information provided in this podcast."
- "This podcast is from a reliable source."
- "I am confident in the accuracy of the information in this podcast."
- "I believe the source of this podcast has expertise in the subject matter."
- "I would use the information from this podcast to make health decisions."

The primary measure of trustworthiness will be calculated as the average score across all five items (range 1−5), with individual item scores reported descriptively as secondary outcomes. The combined score has a noninferiority margin of 0.5 points. AI-assisted podcasts will be considered noninferior if their mean trustworthiness score is no more than 0.5 points lower than that of human-produced podcasts.

Authorship perception will be assessed as an exploratory secondary outcome. After completing each podcast assessment, participants will be asked: ''Who do you think produced this podcast?'' with three response options [1]: a human health expert alone [2], AI with human expert review, or [3] not sure. The distribution of responses will be compared against the null hypothesis of equal distribution (33.3% per option) using a chi-squared goodness-of-fit test, reported separately by arm.

1.3.2.5. Participant timeline. On the day of participation, participants will be recruited, screened, and randomly assigned 1:1 to either the human-produced podcast arm or the AI-assisted podcast arm. They will receive three podcasts on the same three topics, randomized across both arms. After each podcast, participants will complete the comprehension questions. After all three podcasts, participants will complete the format, accessibility, and trustworthiness items for each podcast. All data will be collected remotely. Each participant will complete the study in a single session, estimated to take 45 to 60 minutes. This duration includes providing informed consent, listening to three podcasts, completing the assessments, and answering demographic questions. Authorship will not be disclosed to participants during the study. Prerelease accuracy and safety checks by a topic expert, along with the core safety checklist, will be completed before final recording. Blinded outcome ratings for quality and safety will then be conducted later on deidentified transcripts, without feeding back edits to the released audio. Participants may relisten to podcast episodes during assessment.

1.3.2.6. Sample size. The sample size calculation is based on detecting noninferiority in comprehension scores between the two groups. We assume a baseline comprehension score of 8 out of 10 correct (80%) for participants listening to human-produced podcasts. AI-assisted podcasts will be considered noninferior if their mean score is no more than 1 point lower on the 10-item scale (, at least 7 out of 10 correct). We set a noninferiority margin (Δ) of 1 point, meaning that the AI-assisted arm is acceptable if the lower bound of the two-sided 95% CI excludes a deficit greater than 1 point relative to the human-produced arm.

Sample size calculations indicate that 396 participants (198 per group) are required to achieve 80% power to detect noninferiority with a margin of 1 point on the 10-point comprehension scale, assuming a mean score of 8 of 10 in both groups (80% correct responses) and a two-sided 95% CI.

To account for clustering of responses (three podcasts per participant), we applied a design effect of 1.04, calculated as $1 + (3\text{-}1) \times 0.02$, where 0.02 represents the assumed intraclass correlation coefficient. This increases the required sample size to 412 participants.

Therefore, we will recruit 458 participants (229 per group) to ensure at least 412 evaluable results after accounting for an expected 10% dropout or exclusion rate. This sample size should provide sufficient power for the primary comprehension outcome; analyses of secondary outcomes (including format accessibility, quality, safety, and trust) will be interpreted as exploratory and may be underpowered to detect small between-arm differences.

Sample estimates were generated using Sealed Envelope's sample size calculator for noninferiority trials (www.sealedenvelope.com, n.d.) and were adjusted for the design effect due to clustering.

1.4. Methods: assignment of interventions

1.4.1. Allocation

1.4.1.1. Sequence generation. The allocation sequence will be managed using the QuestionPro Block Randomizer (www.questionpro.com), with participants randomly assigned at a 1:1 ratio to either the human-produced podcast arm or the AI-assisted podcast arm. Randomization will take place after consent and eligibility screening. All participants within each arm will receive the same three podcast topics, which are matched across arms as described earlier.

1.4.1.2. Allocation concealment mechanism. QuestionPro's randomization feature automatically allocates participants to either group, ensuring that the information leaflet and the demographic and outcome-related questions

remain consistent for all participants. The allocation sequence is concealed within the platform and is not accessible to the research team at the point of enrollment. No identifying information will be used to influence allocation.

1.4.1.3. Implementation. The randomization logic in the Block Flow tab of QuestionPro ensures equal allocation across groups. The assignment is automated by the platform and recorded in its audit trail.

1.4.1.4. Blinding. Authorship will be masked for participants and expert raters. File names, voice tags, and metadata that could disclose authorship will be removed. Expert raters will receive deidentified transcripts with neutralized production notes. Because residual differences between human narration and text-to-speech voices may still be present, we refer to this as masking of authorship rather than full blinding. PPI partners involved in prerelease quality checks will also be masked to authorship and receive deidentified materials; they will be briefed only on checklist use and remit, not on arm assignment.

1.5. Methods: data collection, management, and analysis

1.5.1. Data collection methods

All data collection takes place within the QuestionPro and Prolific survey platforms. Participants will listen to three podcasts through the survey interface in their assigned group. Comprehension questions will appear immediately after each podcast, followed by items on format accessibility and perceived trust, in a fixed order. Expert quality and safety ratings will be gathered from deidentified transcripts. Prerelease accuracy and safety checks by a topic expert and a core safety checklist will be completed before the final recording. For study outcomes, blinded expert quality and safety ratings will be obtained later from deidentified transcripts after the audio is frozen, and these ratings will not influence feedback or edits to the published audio.

1.5.2. Data management

Data will be securely stored on the QuestionPro, Prolific, and SharePoint platforms, which offer built-in tools for data entry, coding, and storage. Data security is ensured through encryption and restricted access.

1.5.3. Data sharing and access

At publication, a deidentified dataset, data dictionary, and analysis code will be shared on the Open Science Framework (OSF) under a Creative Commons Attribution 4.0 International (CC BY 4.0). The results will be posted to ISRCTN (International Standard Randomised Controlled Trial Number), shared on OSF, submitted to a peer-reviewed journal, and summarized in a brief plain language podcast. Participants will receive a summary via Prolific messaging.

1.5.4. Statistical methods

The primary analysis will follow the per-protocol principle, excluding participants who meet any of the prespecified criteria below. A sensitivity intention-to-treat analysis will include all randomized participants with at least one postrandomization outcome measurement:.

- Completed listening and answering all questions in less than 25 minutes total
- Demonstrated straight-line answering patterns (defined as selecting the identical response option for $\geq$80% of Likert-scale items within the trustworthiness or format accessibility assessments, calculated separately for each podcast evaluated)

Comparing per-protocol and intention-to-treat estimates will provide complementary evidence on noninferiority, with the per-protocol analysis as primary, consistent with noninferiority trial conventions.

All exclusions will be documented and reported according to the CONSORT guidelines [16], including the number excluded for each criterion, comparison of exclusion rates between study arms, and baseline characteristics of excluded v included participants.

Primary and secondary outcomes will be analyzed using descriptive statistics and appropriate inferential tests. Since participants evaluate three podcasts and multiple participants assess each podcast, we will employ mixed-effects models to account for the hierarchical data structure. For participant-level outcomes (comprehension, format accessibility ratings, and perceived trustworthiness), models will include a fixed effect for study arm and random intercepts for participants and episodes; where convergence permits, we will consider random slopes for episode. For expert-rated outcomes (quality of information and safety), models will include fixed effects for arm and random intercepts for podcast episode and rater. When suitable, we will adjust for prespecified covariates in sensitivity analyses, including age and education.

For the primary outcome (comprehension), we will assess noninferiority using a two-sided 95% CI. Noninferiority will be established for comprehension if the lower bound of the two-sided 95% CI for the difference between groups (AI-assisted minus human-produced) is greater than the negative noninferiority margin, based on the predefined margins for each outcome. Noninferiority margins for secondary outcomes will be applied in exploratory analyses only; no formal adjustment for multiple comparisons will be made, and inferences for these secondary outcomes will be interpreted cautiously.

If noninferiority is established for the primary comprehension outcome, we will explore potential superiority using the same model structure. Superiority will be concluded if the lower bound of the two-sided 95% CI

for the difference (AI-assisted minus human-generated) is greater than zero; this analysis will be considered secondary and will not be alpha-adjusted.

For the primary mixed-effects models, we will use restricted maximum likelihood estimation under a missing-at-random assumption, which allows inclusion of participants with partially observed outcome data.

Subgroup analyses will investigate the effects of age and educational background on outcomes. Adjusted analyses will be conducted to control for potential confounding variables, including demographic factors. These subgroup analyses are prespecified but interpreted as exploratory. Sensitivity analyses may be conducted to evaluate the impact of missing data, various assumptions, and model specifications on the results.

Data analysts will not be blinded; analyses will be conducted by the lead author with statistical support from the trial team, following the prespecified analysis plan.

A Trial Management Group (PI and co-investigators) oversees day-to-day conduct and data quality. A Data Monitoring Committee will not be established for this trial, as the intervention poses minimal risk and involves evaluating audio podcasts without physical interventions. No serious adverse events are expected. The research team will internally oversee trial conduct to ensure adherence to the protocol and maintain data integrity. No interim analyses are planned for this noninferiority trial.

As this trial involves listening to audio health information only, no physical or pharmacological adverse events are anticipated. Safety monitoring is limited to informational safety (accuracy, completeness, and potential for misunderstanding of health content), assessed through the core safety items in the QUEST framework (Supplementary File 7).

1.6. Ethics

Participants will receive detailed information about the study, including its purpose, procedures, potential risks, and benefits. Informed consent will be obtained from all participants prior to the commencement of the study. The consent process will be carried out online via the QuestionPro/Prolific platforms. Participants' personal information and responses will remain confidential. No personally identifiable information beyond platform-assigned IDs will be collected in the research dataset. Participants will listen to audio files only, and no participant voice recordings will be captured. Data will be stored securely, with access restricted to the research team only. No biological samples or clinical procedures are involved, risks are minimal, and participants may withdraw at any time before submitting their responses without penalty. Any changes to the protocol will be documented and communicated to relevant stakeholders, including the ethics committee, trial registry, and participants, where applicable. The HIET trials, of which this is the second trial, have received ethics approval from the University of Galway Research Ethics Committee (2023.05.011).

2. Discussion

This study aims to compare AI-assisted podcasts with human-produced podcasts in the same audio format for communicating evidence-based health information to the public. We adopt a human-in-the-loop approach that combines AI tools with expert review to create podcasts that are accessible, accurate, and trustworthy using Wondercraft AI within a validated workflow. Cochrane PLSs serve as the common evidence source for both groups, ensuring consistent content across different conditions and levels of complexity. This trial complements our prior work (HIET1), which assessed authorship effects in text, by exploring authorship effects in audio.

Our approach addresses the increasing need for accessible health communication that caters to diverse literacy and learning styles. Plain language summaries are created to simplify complex health information for the public, serving as a shared resource for both podcast formats. However, this process demands substantial time, resources, and expertise. AI-assisted podcasts may enable faster production than entirely human-produced ones while still maintaining clarity and balance under expert supervision. Using a noninferiority framework, we compare AI-assisted and human-produced podcasts based on comprehension, format accessibility, information quality, perceived trustworthiness, and safety. Since both are audio, the format remains the same, and differences can be attributed to authorship effects within audio. A potential strength of podcasts, compared to written formats, is their ability to engage listeners through storytelling and narrative features. This aligns with cognitive research showing that emotional narrative styles improve memory retention compared to expository text, which may support learning in audio formats [17]. The conversational tone and interpersonal dynamics typical of podcast hosts can influence how listeners receive information, affecting listener engagement, trust, and understanding [18].

However, AI-assisted podcasts face challenges, such as ensuring factual accuracy and clarity and avoiding oversimplification or overcomplication. Additionally, it is important to recognize that audio content may not be equally accessible to individuals with hearing impairments. If AI-assisted podcasts are proven to be noninferior to human-produced podcasts, they could serve as a scalable option to increase access to evidence-based health care, particularly for individuals with reading difficulties such as dyslexia or visual impairments. Other challenges of this study include the inability to fully mask authorship, as residual vocal cues may differ between human narration and text-to-speech voices. This may introduce bias if listeners respond differently depending on their perceptions of AI-generated content. In HIET-1, participants correctly identified AI-assisted summaries in only 56.3% of cases (near chance for a three-option question), and 55.4% of human-generated summaries

were misattributed as AI-assisted; yet this perception bias did not diminish trust ratings or comprehension scores. We will collect postassessment authorship perception data in the present trial to quantify any analogous effect in audio [13]. Other limitations include potential selection bias from online-only recruitment and the difficulty of creating perfectly parallel scripts and performances even under a shared brief. Additionally, the English language requirement and online platform may exclude populations who could benefit most from accessible health communication formats. This trial maintains a consistent modality by design and does not test alternative audio styles or entirely AI-generated content without human review. Furthermore, some secondary outcomes (such as expert-rated quality and safety and script readability) are based on a small number of podcasts, so estimates for these measures will carry wide uncertainty and should primarily be used to generate hypotheses for future work. Future research could explore additional options, such as different voice technologies or levels of human oversight.

In summary, this research contributes to the growing field of digital health communication and investigates whether AI can be responsibly used to support the creation of public-facing content in a single audio format. The findings will inform future approaches to producing health research that is both accessible and engaging.

CRediT authorship contribution statement

Isabel O'Byrne: Writing − original draft, Conceptualization. **Johanna Pope:** Writing − review & editing. **Paula Byrne:** Writing − review & editing. **Darren Dahly:** Writing − review & editing. **Conor Judge:** Writing − review & editing. **Martin O'Donnell:** Writing − review & editing. **Finn Krewer:** Writing − review & editing. **Mengqi Li:** Writing − review & editing. **Km Saif-Ur-Rahman:** Writing − review & editing. **Tom Conway:** Writing − review & editing. **James Thomas:** Writing − review & editing. **Geraldine Sheils:** Writing − review & editing. **Nikita Burke:** Writing − review & editing. **Marie O'Byrne:** Writing − review & editing. **Aislinn O'Byrne:** Writing − review & editing. **Declan Devane:** Writing − review & editing, Supervision, Funding acquisition, Conceptualization.

Ethics statement

The HIET trials, of which this is the second trial, have received ethics approval from the University of Galway Research Ethics Committee (2023.05.011).

Declaration of competing interest

DD is the principal investigator for the core grant for iHealthFacts, ie, from the Health Research Board (HRB) and Health Service Executive (HSE) grant no. INFO-2021-001.

Acknowledgments

During the preparation of this work, the lead (IOB) and senior author (DD) used ChatGPT for grammatical and language editing purposes. The authors take full responsibility for all the content of this publication.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2026.112272.

Data availability

No data were used for the research described in the article.

References

[1] Gotlieb R, Praska C, Hendrickson MA, Marmet J, Charpentier V, Hause E, et al. Accuracy in patient understanding of common medical phrases. JAMA Network Open 2022;5(11):e2242972.
[2] Amador FLD, Alves GCG, Santos VRD, Moreira RSL. Use of podcasts for health education: a scoping review. Rev Bras Enferm 2024;77(1):e20230096.
[3] Back DA, von Malotky J, Sostmann K, Hube R, Peters H, Hoff E. Superior gain in knowledge by podcasts versus text-based learning in teaching orthopedics: a randomized controlled trial. J Surg Educ 2017;74(1):154−60.
[4] Barton MJ, Okada M, Todorovic M. Podcasts in health education-insights from a scoping review and survey. Anat Sci Educ 2025;18(12):1388−405.
[5] Banić A, Fidahić M, Šuto J, Roje R, Vuka I, Puljak L, et al. Conclusiveness, linguistic characteristics and readability of cochrane plain language summaries of intervention reviews: a cross-sectional study. BMC Med Res Methodol 2022;22(1):240.
[6] Gainey K, Smith J, McCaffery K, Clifford S, Muscat D. Are plain language summaries published in health journals written according to instructions and health literacy principles? A systematic environmental scan. BMJ Open 2024;14(11):e086464.
[7] Robins B, Delaney T, Maher C, Singh B. Podcasts as a tool for promoting health-related behaviours: a scoping review. Digit Health 2024;10:20552076241288630.
[8] Roth J, Chang A, Ricci B, Hall M, Mehta N. Why not a podcast? Assessing narrative audio and written curricula in obstetrical neurology. J Graduate Med Education 2020;12(1):86−91.
[9] Meden E, Radovan M, Štefanc D. Podcasts and informal learning: exploring knowledge acquisition and retention. Education Sci 2024;14(10):1129.
[10] Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D, Bobes-Bascarán J, Fernández-Leal Á. Human-in-the-loop machine learning: a state of the art. Artif Intell Rev 2023;56(4):3005−54.
[11] Monteith S, Glenn T, Geddes JR, Whybrow PC, Achtyes E, Bauer M. Artificial intelligence and increasing misinformation. Br J Psychiatry 2024;224(2):33−5.
[12] Tam TYC, Sivarajkumar S, Kapoor S, Stolyar AV, Polanska K, McCarthy KR, et al. A framework for human evaluation of large

language models in healthcare derived from literature review. NPJ Digit Med 2024;7(1):258.

[13] Devane D, Pope J, Byrne P, Forde E, O'Byrne I, Woloshin S, et al. Comparison of AI-assisted and human-generated plain language summaries for cochrane reviews: a randomised non-inferiority trial (HIET-1) [registered report - stage II]. J Clin Epidemiol 2026;191:112102.

[14] Chan AW, Boutron I, Hopewell S, Moher D, Schulz KF, Collins GS, et al. SPIRIT 2025 statement: updated guideline for protocols of randomized trials. Nat Med 2025;31(5):1784−92.

[15] Pitcher N, Mitchell D, Hughes C. Template and guidance for writing a cochrane plain language summary. 2022. Available at: https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/guidance-writing-cochrane-plain-language-summary.pdf. Accessed June 2, 2026.

[16] Hopewell S, Chan AW, Collins GS, Hróbjartsson A, Moher D, Schulz KF, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ 2025;389:e081123.

[17] Mar RA, Li J, Nguyen ATP, Ta CP. Memory and comprehension of narrative versus expository texts: a meta-analysis. Psychon Bull Rev 2021;28(3):732−49.

[18] Nadora M. Parasocial Relationships with Podcast Hosts: Portland State University. 2019. Available at: https://pdxscholar.library.pdx.edu/cgi/viewcontent.cgi?article=1960&context=honorstheses. Accessed June 2, 2026.