



## Comparison of AI-assisted and human-generated plain language summaries for Cochrane reviews: protocol for a randomized trial (HIET-1)

Title	Comparison of AI-assisted and human-generated plain language summaries for Cochrane reviews: protocol for a randomized trial (HIET-1)
Author(s)	Devane, Declan;Pope, Johanna;Byrne, Paula;Forde, Evan;Woloshin, Steven;Culloty, Eileen;Dahly, Darren;Hess Elgersma, Ingeborg;Munthe-Kaas, Heather;Judge, Conor;O'Donnell, Martin;Krewer, Finn;Galvin, Sandra;Burke, Nikita;Tierney, Theresa;Saif-Ur-Rahman, KM;Conway, Tom;Thomas, James
Publication Date	2025-07-01
Publisher	Elsevier
Repository DOI	<a href="https://doi.org/10.1016/j.jclinepi.2025.111894">https://doi.org/10.1016/j.jclinepi.2025.111894</a>

REGISTERED REPORT STAGE I

# Comparison of AI-assisted and human-generated plain language summaries for Cochrane reviews: protocol for a randomized trial (HIET-1)

Declan Devane<sup>a,b,c,\*</sup>, Johanna Pope<sup>a,b,c</sup>, Paula Byrne<sup>a</sup>, Evan Forde<sup>d</sup>, Steven Woloshin<sup>e,f</sup>, Eileen Culloty<sup>g</sup>, Darren Dahly<sup>h,i</sup>, Ingeborg Hess Elgersma<sup>j</sup>, Heather Munthe-Kaas<sup>j</sup>, Conor Judge<sup>d</sup>, Martin O'Donnell<sup>d</sup>, Finn Krewer<sup>d</sup>, Sandra Galvin<sup>a,c</sup>, Nikita Burke<sup>b,c</sup>, Theresa Tierney<sup>k</sup>, KM Saif-Ur-Rahman<sup>b,l</sup>, Tom Conway<sup>a,b,c</sup>, James Thomas<sup>m</sup>

<sup>a</sup>HRB-Trials Methodology Research Network, University of Galway, Galway, Ireland

<sup>b</sup>Evidence Synthesis Ireland & Cochrane Ireland, University of Galway, Galway, Ireland

<sup>c</sup>School of Nursing and Midwifery, University of Galway, Galway, Ireland

<sup>d</sup>School of Medicine, University of Galway, Galway, Ireland

<sup>e</sup>Center for Medicine in the Media, Dartmouth Institute for Health Policy and Clinical Practice, Lebanon, NH, USA

<sup>f</sup>Lisa Schwartz Foundation for Truth in Medicine, Norwich, VT, USA

<sup>g</sup>FuJo Institute, School of Communications, Dublin City University, Dublin, Ireland

<sup>h</sup>HRB Clinical Research Facility, University College Cork, Cork, Ireland

<sup>i</sup>School of Public Health, University College Cork, Cork, Ireland

<sup>j</sup>Centre for Epidemic Intervention Research, Norwegian Institute of Public Health, Oslo, Norway

<sup>k</sup>Public Partner, HRB Primary Care Clinical Trials Network Ireland, University of Galway, Galway, Ireland

<sup>l</sup>Centre for Health Research Methods, College of Medicine, Nursing and Health Sciences, University of Galway, Galway, Ireland

<sup>m</sup>EPPI Centre, UCL Social Research Institute, University College London, London, UK

Accepted 25 June 2025; Published online 1 July 2025

## Abstract

Plain language summaries (PLSs) of systematic reviews present complex health evidence in accessible language. Advances in artificial intelligence (AI), particularly large language models, may enhance the generation of PLSs. This protocol describes a randomized, parallel-group, two-armed, noninferiority trial comparing AI-assisted vs human-generated PLSs. Adults aged 18 years or older, proficient in English, will be recruited online via an audience recruitment platform. Participants are randomly assigned (1:1 ratio) to (1) the intervention group: three AI-assisted PLSs based on recent Cochrane reviews; or (2) the control group: three human-generated Cochrane PLSs. The primary outcome is comprehension (aligned with QUEST's Understanding dimension), assessed via a 10-item multiple-choice questionnaire for each summary, structured according to Cochrane PLS template sections. Secondary outcomes are readability, quality of

**Funding:** Declan Devane is the principal investigator for the core grant for iHealthFacts that is from the Health Research Board (HRB) and Health Service Executive (HSE) Grant no. INFO-2021-001, which supports core study activities, including participant recruitment and platform costs. Additional support includes PhD funding for Johanna Pope through the College of Medicine, Nursing and Health Sciences, University of Galway. Paula Byrne was funded by a grant for iHealthFacts that is from the Health Research Board (HSE) and Health Service Executive (HSE) Grant no. INFO-2021-001.

**Ethics statement:** This trial has obtained ethics approval from the University of Galway Research Ethics Committee (2023.05.011). Participants will be provided with detailed information about the study, including study purpose, procedures, risks, and benefits. Informed consent will be obtained

from all participants before the study begins. The consent process will be conducted online via the QuestionPro/Prolific platforms. Participants' personal information and responses will be kept confidential. Data will be stored securely, with restricted access granted only to the research team. Any changes to the protocol will be documented and communicated to relevant stakeholders, including the ethics committee, trial registry, and participants, where applicable.

Trial ID: ISRCTN85699985.

Date registered: 04/02/2025.

Link: <https://www.isrctn.com/ISRCTN85699985>.

\* Corresponding author. School of Nursing and Midwifery, Aras Moyola, University of Galway, Galway, Ireland H91E3YV.

E-mail address: [declan.devane@universityofgalway.ie](mailto:declan.devane@universityofgalway.ie) (D. Devane).

information, safety considerations, and perceived trustworthiness. This study aims to provide insights into integrating AI technologies in health communication. Its findings will inform future practices in disseminating evidence-based health information to the public. © 2025 The Author(s). 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; Plain language summaries; Cochrane reviews; Randomized controlled trial; Health communication

## 1. Background and rationale

Effective communication of health information is crucial for enabling individuals to make informed decisions about their health. Systematic reviews, such as those conducted by Cochrane, synthesize vast amounts of health research to provide evidence-based conclusions. However, the complexity of these reviews often makes them inaccessible to the general public [1]. To bridge this gap, Cochrane provides plain language summaries (PLSs) designed to present findings in a clear and understandable manner [2]. However, challenges remain in ensuring that lay summaries are readable, accurate, and trustworthy for nonexpert audiences [3,4].

Recent advancements in artificial intelligence (AI), particularly large language models (LLMs) like Open AI's ChatGPT, Anthropic's Claude, etc, offer the potential for enhancing the communication of health information. Growing evidence suggests that AI-assisted writing can significantly reduce the time required for producing PLSs compared to human-only approaches. McMinn et al demonstrated that AI-assisted abstract writing reduced production time while maintaining comparable quality metrics [5]. Similarly, recent studies have shown that human-in-the-loop AI approaches have potential to assist in developing health communication materials while addressing the critical challenge of time-intensive summary development [6,7].

Emerging evidence demonstrates significant efficiency gains from AI-assisted health communication. Traditional systematic review development requires 6-18 months from protocol to publication, creating substantial bottlenecks for timely evidence dissemination [8,9]. AI-assisted approaches have shown a potential to reduce development time significantly while maintaining quality standards, with studies demonstrating the potential for substantial time reductions in health communication writing [5,6]. This suggests substantial scalability benefits for organizations producing large volumes of evidence summaries. However, existing research has focused primarily on technical metrics, with limited comprehensive evaluation using frameworks like QUEST that address safety, trustworthiness, and real-world usability—gaps our study directly addresses.

Recent systematic reviews have established frameworks for evaluating health care LLMs, notably the QUEST framework, which provides structured guidance for human

evaluation across the following five key dimensions: quality of information, understanding and reasoning, expression style, safety and harm, and trust and confidence [10]. The QUEST framework was developed through a systematic review of 142 health care LLM evaluation studies, identifying critical gaps in current evaluation practices.

The framework provides structured guidance across five key dimensions specifically designed for health care AI applications. Previous research has established that although automated metrics dominate current LLM evaluation, human assessment remains the gold standard for ensuring safety, reliability, and clinical acceptability—particularly crucial for health communication applications where misinformation can have serious consequences [11,12].

These models can generate human-like text and have the potential to assist in creating summaries that are both accurate and easily comprehensible, which would make the complex results of systematic reviews more accessible to a general audience [13]. However, concerns exist about the accuracy and trustworthiness of AI-generated content, especially in health contexts where misinformation can have serious consequences [13–15]. A combined approach, in which AI language models generate initial summaries, which are then refined and verified by human experts, has therefore been proposed. This human-in-the-loop approach (where human experts are involved in refining and verifying AI outputs) can potentially combine AI's strengths—such as rapid text generation and language proficiency—with human expertise to ensure the summaries are accurate, understandable, and aligned with best practices in health communication.

As AI-generated content becomes increasingly integrated into health care, it is essential to evaluate its effectiveness and reliability in real-world applications [16] using rigorous trials. Such evaluations must assess not only technical accuracy but also safety considerations and potential for harm, particularly in health communication contexts. Such comparisons can provide empirical evidence on how AI-assisted summaries perform relative to traditional human-generated summaries.

Our evaluation approach is guided by the QUEST framework, ensuring a comprehensive assessment across key dimensions critical for health communication.

This study compares two methods of generating PLSs of Cochrane reviews.

**What is new?****Key findings**

- We will compare artificial intelligence (AI)-assisted and human-generated plain language summaries (PLSs) of Cochrane reviews in a randomized trial to establish whether AI assistance yields summaries that are noninferior in terms of comprehension, readability, quality, safety, and trustworthiness.

**What this add to what is known?**

- PLSs are crucial for making complex health evidence accessible to the general public, but creating them requires significant expertise and resources.
- Large language models show promise for enhancing health information communication, but their effectiveness in creating accurate, comprehensible, and trustworthy summaries has not been rigorously evaluated in randomized trials.

**What is the implication and what should change now?**

- If AI-assisted summaries prove noninferior to human-generated ones, this could significantly improve the scalability and efficiency of disseminating evidence-based health information to the public and other stakeholders.
- This study will provide empirical evidence to guide the integration of AI technologies in health communication and inform future practices in making health evidence more accessible.

1. AI-assisted summaries: human experts collaborate with an LLM in a human-in-the-loop model to create summaries intended for a general audience.
2. Standard human-generated summaries: traditional summaries produced solely by human experts following Cochrane's established guidelines.

The evaluation of these summaries will systematically assess all five QUEST dimensions while maintaining focus on our primary outcome of comprehension.

**2. Objective**

This study, which is part of a larger initiative known as the Health Information Effectiveness Trials, is the first in a series of studies evaluating methods for enhancing health information communication. The aim is to evaluate whether integrating AI in generating PLSs enhances the communication of synthesized evidence. Specifically, this study will compare the effectiveness of AI-assisted vs human-generated summaries of Cochrane reviews, testing for

noninferiority in five key dimensions of health communication among the general public, guided by the QUEST framework for health care LLM evaluation [10].

**2.1. Questions**

1. Are AI-assisted summaries noninferior to human-generated summaries in terms of how well participants understand the health information provided? (comprehension)
2. Are AI-assisted summaries noninferior to human-generated summaries in terms of expression style, as measured by the standardized readability metrics? (expression style)
3. Are AI-assisted summaries noninferior to human-generated summaries in terms of quality of information in representing the Population, Intervention, Comparator, Outcomes (PICO) elements and overall findings of the Cochrane reviews? (quality of information)
4. Do participants rate the trustworthiness of AI-assisted summaries as noninferior compared to human-generated summaries? (perceived trustworthiness)
5. Are AI-assisted summaries noninferior to human-generated summaries in terms of safety considerations, including potential for harm, bias, and appropriate presentation of limitations? (safety)

**3. Trial design**

This study is a randomized, parallel-group, two-armed, noninferiority trial designed to compare the effectiveness of AI-assisted Cochrane PLS and standard human-generated Cochrane PLS in a general public audience. The protocol's reporting follows the Standard Protocol Items: Recommendations for Interventional Trials guidelines [17]. The outcome assessment is structured according to the QUEST framework for evaluating health care LLMs [10].

Participants will be randomized in a 1:1 allocation ratio to one of two groups.

1. Intervention group: AI-assisted summaries generated iteratively using a LLM.
2. Comparator group: standard human-generated Cochrane PLS prepared solely by experts in systematic reviews who are independent of the study team.

Both groups will receive summaries based on the same three Cochrane reviews. This ensures that each summary is evaluated in both AI-assisted and human-generated forms, facilitating a direct comparison between the two methods."

**3.1. Blinding**

Due to the nature of the interventions and the public availability of human-generated Cochrane PLSs, blinding

participants to their group allocation is not feasible in this study. Participants might recognize the human-generated summaries if they have prior familiarity with Cochrane reviews or if they search for them online.

Researchers administering the study and analyzing outcome data will also be aware of group assignments due to the nature of the intervention materials and the data collection processes.

## 4. Methods: participants, interventions, and outcomes

### 4.1. Study setting

This trial will be conducted fully online, leveraging digital platforms to recruit participants and deliver the study materials. The Prolific platform (<https://www.prolific.com>) will be used to recruit diverse participants, including individuals from various geographic locations, age groups, and educational backgrounds.

The trial will use Questionpro, an online survey system that will randomly divide participants into two groups: those receiving AI-assisted summaries and those receiving human-generated Cochrane PLS. All participants will complete demographic, baseline, and outcome assessments directly within the online platform, ensuring accessibility and convenience.

### 4.2. Eligibility criteria

#### 4.2.1. Inclusion criteria

1. Participants must be 18 years or older.
2. Participants must be proficient in English, as the study materials—including the intervention, comparator, and assessments—will be provided in English. To ensure adequate comprehension, participants will be asked to self-report their English reading proficiency on a scale from 1 (very poor) to 10 (excellent). Only those who rate their reading proficiency as 7 or higher will be eligible to participate in the study.
3. Participants must have access to the internet and a device capable of completing an online survey (eg, computer, tablet, or smartphone).
4. Participants must provide informed consent before starting the study.

#### 4.2.2. Exclusion criteria

1. Individuals unable to complete the online survey or who fail to meet the minimum participation requirements will be excluded.
2. Responses will be excluded if they meet any of the following criteria:
  - Total completion time <10 minutes (combined reading and question time). Responses will be excluded if the total completion time is < 10 minutes. The threshold is based on the

average length of pilot Cochrane PLSs (672 words; range 265–1324) reported in the 2021 PLS pilot evaluation report [18]. Reading three such summaries ( $\approx 2\,020$  words) at a conservative adult silent-reading speed of 238 words per minute for nonfiction requires about 8½ minutes. Even a very fast, focused participant requires at least another 1½ minutes to answer the 10-item comprehension quiz and demographics, giving a realistic lower-bound engagement time of  $\approx 10$  minutes. We, therefore, treat a total task time <10 minutes as evidence the respondent is unlikely to have read the material attentively.

- Evidence of straight-line answering patterns, defined as selecting the same response multiple times across unrelated items.

### 4.3. Interventions

#### 4.3.1. AI-assisted PLSs

Each participant in the intervention group will receive three health information summaries generated through an iterative process involving human experts collaborating with an AI LLM. The specific LLM will be selected based on available benchmarking of available models at the time of protocol implementation, prioritizing those models that show superior performance in scientific summarization, medical knowledge accuracy, and plain language generation. To minimize potential order effects, the presentation sequence of the three summaries will be randomized for each participant. They will receive the same three summaries, deliberately chosen to represent varying levels of complexity and different health topics. This human-in-the-loop process aims to create a PLS of Cochrane intervention reviews that is clear, structured, accurate, and suitable for a general audience.

Reviews will be selected based on the following criteria.

- Publication date: Cochrane reviews should have been published after the last training date (as best as can be established) of the LLM used to ensure the AI model has no prior exposure to the content.
- Focus on the effectiveness of health care interventions, measuring clear outcomes such as treatment success or failure.
- Relevance to common public health concerns and the ability to generate summaries that a general, nonexpert audience can understand.
- Diversity in complexity levels, assessed using standardized readability metrics.
- Representation of different types of interventions (eg, pharmacological, behavioral, and surgical)

Three Cochrane reviews will be purposefully selected based on these criteria to ensure diversity in health topics and complexity levels. All participants will receive summaries of these same three reviews.

The process of developing the PLS will consist of the following steps, all of which will be piloted.

### 1. Initial prompt development and standardization:

- A comprehensive, draft structured prompt has been developed based on Cochrane PLS guidance [19], enhanced with best practices from AI prompting guidelines and plain language resources [20–23] (see [Supplementary file 1](#) for draft of full prompt)
- The prompt includes detailed specifications for:
  - Core requirements (length, reading level, and audience)
  - Preparation steps
  - Required sections with word counts
  - Language guidelines
  - Formatting requirements
  - Quality checks
- The prompt will be tested on pilot reviews and refined iteratively
- The final standardized prompt ensures adherence to:
  - Cochrane PLS structure and guidelines
  - Eighth-grade reading level requirement
  - Maximum 800–850 word limit
  - Clear specification of certainty of evidence
  - Balanced presentation of benefits and harms

We will provide the AI model with relevant sections of Cochrane reviews. If the review is short, we will include the full text, covering background, methods, results, and conclusions. For longer reviews that exceed the AI's context window, we will prioritize key sections such as the abstract, main results, and conclusions. The background and methods will be summarized concisely, either manually or using the AI itself to generate streamlined versions. Any AI-generated content used as input will be carefully reviewed and verified by human experts to ensure accuracy before being incorporated into the process. This approach ensures critical information is included while adhering to input limitations. The existing Cochrane PLS included in the original Cochrane review will not be provided to the LLM as part of the file input.

### 2. Iterative refinement process:

- Two independent experts will review the initial AI output against the detailed quality checklist provided in the prompt
- Structured feedback will use standardized assessment covering:
  - Content completeness and accuracy
  - Language clarity and accessibility
  - Formatting and structure
  - Readability metrics

- Term consistency
- Adherence to word count limits per section
- A maximum of three refinement rounds will be conducted, with refinement stopping if no substantial improvements (defined as <5% change in readability scores) are observed between rounds.
- Each iteration will be documented and assessed using objective metrics:
  - Readability scores
  - Sentence length averages
  - Active voice percentage
  - Medical-term explanation completeness
  - Section-specific word counts
- As part of process reporting, we will measure the similarity between the AI-generated text and the final AI-assisted text using the Jaccard Similarity Index [24] or other text similarity measures to help demonstrate the extent to which the human expert's input modified the original LLM output.
- Example refinement prompts:
  - "Include information on the side effects of the intervention."
  - "Simplify the language used to describe the study results."
  - "Clarify the sample size and demographic details of the study participants."
  - "Summarize the key findings in bullet points for easier understanding."

### 3. Final expert review and enhancement: The human expert will integrate their own knowledge to ensure the final summary is accurate, comprehensive, and understandable. This involves:

- Adjusting complex health terms to be more accessible to the general public and avoiding medical terminology
- Ensuring all key points from the Cochrane review are covered, and the summary aligns with the Cochrane PLS guidance.
- Ensuring that the PICO elements and findings are accurately represented.
- Checking against QUEST framework criteria for quality and safety, which means, verifying appropriate presentation of uncertainties and limitations

### 4. Public and patient involvement (PPI): A PPI partner will review the summary, focusing on readability, clarity, and accessibility for a general audience. The PPI partner will specifically assess whether medical terms are adequately explained and whether the information would be helpful for decision-making. Their feedback will be incorporated into the final version of the PLS. The PPI partner will use a structured

checklist aligned with the prompt requirements to ensure systematic feedback.

#### 5. Documentation and quality control

- Each stage of the interaction with the AI, including the initial prompt, the follow-up refinements, and the final integrated summary, will be documented and reported. Documentation will include the rationale for major changes between versions. The duration of each human–AI iteration will be automatically time-stamped, and the median (interquartile range) total generation time per summary will be reported descriptively as part of the process evaluation. Because the historical human-generated Cochrane PLSs were produced outside the trial and their development time is not recoverable, these timing data will be presented for transparency only and will not be used for statistical comparison between arms.
- Text similarity metrics will be calculated between versions.
- Changes in readability metrics will be tracked across iterations.
- Final verification against the QUEST framework criteria will be performed.
- Independent review by a second expert will confirm adherence to all quality standards.

#### 4.3.2. Human-generated Cochrane PLSs

Each participant in the comparator group will receive the existing published Cochrane PLS for each review prepared by human experts without AI assistance. Participants will receive the same three summaries based on the identical Cochrane reviews used in the intervention group, ensuring direct comparability between the AI-assisted and human-generated summaries. These summaries adhere to the structure and principles in the Cochrane guidance for writing PLSs [19]. The Cochrane reviews selected will be the same as those used in the intervention group.

1. Each summary is created by systematic review experts following the detailed guidance in the Cochrane Handbook. This includes efforts to:
  - Avoid medical/health jargon and use language easily understandable by a general, nonexpert audience.
  - Adhere to the standard Cochrane PLS template, which includes key headings such as:
    - Plain language summary title
    - Key messages
    - What did we want to find out?

- What did we do?
- What did we find?
- Main results, including narrative statement on confidence (certainty) in the evidence  
What are the limitations of the evidence?
- How up-to-date is this evidence?

2. These PLSs undergo peer review to ensure the quality, consistency, and accuracy of the information presented before publication. The review process ensures that the final summary faithfully represents the findings of the Cochrane review in a clear and structured manner.

#### 4.4. Outcomes

The following outcomes will be assessed, aligned with the QUEST framework for evaluating health care LLMs [10].

##### 4.4.1. Primary

1. Comprehension (aligned with QUEST's understanding dimension)

Participants will complete a standardized 10-item multiple-choice questionnaire for each summary, structured to align with the Cochrane PLS template (see [Supplementary file 2](#)). The questionnaire systematically assesses understanding across five key domains.

1. Understanding of review topic (2 items)
  - Understanding of the health condition/problem
  - Recognition of review importance
2. Review aims and methods (2 items)
  - Comprehension of the main review question
  - Basic understanding of evidence-gathering approach
3. Main results (3 items)
  - Understanding of key benefits
  - Understanding of unwanted effects/harms
  - Grasp of the size of the evidence base
4. Evidence quality and limitations (2 items)
  - Recognition of main limitations
  - Understanding of evidence strength/certainty
5. Currency of evidence (1 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. The questionnaire will be piloted with public participants to ensure clarity.

For the primary outcome of comprehension, a noninferiority margin of 10% will be used. AI-assisted summaries will be considered noninferior if the comprehension score is not more than 10% worse than that of human-generated summaries.

#### 4.4.2. Secondary

##### 2 Readability (aligned with QUEST's expression style dimension)

Readability will be assessed automatically in [readabilityformulas.com](https://readabilityformulas.com), using multiple formulas to measure the ease with which the text can be read and understood [25]. These tests evaluate sentence length, word complexity, and the overall grade level required for comprehension. We will measure and report the following readability measures in [Table 1](#).

Primary readability outcome will be the Flesch–Kincaid Grade Level, with other metrics reported as secondary outcomes. To assess the impact of human intervention, we will compare the readability scores of the initial AI-generated summaries with those of the final AI-assisted summaries.

For the outcome of readability, a noninferiority margin of 1-grade level will be used. AI-assisted summaries will be considered noninferior if their mean Flesch–Kincaid Grade Level is not more than 1-grade level higher than that of human-generated summaries.

##### 3. Quality of Information

Two independent systematic review experts will compare both the AI-assisted and human-generated summaries directly against the original Cochrane review. This prevents the human summary from being treated as an implicit “gold standard” and allows errors in either version to be detected.

The assessment will focus on four main types of errors.

1. Incorrect output (where the LLM generated wrong information).
2. Irrelevant output (where the LLM generated unnecessary or off-topic information).
3. Omissions (where the LLM 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 reader)
- 2 .Moderate quality (minor errors that do not significantly alter understanding)
- 3 .High quality (no errors)

Inter-rater reliability will be calculated using Cohen's Kappa, with a minimum threshold of 0.7 required. A third systematic review expert will arbitrate all disagreements. The assessment process will be piloted and refined with three test summaries before full implementation.

We assume a baseline accuracy rate of 80% in the human-generated summaries (group A) and set a noninferiority margin of 10%. AI-assisted summaries will be considered noninferior if their accuracy rate is not more than 10% lower than that of human-generated summaries.

##### 4 Safety (aligned with QUEST's safety and harm dimension)

Two independent expert raters will evaluate each summary using two safety assessment frameworks:

Core safety criteria (applied to both AI-assisted and human-generated summaries).

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

Additional AI-specific safety criteria (applied only to AI-assisted summaries).

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

Safety will be assessed as present/absent for each criterion, generating a percentage score of safety criteria met. A noninferiority margin of 10% will be used on core safety criteria only. AI-specific safety concerns will be reported descriptively and used to inform process improvements. Any critical safety issues identified during the assessment will be documented and reported separately.

##### 5. Perceived trustworthiness (aligned with QUEST's trust and confidence dimension)

Participants will be asked to assess the trustworthiness of reviews using a 5-point Likert scale (1 = "strongly disagree" to 5 = "strongly agree") based on items adapted from existing scales measuring trust in online health information. Participants will rate their agreement with the following statements.

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

**Table 1.** Readability tests

Readability test	Measures	Aims to measure
Flesch Reading Ease	<ul style="list-style-type: none"> <li>• Sentence length: average number of words per sentence</li> <li>• Word length: average number of syllables per word</li> <li>• Overall text readability: provides a score between 0 and 100, with higher scores indicating easier readability</li> </ul>	How easy the text is to read; higher scores suggest the text is easier to understand
Flesch–Kincaid Grade Level	<ul style="list-style-type: none"> <li>• Sentence length: average number of words per sentence</li> <li>• Word length: average number of syllables per word</li> <li>• Overall text readability: provides a US school grade level, indicating the minimum education level needed to understand the text</li> </ul>	The educational grade level required to understand the text
Gunning Fog Index	<ul style="list-style-type: none"> <li>• Sentence length: average number of words per sentence</li> <li>• Vocabulary difficulty: percentage of complex or polysyllabic words</li> <li>• Overall text readability: estimates the years of formal education needed to understand the text</li> </ul>	The number of years of education needed to understand the text
Automated Readability Index	<ul style="list-style-type: none"> <li>• Sentence length: average number of words per sentence</li> <li>• Word length: average number of characters per word</li> <li>• Overall text readability: provides a US school grade level</li> </ul>	The US grade level required to comprehend the text
Coleman–Liau Index	<ul style="list-style-type: none"> <li>• Sentence length: average number of sentences per 100 words</li> <li>• Word length: average number of letters per word</li> <li>• Overall text readability: provides a US school grade level</li> </ul>	The educational grade level required to understand the text, focusing on characters per word and sentences per 100 words
SMOG Index	<ul style="list-style-type: none"> <li>• Vocabulary difficulty: number of polysyllabic words</li> <li>• Overall text readability: estimates the years of education needed to understand the text</li> </ul>	The number of years of education needed to comprehend the text based on polysyllabic words
Linsear Write Readability Formula	<ul style="list-style-type: none"> <li>• Sentence length: average number of words per sentence</li> <li>• Word length: number of complex words</li> <li>• Overall text readability: provides a grade level score</li> </ul>	The grade level required to understand the text, emphasizing the number of complex words

SMOG, Simple Measure of Gobbledygook.

After completing all assessments for all summaries, participants will be asked two additional questions.

1. Do you think this summary was written by:

- A human expert alone
- An AI system with human expert review
- An AI system alone
- Not sure

2. How much would it matter to you if health information was written by each of the following? Please rate from 1 (does not matter at all) to 5 (matters a great deal):

- A human expert alone
- An AI system with human expert review
- An AI system alone

The primary trustworthiness measure will be calculated as the mean score across all five items (range 1-5), with individual item scores reported descriptively as secondary outcomes. In the human-generated summaries, we expect a mean score of 4.5 out of 5. The composite score has a noninferiority margin of 0.5 points. AI-assisted summaries will be considered noninferior if their mean trustworthiness score is not more than 0.5 points lower than that of human-generated summaries.

To ensure unbiased scoring, the two expert-rated domains—quality of information and safety/harm—will be evaluated by comparing every statement in each summary with the full Cochrane review. This prevents the existing human PLS from being treated as a de-facto “gold standard”: any incorrect, irrelevant, omitted, or outdated content (quality-of-information errors) and any safety

concerns are identified equally in both versions. Participant-rated outcomes (comprehension, perceived trustworthiness) and software-derived readability metrics are assessed directly on the summaries themselves.

#### 4.5. Participant timeline

On day 1, participants will be recruited, screened, and randomly assigned to the intervention (AI-assisted summary) or comparator (human-generated summary) groups. They will receive three summaries (the same for all participants) and complete the comprehension and trustworthiness assessments for each summary on day 1. All data will be collected remotely. Data collection will occur in a single session per participant, with an estimated total participation time of 45-60 minutes, including consent, reading three summaries, completing assessments, and demographic questions.

#### 4.6. Sample size

The sample size calculation for this trial is based on detecting noninferiority in comprehension scores between the two groups. We assume a baseline comprehension score of 80% in the group receiving human-generated summaries (group A) and expect that the comprehension score in the AI-assisted summaries (group B) will be no worse than 10% below this score. We set a noninferiority margin ( $\Delta$ ) of 10%, which means that AI-assisted summaries will be considered noninferior if the comprehension score is not more than 10% lower than that of human-generated summaries.

Sample size calculations indicate that 396 evaluable participants (198 per group) are needed for 80% power to detect noninferiority with a margin of 10% (1 point on the 10-point comprehension scale), assuming a mean score of 8/10 in both groups and using a two-sided 95% confidence interval (CI) approach.

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

Therefore, we will recruit 454 participants (227 per group) to ensure at least 412 evaluable cases after accounting for an anticipated 10% dropout or exclusion rate. This sample size should provide adequate power for both primary and secondary outcomes, although it may be underpowered to detect small differences in readability measures.

Sample estimates were generated using Sealed Envelope's sample size calculator for noninferiority trials [26] and adjusted for the design effect due to clustering.

#### 4.7. Recruitment

Participants will be recruited via an audience recruitment platform (Prolific). They will be compensated according to Prolific's usual compensation standards.

### 5. Methods: assignment of interventions

#### 5.1. Allocation

##### 5.1.1. Sequence generation

The allocation sequence will be handled using the QuestionPro block randomizer, with participants randomly assigned to either the control (human-generated summaries) or intervention (AI-assisted summaries) group on a 1:1 ratio. All participants within each group will receive the same three summaries, which have been purposefully selected as described previously.

##### 5.1.2. Allocation concealment mechanism

QuestionPro's randomization feature will automatically assign participants to either group while ensuring that the information leaflet and demographic and outcome-related questions remain consistent for all participants.

##### 5.1.3. Implementation

The randomization logic, available in the block flow tab within QuestionPro, ensures equal allocation between groups. The platform will fully automate group assignment.

### 6. Methods: data collection, management, and analysis

#### 6.1. Data collection methods

Data collection will be conducted via QuestionPro. Each participant will receive three summaries. After viewing each summary, participants will immediately complete baseline demographics, including health care background (yes/no) to assess potential response differences based on background knowledge, and self-reported outcome assessments (ie, comprehension and trustworthiness). Standardized questionnaires will assess comprehension, and Likert scales will assess trustworthiness. Readability metrics will be calculated automatically in [readabilityformulas.com](https://readabilityformulas.com) [25], and accuracy by the research team, as noted earlier. All data collection is embedded in the survey platform.

#### 6.2. Data quality monitoring

We will monitor data quality through multiple mechanisms.

- Before beginning the main assessment tasks, participants will complete an instructional attention check [27]. This involves a page titled "Introduction to Plain

Language Summaries in Health Research" containing background information about PLS and specific instructions to click the title rather than answer a decoy multiple-choice question.

- Responses will be monitored for data quality with the following criteria:
  - Total completion time <10 minutes (combined reading and question time)
  - Responses showing straight-line answering patterns (same answer selected for all questions)
  - Responses with inconsistent answers to related questions
- Flagged responses will be reviewed independently by two researchers to determine exclusion.

### 6.3. Data management

Data will be securely stored in the QuestionPro and Prolific platforms, which provide built-in data entry, coding, and storage tools. Data security is ensured through encryption and limited access.

### 6.4. Data sharing and access

Study data will be collected anonymously through Prolific. The dataset, including responses, comprehension scores, and quality assessments, will be made available for secondary research through a publicly accessible OSF Project website at time of results journal publication.

### 6.5. Statistical methods

The primary analysis population will exclude participants who meet any of the following prespecified criteria.

- Failed the instructional attention check
- Completed reading and answering the questions in less than 10 minutes total
- Demonstrated straight-line answering patterns (same answer selected for all questions)

A sensitivity analysis will be conducted, including all randomized participants (full intention-to-treat [ITT] population) regardless of these exclusion criteria.

All exclusions will be documented and reported according to Consolidated Standards of Reporting Trials guidelines, including.

- Number excluded for each criterion
- Comparison of exclusion rates between study arms
- Baseline characteristics of excluded vs included participants

Primary and secondary outcomes will be analyzed using descriptive statistics and appropriate inferential tests. Given

that participants evaluate three summaries each, and each summary is evaluated by multiple participants, we will use mixed-effects models to account for the hierarchical data structure. This approach will include random effects for participants and summaries to address within-participant and within-summary variability.

For the primary outcome (comprehension) and the secondary outcomes (accuracy, readability, and perceived trustworthiness), we will conduct one-sided tests for noninferiority using the predefined noninferiority margins for each outcome.

If noninferiority is established for any outcome, we will test for superiority for that outcome 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.

No interim analyses are planned for this noninferiority trial.

#### 6.5.1. Comprehension and accuracy

- We assume baseline rates of 80% in the human-generated summaries and set a noninferiority margin of 10%.
- Noninferiority will be concluded if the upper limit of the two-sided 95% CI for the difference in proportions (AI-assisted minus human-generated) does not exceed 10%.

#### 6.5.2. Readability

- Primary readability outcome will be Flesch–Kincaid Grade Level.
- A noninferiority margin of 1-grade level is set.
- Noninferiority will be concluded if the upper limit of the two-sided 95% CI for the difference in mean grade levels does not exceed 1.

#### 6.5.3. Perceived trustworthiness

- We expect a mean score of 4.5 out of 5 in the human-generated summaries and set a noninferiority margin of 0.5 points.
- Non-inferiority will be concluded if the upper limit of the two-sided 95% CI for the difference in mean scores does not exceed 0.5.

Missing data will be handled using multiple imputation techniques. Both ITT) and per-protocol (PP) analyses will be performed to assess the robustness of the findings, as recommended for noninferiority trials. The ITT analysis will include all participants as originally allocated, whereas the PP analysis will include only those who fully adhered to the study protocol.

For outcomes requiring expert assessment (quality of information and safety), inter-rater reliability will be

calculated using Cohen's Kappa, with a minimum threshold of 0.7 considered acceptable. Disagreements will be resolved through discussion with a third expert rater.

Subgroup analyses will explore the effects of age and educational background and health care background on outcomes. Adjusted analyses will be conducted to control for potential confounding variables such as demographic factors. Sensitivity analyses may be performed to assess the impact of missing data, different assumptions, and model specifications on the results.

A data monitoring committee will not be established for this trial, as the intervention poses minimal risk and involves evaluating written summaries without physical interventions. No serious adverse events are anticipated. Trial conduct will be internally monitored by the research team to ensure protocol adherence and data integrity.

## 7. Discussion

This study seeks to evaluate the effectiveness of integrating AI in generating PLSs of Cochrane reviews for a general audience. By leveraging a human-in-the-loop approach, we combine the potential efficiency and language capabilities of AI models with the expertise of human reviewers, aiming to produce summaries that are accessible, accurate, and trustworthy.

An innovative aspect of our study design is the use of purposefully selected Cochrane reviews to ensure diversity in health topics and complexity levels. This approach enhances the generalizability of our findings by testing the AI-assisted method across various health topics and content complexities. It allows us to assess the consistency and reliability of AI-assisted summaries in different contexts, making our results more applicable to real-world settings.

We anticipate that our findings will provide valuable insights into the potential of AI-assisted summarization in health communication. If AI-assisted summaries are found to be noninferior to human-generated summaries, this could have significant implications for the scalability and efficiency of disseminating evidence-based health information to the public.

However, our study has some limitations. Although we chose reviews published within the prior 3 months, given the rapid pace of LLM development, we cannot be sure if the LLM would have included this review in its training data. In addition, using only three purposefully selected reviews may limit the generalizability of our findings to other health topics not included in the study. We have attempted to mitigate this by selecting reviews that cover diverse topics and complexity levels. Finally, the public availability of the human-generated Cochrane PLSs means participants could potentially recognize these summaries or find them online, which may influence their responses.

In summary, this study aims to contribute to the understanding of how AI can be effectively integrated into health

information communication, potentially enhancing accessibility and engagement for nonexpert audiences.

## CRedit authorship contribution statement

**Declan Devane:** Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Johanna Pope:** Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Paula Byrne:** Writing – review & editing, Writing – original draft, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Evan Forde:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Conceptualization. **Steven Woloshin:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization. **Eileen Culloty:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Data curation, Conceptualization. **Darren Dahly:** Methodology, Investigation, Formal analysis, Data curation. **Ingeborg Hess Elgersma:** Writing – review & editing, Writing – original draft, Methodology, Investigation. **Heather Munthe-Kaas:** Visualization, Methodology, Investigation. **Conor Judge:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Martin O'Donnell:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Finn Krewer:** Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis. **Sandra Galvin:** Writing – review & editing, Writing – original draft, Resources, Methodology, Investigation. **Nikita Burke:** Supervision, Resources, Project administration, Methodology, Investigation, Conceptualization. **Theresa Tierney:** Writing – original draft, Methodology, Investigation. **KM Saif-Ur-Rahman:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Tom Conway:** Writing – original draft, Methodology, Investigation. **James Thomas:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Data curation, Conceptualization.

## Declaration of competing interest

D.D. declares that he holds several publicly funded grants, including those for evidence synthesis capacity building, but he has no conflicts of interest to declare. C.J. reports that he holds a number of publicly funded grants but no conflicts of interest to declare. S.G., since

submitting this manuscript, has taken on a new professional affiliation in the pharmaceutical sector. S.G.'s contributions to this study were made before this affiliation, and S.G. has no financial or commercial interests related to the subject matter of this research. J.H. declares he holds several grants related to evidence synthesis but has no conflicts of interest to declare. There are no competing interests for any other author.

## Supplementary data

Supplementary data related to this article can be found at <https://doi.org/10.1016/j.jclinepi.2025.111894>.

## References

- [1] 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;22:240. <https://doi.org/10.1186/s12874-022-01721-7>.
- [2] Langendam MW, Akl EA, Dahm P, Glasziou P, Guyatt G, Schünemann HJ. Assessing and presenting summaries of evidence in Cochrane Reviews. *Syst Rev* 2013;2:81. <https://doi.org/10.1186/2046-4053-2-81>.
- [3] Yi L, Yang X. Are lay abstracts published in Autism readable enough for the general public? A short report. *Autism* 2023;27:2555–9. <https://doi.org/10.1177/13623613231163083>.
- [4] Al-Hussaini I, Wu A, Mitchell C. Pathology dynamics at Bio-LaySumm: the trade-off between readability, relevance, and factuality in lay summarization. In: Demner-fushman D, Ananiadou S, Cohen K, editors. *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*. Toronto, Canada: Association for Computational Linguistics; 2023:592–601. <https://doi.org/10.18653/v1/2023.bionlp-1.63>.
- [5] McMinn D, Grant T, DeFord-Watts L, Porkess V, Lens M, Rapier C, et al. Using artificial intelligence to expedite and enhance plain language summary abstract writing of scientific content. *JAMIA Open* 2025;8:ooaf023. <https://doi.org/10.1093/jamiaopen/ooaf023>.
- [6] Ovelman C, Kugley S, Gartlehner G, Viswanathan M. The use of a large language model to create plain language summaries of evidence reviews in healthcare: a feasibility study. *Cochrane Evid Synth Methods* 2024;2:e12041. <https://doi.org/10.1002/cesm.12041>.
- [7] Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. Are ChatGPT and large language models “the answer” to bringing us closer to systematic review automation? *Syst Rev* 2023;12:72. <https://doi.org/10.1186/s13643-023-02243-z>.
- [8] Ganann R, Ciliska D, Thomas H. Expediting systematic reviews: methods and implications of rapid reviews. *Implement Sci* 2010;5:56. <https://doi.org/10.1186/1748-5908-5-56>.
- [9] Clark J, Glasziou P, Mar CD, Bannach-Brown A, Stehlik P, Scott AM. A full systematic review was completed in 2 weeks using automation tools: a case study. *J Clin Epidemiol* 2020;121:81–90. <https://doi.org/10.1016/j.jclinepi.2020.01.008>.
- [10] 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;7:1–20. <https://doi.org/10.1038/s41746-024-01258-7>.
- [11] Daraz L, Morrow AS, Ponce OJ, Beuschel B, Farah MH, Katabi A, et al. Can patients trust online health information? A meta-narrative systematic review addressing the quality of health information on the internet. *J Gen Intern Med* 2019;34:1884–91. <https://doi.org/10.1007/s11606-019-05109-0>.
- [12] Wang Y, McKee M, Torbica A, Stuckler D. Systematic literature review on the spread of health-related misinformation on social media. *Soc Sci Med* 2019;240:112552. <https://doi.org/10.1016/j.socscimed.2019.112552>.
- [13] Xie Q, Schenck EJ, Yang HS, Chen Y, Peng Y, Wang F. Faithful AI in medicine: a systematic review with Large Language models and beyond. *medRxiv* 2023;2023:23288752. <https://doi.org/10.1101/2023.04.18.23288752>.
- [14] Dunn AG, Shih I, Ayre J, Spallek H. What generative AI means for trust in health communications. *J Commun Healthc* 2023;16:385–8. <https://doi.org/10.1080/17538068.2023.2277489>.
- [15] Sorich MJ, Menz BD, Hopkins AM. Quality and safety of artificial intelligence generated health information. *BMJ* 2024;384:q596. <https://doi.org/10.1136/bmj.q596>.
- [16] Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019;17:195. <https://doi.org/10.1186/s12916-019-1426-2>.
- [17] Chan A-W, Tetzlaff JM, Altman DG, Laupacis A, Gøtzsche PC, Krleža-Jerić K, et al. SPIRIT 2013 statement: defining standard protocol items for clinical trials. *Ann Intern Med* 2013;158:200–7. <https://doi.org/10.7326/0003-4819-158-3-201302050-00583>.
- [18] Cochrane. Plain Language Summary pilot Project: final evaluation report. Cochrane. 2021. Available at: <https://community.cochrane.org/sites/default/files/uploads/inline-files/PLSPilotFinalEvaluationReportFINAL.pdf>. Accessed February 7, 2025.
- [19] Pitcher N, Mitchell D, Hughes C. *Template and guidance for writing a Cochrane Plain language summary, Version 1*. London: Cochrane; 2022.
- [20] Best practices for prompt engineering with the OpenAI API | OpenAI help center. Available at: <https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api>. Accessed October 28, 2024.
- [21] General tips for designing prompts. 2024. Available at: <https://www.promptingguide.ai/introduction/tips>. Accessed October 28, 2024.
- [22] Writing a plain language (lay) summary of your research findings. In: Health Research Authority. Available at: <https://www.hra.nhs.uk/planning-and-improving-research/best-practice/writing-plain-language-lay-summary-your-research-findings/>. Accessed October 28, 2024.
- [23] Plain English summaries | NIHR. Available at: <https://www.nihr.ac.uk/plain-english-summaries>. Accessed October 28, 2024.
- [24] Jaccard P. Nouvelles Recherches Sur la Distribution Florale. *Bull Soc Vaudoise Sci Nat* 1908;44:223–70. <https://doi.org/10.5169/seals-268384>.
- [25] Scott SB. ReadabilityFormulas.com. Available at: <https://readabilityformulas.com/about-us/>. Accessed September 26, 2024.
- [26] Sealed Envelope | Power calculator for binary outcome non-inferiority trial. 2012. Available at: <https://www.sealedenvelope.com/power/binary-noninferior/>. Accessed October 5, 2024.
- [27] Oppenheimer DM, Meyvis T, Davidenko N. Instructional manipulation checks: detecting satisficing to increase statistical power. *J Exp Soc Psychol* 2009;45:867–72. <https://doi.org/10.1016/j.jesp.2009.03.009>.