ISRCTN85699985 https://doi.org/10.1186/ISRCTN85699985

Which works better - summaries of health research written by people alone or with computer help?

Submission date 12/11/2024	**Recruitment status** Recruiting	[X] Prospectively registered [ ] Protocol
Registration date 04/02/2025	**Overall study status** Ongoing	[ ] Statistical analysis plan [ ] Results
Last Edited 06/06/2025	**Condition category** Other	[ ] Individual participant data [X] Record updated in last year

Plain English summary of protocol

Background and study aims
Effective communication of health information enables informed decision-making. Plain Language Summaries (PLS) of systematic reviews present complex health evidence in accessible language for the general public. Advances in artificial intelligence (AI), particularly large language models like Open AI's ChatGPT, offer potential enhancements in generating these summaries. This protocol outlines a trial comparing the effectiveness of AI-assisted versus human-generated PLS. By evaluating whether AI-assisted summaries are non-inferior to human-generated summaries in these five key dimensions, this study aims to provide insights into the integration of AI technologies in health communication. Findings will inform future practices in disseminating evidence-based health information to the public.

Who can participate?
Adults aged 18 years or older who are proficient in English

What does the study involve?
Participants will be randomly assigned to one of two groups:
1. Intervention Group: Receives three AI-assisted PLS summaries created through a human-in-the-loop process with a large language model based on Cochrane reviews published after the model's training cutoff date.
2. Control Group: Receives three standard human-generated Cochrane PLS matching the summaries in the intervention group.
Each participant will receive three PLS purposefully selected from Cochrane intervention reviews, representing different health conditions and varying levels of evidence certainty. Comprehension, readability, quality of information, safety considerations, and perceived trustworthiness are all assessed.

What are the possible benefits and risks of participating?
There are no big risks in joining this study. Sometimes, people might feel upset if the information they are looking at doesn't match their own ideas or experiences.
If you choose to take part, you will receive compensation as per Prolific's standard payment

guidelines. By taking part, you'll help us understand more about how people find and trust health information, which may shape future health communication and support services. You also might have some fun!

Where is the study run from?
University of Galway (Ireland)

When is the study starting and how long is it expected to run for?
February 2024 to January 2026

Who is funding the study?
College of Medicine, Nursing and Health Sciences, University of Galway (Ireland)

Who is the main contact?
Prof. Declan Devane, declan.devane@universityofgalway.ie

Contact information

Type(s)
Public, Scientific, Principal investigator

Contact name
Prof Declan Devane

ORCID ID
https://orcid.org/0000-0002-9393-7075

Contact details
University of Galway
Galway
Ireland
H91 TK33
+353 91 524411
declan.devane@universityofgalway.ie

Additional identifiers

Clinical Trials Information System (CTIS)
Nil known

ClinicalTrials.gov (NCT)
Nil known

Protocol serial number
Nil known

Study information

Scientific Title

Comparison of AI-assisted and human-generated plain language summaries for Cochrane reviews: protocol for a randomised trial

Acronym
HEIT-1

Study objectives
This study, which is part of a larger initiative known as the Health Information Effectiveness Trials (HIET), 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 plain language summaries enhances the communication of synthesised evidence. Specifically, this study will compare the effectiveness of AI-assisted versus human-generated summaries of Cochrane reviews, testing for non-inferiority in five key dimensions of health communication among the general public, guided by the QUEST framework for healthcare large language model (LLM) evaluation.

Ethics approval required
Ethics approval required

Ethics approval(s)
approved 11/10/2024, University of Galway Research Ethics Committee (University Road, Galway, H91 TK33, Ireland; +353 91 524411; ethics@universityofgalway.ie), ref: 2023.05.011 Amend 2410

Study design
Randomized parallel-group two-armed non-inferiority trial

Primary study design
Interventional

Study type(s)
Other

Health condition(s) or problem(s) studied
Health information

Interventions
This study is a randomised, parallel-group, two-armed, non-inferiority trial designed to compare the effectiveness of AI-assisted Cochrane plain language summarises (PLS) and standard human-generated Cochrane PLS in a general public audience.

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

Arm 1 - AI-assisted summaries:
Participants will receive three health information summaries created through a hybrid approach combining artificial intelligence with human expertise. Each summary is generated by an AI language model and then refined through multiple rounds of expert review. A human expert first feeds the Cochrane review into the AI system using standardised prompts. The resulting summary undergoes up to three rounds of refinement by healthcare experts, with each version checked for accuracy and readability. A patient/public representative then reviews the summary

for clarity and accessibility. The final version incorporates all expert and public feedback to ensure it's both accurate and easy to understand.

Arm 2 - traditional summaries:
Participants will receive three standard Cochrane Plain Language Summaries created using the traditional human-only approach. These summaries are written by systematic review experts following Cochrane's established guidelines, without any AI assistance. They undergo standard peer review before publication and represent the current best practice in creating health information summaries.

Both arms use the same three underlying Cochrane reviews to ensure fair comparison. All summaries, regardless of creation method, aim to explain complex health information at an 8th-grade reading level and include similar sections covering the review's purpose, methods, findings, and conclusions.

The study's objectives are to assess:
1. Comprehension (aligned with QUEST's Understanding dimension)
Participants will complete a standardised 10-item multiple-choice questionnaire for each summary, structured to align with the Cochrane Plain Language Summary template. The questionnaire systematically assesses understanding across five key domains:

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

Each question will use plain language as defined in 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 for clarity.

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

2. Readability (aligned with QUEST's Expression style dimension)
Readability will be assessed automatically in readabilityformulas.com, according to multiple readability formulas that measure the ease with which the text can be read and understood. These tests evaluate sentence length, word complexity, and the overall grade level required for comprehension. We will measure and report the following readability measures:

Readability Test Measures

Flesch Reading Ease

• Sentence length: average number of words per sentence
• Word length: average number of syllables per word
• Overall text readability: provides a score between 0 and 100, with higher scores indicating easier readability

Flesch-Kincaid Grade Level

• Sentence length: average number of words per sentence
• Word length: average number of syllables per word
• Overall text readability: provides a US school grade level, indicating the minimum education level needed to understand the text

Gunning Fog Index

• Sentence length: average number of words per sentence
• Vocabulary difficulty: percentage of complex or polysyllabic words
• Overall text readability: estimates the years of formal education needed to understand the text

Automated Readability Index (ARI)

• Sentence length: average number of words per sentence
• Word length: average number of characters per word
• Overall text readability: provides a US school grade level

Coleman-Liau Index

• Sentence length: average number of sentences per 100 words
• Word length: average number of letters per word
• Overall text readability: provides a US school grade level required to understand the text, focusing on characters per word and sentences per 100 words

SMOG Index

• Vocabulary difficulty: number of polysyllabic words
• Overall text readability: estimates the years of education needed to comprehend the text based on polysyllabic words

Linsear Write Readability Formula

• Sentence length: average number of words per sentence
• Word length: number of complex words
• Overall text readability: provides a grade level score required to understand the text, emphasizing the number of complex words

The 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 non-inferiority margin of 1 grade level will be used. AI-assisted summaries will be considered non-inferior 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
The quality of information will be assessed by two independent systematic review experts using a standardised assessment form. Raters will compare the AI-assisted summaries with the original human-generated Cochrane PLSs. 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. All disagreements will be resolved through arbitration by a third systematic review expert. The assessment process will be piloted and refined with 3 test summaries before full implementation.

We assume a baseline accuracy rate of 80% in the human-generated summaries (Group A) and set a non-inferiority margin of 10%. AI-assisted summaries will be considered non-inferior 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
Expert raters will evaluate each summary for potential risks and biases, focusing on:

• Risk of misinterpretation
• Presence of bias or inappropriate recommendations
• Appropriate presentation of limitations and uncertainties
• Evidence of fabrication or hallucination

Safety will be assessed as present/absent for each criterion, generating a percentage score of safety criteria met. A non-inferiority margin of 10% will be used.

4. 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 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."

After completing all assessments for all summaries, participants will be asked two additional questions:
1. Do you think this summary was written by:

o A human expert
o An AI system with human expert review
o 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):
o A human expert alone
o An AI system with human expert review

For perceived trustworthiness, we expect a mean trustworthiness score of 4.5 out of 5 in the human-generated summaries. We will set a non-inferiority margin of 0.5 points. AI-assisted summaries will be considered non-inferior if their mean trustworthiness score is not more than 0.5 points lower than that of human-generated summaries.

Intervention Type
Other

Primary outcome(s)
1. Comprehension measured using a standardised 10-item multiple-choice questionnaire for each summary at one timepoint
2. Readability measured using the Flesch-Kincaid Grade Level at one timepoint

Key secondary outcome(s))
1. Readability measured using the following at one timepoint:
1.1. Flesch-Kincaid Grade Level
1.2. Gunning Fog Index
1.3. Automated Readability Index (ARI)
1.4. Coleman-Liau Index
1.5. SMOG Index
1.6. Linsear Write Readability Formula

2. Quality of information will be assessed by two independent systematic review experts using a standardised assessment form and rating scale at one timepoint focusing on four main types of errors:
2.1. Incorrect output (where the LLM generated wrong information)
2.2. Irrelevant output (where the LLM generated unnecessary or off-topic information)
2.3. Omissions (where the LLM failed to include key information that should be present)
2.4. Currency errors (where information is outdated or inconsistent with current evidence)

3. Safety measured by expert raters who will evaluate each summary for potential risks and biases at one timepoint, focusing on:
3.1. Risk of misinterpretation
3.2. Presence of bias or inappropriate recommendations
3.3. Appropriate presentation of limitations and uncertainties
3.4. Evidence of fabrication or hallucination

4. Perceived trustworthiness by asking participants to assess the trustworthiness of the reviews using a 5-point Likert scale based on items adapted from existing scales measuring trust in online health information. Participants will rate their agreement with the following statements:
4.1. "I trust the information provided in this summary."
4.2. "This summary is from a reliable source."

4.3. "I am confident in the accuracy of the information in this summary."
4.4. "I believe the source of this summary has expertise in the subject matter."
4.5. "I would use the information from this summary to make health decisions."

After completing all assessments for all summaries, participants will be asked two additional questions:
1. Do you think this summary was written by:
1.1. A human expert
1.2. An AI system with human expert review
1.3. 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):
2.1. A human expert alone
2.2. An AI system with human expert review

Completion date
30/01/2026

Eligibility

Key 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), and 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 (e.g., computer, tablet, or smartphone).
4. Participants must provide informed consent before starting the study.

Participant type(s)
Healthy volunteer, All

Healthy volunteers allowed
No

Age group
Mixed

Lower age limit
18 years

Upper age limit
100 years

Sex
All

Key exclusion criteria

1. Individuals with formal education or professional experience in health-related fields, such as healthcare professionals or academic researchers in health or medicine, will be excluded. This ensures the study targets individuals who may have less familiarity with the topic, thereby maximising the potential impact of the findings.
2. Individuals unable to complete the online survey or who fail to meet the minimum participation requirements will be excluded.
3. Responses will be excluded if they meet any of the following criteria:
3.1. Total completion time <10 minutes per summary (combined reading and question time)
3.2. Evidence of straight-line answering patterns
3.3. Inconsistent answers to related questions

Date of first enrolment
01/09/2025

Date of final enrolment
30/01/2026

Locations

Countries of recruitment
United States of America

Study participating centre
Participants will be recruited via an audience recruitment platform (Prolific)
-
United States of America
-

Sponsor information

Organisation
Ollscoil na Gaillimhe – University of Galway

ROR
https://ror.org/03bea9k73

Funder(s)

Funder type
University/education

Funder Name

College of Medicine, Nursing and Health Sciences, University of Galway

Alternative Name(s)
College of Medicine, Nursing and Health Sciences, National University of Ireland, Galway, College of Medicine Nursing & Health Sciences, NUI Galway - College of Medicine, Nursing and Health Sciences, College of Medicine, Nursing & Health Sciences - NUI Galway

Funding Body Type
Government organisation

Funding Body Subtype
Universities (academic only)

Location
Ireland

Results and Publications

Individual participant data (IPD) sharing plan
The datasets generated during and/or analysed during the current study will be stored in a publically available repository (https://osf.io/srwdk/)

• The type of data stored: Survey responses including comprehension scores, trustworthiness ratings, demographic data
• Dates of availability: Available upon publication of study results
• Whether consent from participants was required and obtained: Informed consent was obtained from all participants, including explicit information about public data sharing
• Comments on data anonymization: All data are fully de-identified and anonymised: no identifiers collected, demographics provided in ranges only, Prolific payment data handled separately

IPD sharing plan summary
Stored in publicly available repository

Study outputs

Output type	Details	Date created	Date added	Peer reviewed?	Patient-facing?
Participant information sheet	version 24	13/11/2024	26/11/2024	No	Yes
Participant information sheet	Participant information sheet	11/11/2025	11/11/2025	No	Yes