Journal of Clinical Epidemiology 191 (2026) 112102

REGISTERED REPORT STAGE II

Comparison of AI-assisted and human-generated plain language summaries for Cochrane reviews: a randomised non-inferiority trial (HIET-1) [Registered Report - stage II]

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

[a]*Health Research Board-Trials Methodology Research Network, University of Galway, Galway, Ireland*
[b]*Evidence Synthesis Ireland & Cochrane Ireland, University of Galway, Galway, Ireland*
[c]*School of Nursing and Midwifery, University of Galway, Galway, Ireland*
[d]*School of Medicine, University of Galway, Galway, Ireland*
[e]*Dartmouth Institute, Lebanon, NH, USA*
[f]*Lisa Schwartz Foundation for Truth in Medicine, Norwich, VT, USA*
[g]*FuJo Institute, School of Communications, Dublin City University, Dublin, Ireland*
[h]*HRB Clinical Research Facility, University College Cork, Cork, Ireland*
[i]*School of Public Health, University College Cork, Cork, Ireland*
[j]*Centre for Epidemic Intervention Research, Norwegian Institute of Public Health, Oslo, Norway*
[k]*Public Partner, Health Research Board-Primary Care Clinical Trials Network Ireland, University of Galway, Galway, Ireland*
[l]*Centre for Health Research Methods, University of Galway, Galway, Ireland*
[m]*EPPI Centre, UCL Social Research Institute, University College London, London, UK*

Accepted 7 December 2025; Published online 15 December 2025

Abstract

Objectives: To compare the comprehension, readability, quality, safety, and trustworthiness of artificial intelligence (AI)-assisted vs human-generated plain language summaries (PLSs) for Cochrane systematic reviews.

Study Design: Randomized, parallel-group, two-arm, noninferiority trial (ISRCTN85699985).

Setting: Online survey platform, September 2025.

Participants: Adults aged 18 years or older with a minimum English reading proficiency of 7 out of 10, recruited via Prolific. Of the 500 individuals screened, 465 were randomized and 453 completed per-protocol analysis.

Interventions: Participants were randomly assigned to three AI-assisted PLSs developed with ChatGPT and human-in-the-loop verification, or to three published human-generated Cochrane PLSs for the same reviews.

Outcomes: Primary: comprehension (10-item questionnaire, noninferiority margin 10%). Secondary: readability quality and safety, trustworthiness, and authorship perception.

Results: Mean comprehension scores were 88.9% (n = 228) in the AI-assisted group and 89.0% (n = 225) in the human-generated group (mean difference −0.03 percentage points, 95% CI: −1.9% to 2.0%); the upper CI bound (2.0 percentage points) did not exceed the +10 percentage-point noninferiority margin, demonstrating noninferiority. Flesch-Kincaid Grade Level showed no significant difference (8.20 vs 8.38, P = .722), although formal noninferiority was missed (upper 95% CI bound 1.72 exceeded the 1.0 grade level margin). AI-assisted summaries scored higher on Flesch Reading Ease (63.33 vs 50.00, P = .008) and lower on the Coleman-Liau Index. All summaries met prespecified quality and safety standards (100% in both groups). Trustworthiness scores were comparable (3.98 vs 3.91,

Funding: This study is supported by the Health Research Board (HRB) Grant nos. TMRN-2021-001 & ESI-2021-001 and the Public Health Agency (Northern Ireland) Grant No. ESI-2021-001 (Principal Investigator: DDe), which supports core study activities, including participant recruitment and platform costs. Additional support includes PhD funding for JP from the College of Medicine, Nursing and Health Sciences at the University of Galway. The funders had no role in study design, data collection, data analysis, interpretation of data, manuscript writing, or the decision to submit for publication.

Trial ID: ISRCTN85699985 (https://www.isrctn.com/ISRCTN85699985); Date registered: 04/02/2025.

* Corresponding author. School of Nursing and Midwifery, University of Galway, Áras Moyola, Galway H91 E3YV, Ireland.

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

https://doi.org/10.1016/j.jclinepi.2025.112102

difference 0.068, 95% CI: −0.043 to 0.179; meeting noninferiority). Participants demonstrated limited ability to distinguish between authorship, correctly identifying AI-assisted summaries in 56.3% of cases and human-generated summaries in 34.7% (≈ chance for a three-option question), with 55.4% of human-generated summaries misattributed as AI-assisted. Exploratory subgroup analysis showed an age interaction (P = .023), though based on a small subgroup (n = 14, 3%).

Conclusion: AI-assisted PLSs with human oversight achieved comprehension levels noninferior to those of human-generated Cochrane summaries, with comparable quality, safety, and trust ratings. AI summaries were largely indistinguishable from those generated by humans. Pretrial verification identified and corrected numerical errors, confirming the need for human oversight. These findings support human-in-the-loop AI workflows for PLS production, though formal evaluation of the time and resource implications is needed to establish efficiency gains over traditional manual methods.

Keywords: Artificial intelligence; Plain language summaries; Cochrane reviews; Health communication; Noninferiority trial; Human-in-the-loop; Comprehension

1. Background and rationale

Evidence synthesis involves structured reviews that gather and analyze multiple sources to summarize findings on a given topic. In health care, formats include systematic reviews, overviews of reviews, qualitative and mixed-methods syntheses, and broader approaches such as scoping, mapping, and evidence gap maps [1].

Systematic reviews employ clear, reproducible methods to answer specific questions about the effects of interventions, diagnostic accuracy, prognosis, or prevalence [1]. Cochrane published 12,150 reviews from 1998 to 2024 [2]. With 5.14 million articles published in 2022, representing a 71% increase since 2010, practitioners and policymakers find it impossible to stay current without synthesized evidence [3,4]. However, the technical complexity and specialized language of these reviews and most evidence syntheses often make them difficult for patients, carers, and the public to understand, even though they rely on this information for decision-making. Studies show that Cochrane plain language summaries (PLSs) have a Simple Measure of Gobbledygook readability index of 14.7, indicating that they require over 14 years of education, despite aiming for a reading level of 11−12 years old [5]. Qualitative research highlights key barriers, including scientific complexity and a disconnect from patients' real experiences [6].

Cochrane has attempted to address this by requiring PLSs for all reviews, following a standard format. These summaries cover the review question, main findings, evidence quality, and implications, asking authors to strike a balance between completeness, brevity, and readability [7].

Producing high-quality PLSs is time-consuming. Systematic reviews typically take approximately 67.3 weeks (15.5 months) from protocol to publication [8], and many require updates or translation; therefore, manual PLS writing can delay dissemination and, in turn, hinder practice change. Automation can shorten timelines; 1 case study reported completing a full systematic review in just 2 weeks using automation tools [9], resulting in approximately a 70% reduction in the end-to-end review workflow (not PLS writing specifically).

Recent advances in large language models (LLMs) offer opportunities to tackle these challenges through automated text generation and summarization capabilities. Van Veen et al [10] found that adapted LLMs outperform clinicians in clinical text summarization, with artificial intelligence (AI) summaries regarded as equal to or better than those of clinicians in 81% of cases. Importantly, expertise in systematic reviewing does not automatically confer proficiency in plain-language writing.

Concerns about AI-generated health content include varying hallucination rates (1.5% in clinical consultation summaries [11] and 61.6% for reference relevancy [12]), as well as the use of black-box approaches that restrict accountability in health-care communication [13]. These issues have led to growing interest in human-in-the-loop methods that combine AI efficiency with human expertise and oversight. Experts review, verify, and refine AI drafts to strike a balance between efficiency and quality assurance [14,15]. The QUEST framework, developed through systematic analysis of 142 health-care LLM evaluation studies, provides criteria across five principles: quality of information, understanding and reasoning, expression, safety, and trust [16].

Currently, evidence from rigorous trials comparing AI-assisted and human-generated health summaries is limited. However, a recent single-blinded, randomized, noninferiority trial comparing ChatGPT-4-generated and human-written Cochrane PLSs reported that ChatGPT-4 was noninferior and, in some cases, superior, with higher scores for informativeness, readability, and detail, as judged by Cochrane editorial group members and laypersons [17]. In addition, Ayers et al [18] found that health-care professionals preferred ChatGPT responses 78.6% of the time in a cross-sectional study of patient questions. Recent exploratory studies suggest that LLMs can achieve high sensitivity in literature screening, with the potential to reduce reviewer workload [19]; however, trials remain few and small.

This study, which is the first trial within a larger initiative known as the Health Information Effectiveness Trials

What is new?

Key findings

- AI summaries achieved noninferior comprehension vs human-generated versions.
- Human verification corrected critical numerical errors in AI drafts.

What this adds to what is known?

- Randomized trial testing public comprehension of AI health summaries.

What is the implication and what should change now?

- Organizations can adopt AI workflows with mandatory human verification.
- Human oversight is essential for safety, not an optional quality enhancement.

(HIET), aimed to compare the effectiveness of AI-assisted vs human-generated PLSs of Cochrane reviews, testing for noninferiority in comprehension, readability, quality, safety, and trustworthiness among the general public, guided by the QUEST framework.

2. Methods

Portions of the Methods section are reproduced from our published Stage 1 Registered Report protocol [20] to ensure methodological transparency and facilitate comparison between the preregistered protocol and final report, consistent with text recycling best practices (https://textrecycling.org/resources/best-practices-for-researchers/).

2.1. Study design

We conducted a randomized, two-arm, noninferiority trial comparing AI-assisted and human-generated PLSs of Cochrane reviews (ISRCTN85699985). The trial is reported in accordance with the CONSORT (CONsolidated Standards Of Reporting Trials) 2025 statement [21]. We chose a noninferiority design because the primary decision problem is whether AI-assisted outputs can safely substitute for the current human standard rather than whether they are strictly better or exactly equivalent. By presenting confidence intervals against a prespecified noninferiority margin, we show whether AI-assisted performance is not unacceptably worse than human production, while still enabling readers to see whether the results are compatible with equal or better performance. Detailed methods are provided in the published protocol [20] (Stage 1 Registered Report). Participants were randomized 1:1 to receive either three AI-assisted PLSs or three human-generated Cochrane PLSs from the same reviews. This study was unblinded, with participants being aware that they were comparing different summary types but not informed which specific summaries they received or their authorship source.

2.2. Participants and setting

Adults aged 18 years or older, proficient in English (self-rated proficiency $\geq$ 7/10), were recruited via Prolific (an online research participant recruitment platform; www.prolific.com) in September 2025. Prolific was used for participant recruitment, and QuestionPro (an online survey platform; www.questionpro.com) was used for survey delivery and data collection. Participants were compensated €12 for their time (approximately 50 minutes). We excluded responses with completion times $<$10 minutes or evidence of straight-line answering patterns, as prespecified in the protocol.

2.3. Interventions

2.3.1. Development of plain language summaries

We selected three Cochrane intervention reviews following prespecified criteria outlined in our protocol [20]: alcohol use disorders [22], antidepressants for low back pain and spine-related leg pain [23], and music-based therapeutic interventions for people with dementia [24]. These reviews were purposefully chosen to ensure diversity in health topics, complexity levels, and intervention types while addressing common public health concerns.

2.3.2. AI-assisted summaries (intervention arm)

We developed a structured prompt by iteratively testing it on three pilot Cochrane reviews not used in the main trial, documenting the changes after each pilot and then locking a final prompt (version 3) before randomization. The prompt specifications included an 400-850-word limit, an eighth grade (11−12 years old) reading level [25], preservation of the Cochrane PLS structure, plain absolute numbers (eg, '64 in 100 vs 74 in 100'), and standardized certainty language ('shows,' 'probably,' 'may,' 'uncertain'). The AI model received the full Cochrane review text as a PDF with the published PLS removed. Only version 3 was used for all trial summaries. The complete prompt text and formatting specifications are provided in the supplementary file to enable replication. We used ChatGPT (model o3-pro-2025-06-10).

Our development process followed the steps outlined in our protocol [20]: initial AI generation using the standardized prompt; expert review using a structured checklist (two independent reviewers; a protocol enhancement) assessing Population, Intervention, Comparator, Outcomes (PICO) completeness, numerical accuracy, balanced presentation of benefits and harms, and consistency with the source; refinements based on findings such as term simplification and

clarification of comparators; public and patient involvement (PPI) partner review focusing on readability and accessibility; and final verification. Based on PPI feedback, 1 terminology clarification was added.

2.3.3. Human-generated plain language summaries (comparator arm)

The currently published Cochrane PLSs for the same three reviews served as comparators. These were developed in accordance with the Cochrane guidelines. These reviews were published before the widespread adoption of AI assistance in medical writing, and their author declarations did not mention AI tool usage.

2.3.4. Pretrial verification of quality and safety

Before the trial, we conducted blinded quality and safety assessments of all six summaries against their source reviews. While the protocol specified quality and safety as outcomes, it did not explicitly state whether failing summaries would be corrected pretrial. We elected to correct identified issues before randomization to ensure participant safety and maintain scientific integrity, ie, all participants deserved access to accurate summaries. The quality and safety assessment checklist is provided in the Supplementary File.

Safety, as used here, refers to the risk that a summary could mislead readers through misinterpretation, bias, inappropriate recommendations, poor handling of uncertainty, or inconsistency with the source review.

Two independent reviewers evaluated all summaries using standardized criteria. Quality was assessed on a three-level scale (poor, moderate, high), while safety was evaluated based on four core criteria: risk of misinterpretation, presence of bias or inappropriate recommendations, appropriate presentation of uncertainty, and consistency with the source review. We required Cohen's $\kappa \geq 0.7$ for interrater agreement, with a third reviewer arbitrating disagreements.

This pretrial verification identified two issues requiring attention. In the AI-assisted music-based interventions, PLS, key messages cited improvements in social interaction without providing corresponding details; we added a clarifying sentence on improvements during group music activities, based on findings from small studies. In the human-generated alcohol PLS, the background used 'car accidents' as an illustrative example, though the source review did not explicitly mention this term. We retained the published Cochrane PLS verbatim as it appropriately illustrated the types of harms described in the review.

2.4. Outcomes

2.4.1. Comprehension (primary)

Participants completed a ten-item multiple-choice questionnaire for each summary aligned with the Cochrane PLS template sections. Each item had 1 correct answer (score 0/1; total 0−10 per summary). The questionnaire assessed understanding across five key domains: understanding of review topic (2), review aims and methods (2), main results (3), evidence quality and limitations (2), and currency of evidence (1). The same questionnaire was used for AI and human versions to ensure comparability. Prior to the main trial, we piloted all items with 40 participants, refining questions that fell outside our preset difficulty thresholds (ie, those with a response rate of either $>$ 90% correct or $<$40% correct). The pilot confirmed feasibility and timing, with participants taking approximately 50 minutes to complete all three summaries and quizzes. Complete questionnaires are in Supplementary File 3.

2.4.2. Secondary outcomes

We assessed perceived trustworthiness using a five-item scale with 5-point Likert responses, examining participants' trust in the information, confidence in the source's reliability and accuracy, perception of the source's expertise, and willingness to use the information for health decisions. Items were averaged to form a composite score. Following completion of all assessments, participants answered questions about perceived authorship, choosing whether they believed each summary was written by 'a human expert alone,' 'an AI system with human expert review,' or indicating they were 'not sure.' Document readability was evaluated using seven standardized metrics calculated through readabilityformulas.com, with Flesch-Kincaid Grade Level serving as our primary readability measure.

The prespecified authorship preference question was not included because pilot testing showed it was difficult to understand and could introduce bias in subsequent responses. We instead focused on postassessment authorship perception, which better captured genuine reactions without priming.

2.5. Sample size

A priori sample size calculations indicated that 396 participants (198 per group) were needed to achieve 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 a two-sided 95% confidence interval. To account for clustering of responses (three summaries per participant), we applied a design effect of 1.04, calculated as $1 + (3\text{-}1) \times 0.02$, where 0.02 was the assumed intraclass correlation coefficient, increasing the required sample to 412 evaluable participants. Therefore, we recruited 454 participants (227 per group) to ensure at least 412 evaluable cases after accounting for an anticipated 10% dropout or exclusion rate.

2.6. Randomization

Randomization sequence was generated using QuestionPro's built-in randomization function. Allocation was

concealed through automated assignment at the time of survey access.

2.7. Statistical analyses

2.7.1. Primary analysis

Our per-protocol analysis excluded participants who failed the attention checks built into the survey, completed the survey in under 10 minutes, or exhibited straight-line responding. We assessed comprehension using linear mixed-effects models (assuming normally distributed random effects and residual errors) fitted by restricted maximum likelihood, with the treatment group as the fixed effect and random intercepts for participants and topics to account for the hierarchical data structure. Each of the 453 participants evaluated 3 summaries, yielding a total of 1359 observations. We concluded noninferiority if the lower bound of the two-sided 95% confidence interval for the difference (AI-assisted minus human-generated) was greater than our prespecified margin of −10 percentage points (equivalent to −1 point on the 10-point scale). For readability outcomes where lower scores indicate better readability, noninferiority required that the upper bound of the two-sided 95% CI did not exceed +1.0 grade level. For quality and safety to be considered noninferior, the difference between groups could not exceed 10 percentage points in the proportion of summaries rated as having poor quality or failing safety criteria.

2.7.2. Sensitivity analyses

The intention-to-treat analysis included all 465 participants who were randomly assigned. Although our protocol specified multiple imputation for missing outcome data, perfect collinearity between total scores and subscores led to model instability, necessitating imputation for only six cases. Therefore, for the six participants who discontinued (1.3%), we imputed missing comprehension scores using treatment group means and bootstrap resampling with 1000 iterations to reflect the uncertainty of the imputation. The six participants who completed the survey in under 10 minutes were included in the intention to treat (ITT) analysis, using their observed data without imputation.

2.7.3. Subgroup analyses

We investigated treatment effect heterogeneity through predefined subgroup analyses based on age and education, and tested interactions using mixed-effects models that included treatment-by-subgroup interaction terms. The predefined subgroup analysis of the health-care background was not feasible because no participants reported health care−related education. Gender and ethnicity were not prespecified as subgroup analyses (ie, for testing interactions) but were included as covariates in adjusted analyses as described below.

2.7.4. Adjusted analyses

As prespecified, we conducted adjusted analyses, including age, education, and ethnicity as covariates, while maintaining the same random-effects structure. These adjusted analyses control for potential confounding but do not test for heterogeneity in treatment effects across demographic subgroups. All models used F-tests with Kenward-Roger approximation for degrees of freedom, and statistical significance was set at $P < .05$ for interaction tests. As these were predefined exploratory analyses, we did not adjust for multiple comparisons. All analyses were performed using R (v.4.3.0).

2.7.5. Data management

We collected all data using QuestionPro's secure platform, which has built-in encryption and restricted access. The anonymized dataset, including participant responses, comprehension scores, and quality assessments, is available in our OSF project repository (https://osf.io/srwdk/).

2.8. Public and patient involvement

This study was developed with PPI. A PPI partner (T.T., co-author) was involved throughout the study and contributed to all key aspects of the research. During protocol development, the PPI partner reviewed and provided input on the study design and outcomes. They also reviewed all six summaries (both AI-assisted and human-generated) for readability and accessibility during the summary development phase. During the manuscript review process, the PPI partner provided feedback on questionnaire clarity and appropriateness, leading to the addition of 1 terminology clarification.

In addition, they reviewed and commented on the study findings and their presentation during the manuscript review stage. The PPI partner was compensated for their time at standard rates. Participants were not involved in the design, conduct, or interpretation of the study beyond completing the assessments.

3. Results

3.1. Participant flow

Of the 500 individuals who accessed the online survey, five declined consent and 30 exited before randomization, leaving 465 individuals randomized to either AI-assisted PLSs ($n = 233$) or human-generated PLSs ($n = 232$).

Following randomization, three in each group discontinued after starting their first summary but before completing all three assigned summaries, leaving 459 completers. Of these 459, six were excluded from the per-protocol analysis because their completion times were less than 10 minutes (range: 5.0−9.9 minutes), as prespecified in the protocol. No participants were excluded for failing attention checks or straight-line responding.

The per-protocol population, therefore, consisted of 453 participants: 228 in the AI-assisted group and 225 in the human-generated group (see Fig. 1).

The intention-to-treat population included all 465 randomized participants. Missing data were handled using multiple imputation as prespecified. In the ITT analysis, outcomes were imputed for the six discontinuities. In the per-protocol analysis, no imputation was required as all 453 adherent participants had complete data. Primary analyses used the per-protocol set, with ITT as a prespecified sensitivity analysis. All analyses employed mixed-effects models with random intercepts for participants and summaries.

3.2. *Human verification process*

All AI-generated summaries underwent structured human review before trial commencement. This verification process revealed specific patterns requiring correction. The most consequential errors were numerical inaccuracies in two summaries: in the alcohol use disorder summary, serious adverse event rates were initially reported in reverse, and in the antidepressants summary, side-effect frequencies were incorrectly reported. These data errors were identified during systematic fact-checking against source review tables and corrected prior to randomization. Additional revisions included adjusting section headings for template consistency, expanding abbreviations on first mention, removing redundant phrasing, and refining uncertainty statements to align with grading of recommendations assessment, development, and evaluation (GRADE) terminology. In the music-based activities summary, a standalone subsection duplicating content from the limitations section was deleted.

Reviewers verified every numerical claim against source data, assessed alignment with certainty language

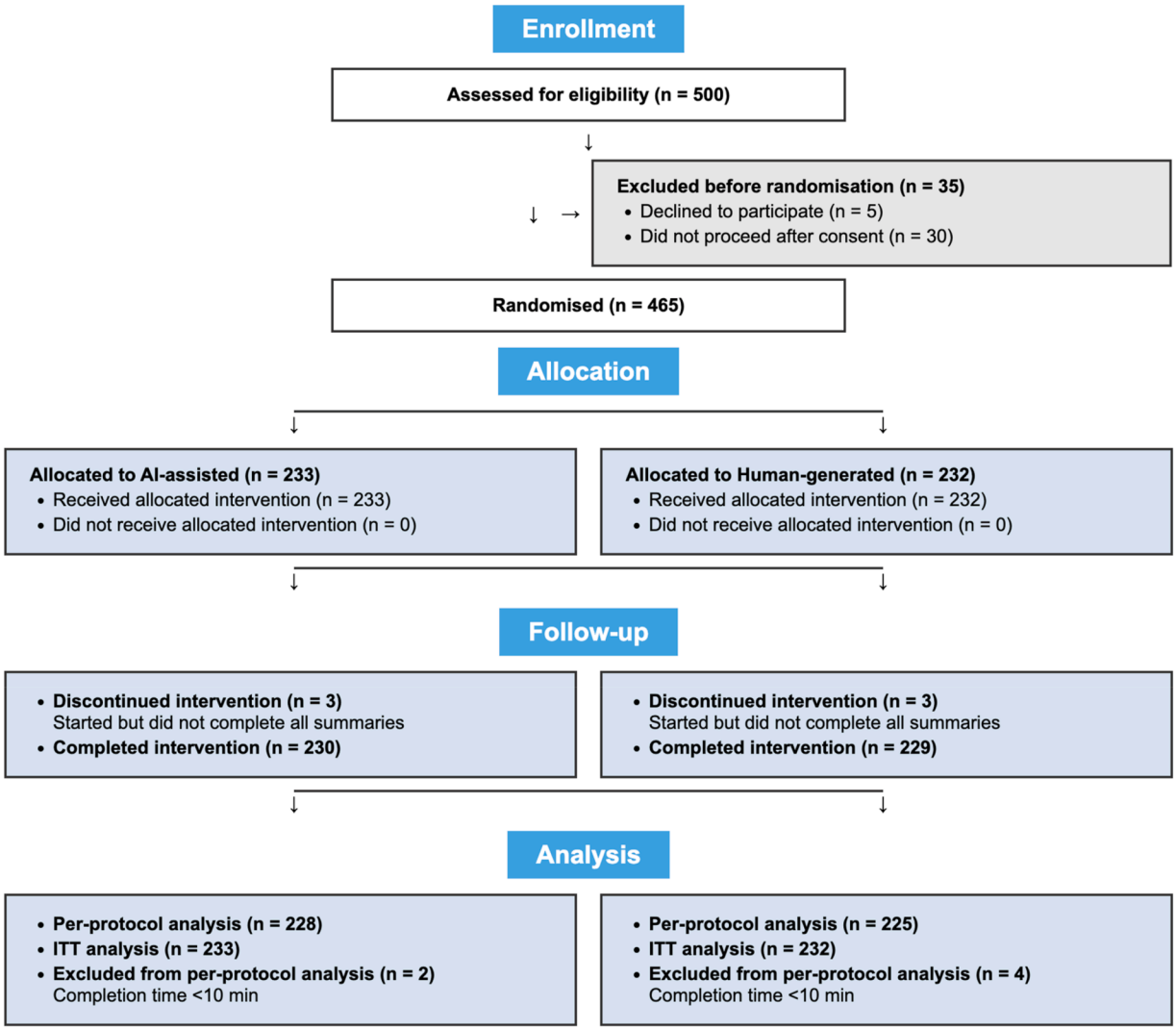

Figure 1. Participant flow.

conventions, evaluated balance in the presentation of benefits and harms, and ensured appropriate acknowledgment of uncertainty.

3.3. Baseline characteristics

Table 1 displays the baseline characteristics of the 453 per-protocol participants. Participants were mainly aged 25–34 years (45%), women (57%), and Black (74%); 85% held a university degree or higher. English proficiency was high (48% native/near-native; 26% excellent). The mean completion time was 46 minutes (SD = 18), and the groups were well balanced. Participants were recruited from 27 countries, with South Africa (72%) as the predominant country of representation, followed by the United

Table 1. Baseline characteristics

Characteristic	Overall (N = 453)[a]	AI-assisted (n = 228)[a]	Human-generated (n = 225)[a]	P value[b]
Age group				0.8
18–24	94 (21%)	42 (18%)	52 (23%)	
25–34	205 (45%)	104 (46%)	101 (45%)	
35–44	68 (15%)	35 (15%)	33 (15%)	
45–54	46 (10%)	25 (11%)	21 (9.3%)	
55–64	26 (5.7%)	15 (6.6%)	11 (4.9%)	
65 y and older	14 (3.1%)	7 (3.1%)	7 (3.1%)	
Gender				0.6
Man (including trans male/trans man)	193 (43%)	100 (44%)	93 (41%)	
Nonbinary	1 (0.2%)	0 (0%)	1 (0.4%)	
Rather not say	1 (0.2%)	0 (0%)	1 (0.4%)	
Woman (including trans female/trans woman)	258 (57%)	128 (56%)	130 (58%)	
Ethnicity				0.11
Black	333 (74%)	170 (75%)	163 (72%)	
White	84 (19%)	46 (20%)	38 (17%)	
Asian	15 (3.3%)	4 (1.8%)	11 (4.9%)	
Mixed	15 (3.3%)	5 (2.2%)	10 (4.4%)	
Prefer not to say	2 (0.4%)	0 (0%)	2 (0.9%)	
Prefer to self-describe	4 (0.9%)	3 (1.3%)	1 (0.4%)	
Country of origin				0.8
South Africa	326 (72%)	165 (72%)	161 (72%)	
United Kingdom	42 (9.3%)	20 (8.8%)	22 (9.8%)	
Poland	13 (2.9%)	7 (3.1%)	6 (2.7%)	
Other (24 countries)	69 (15%)	36 (16%)	33 (15%)	
Education level				0.4
University/college degree	229 (51%)	116 (51%)	113 (50%)	
Advanced degree	155 (34%)	83 (36%)	72 (32%)	
Secondary education (ages 12–18 y)	28 (6.2%)	12 (5.3%)	16 (7.1%)	
Vocational training	29 (6.4%)	14 (6.1%)	15 (6.7%)	
Currently studying	10 (2.2%)	2 (0.9%)	8 (3.6%)	
Prefer not to say	2 (0.4%)	1 (0.4%)	1 (0.4%)	
English reading proficiency				0.3
10 - native/near-native	217 (48%)	110 (48%)	107 (48%)	
9 - Excellent	116 (26%)	64 (28%)	52 (23%)	
8 - Very good	107 (24%)	46 (20%)	61 (27%)	
7 - Good	13 (2.9%)	8 (3.5%)	5 (2.2%)	
Time to complete (minutes)	46 (18)	44 (19)	47 (18)	0.2

AI, artificial intelligence.
[a] n (%); mean (SD).
[b] Pearson's Chi-squared test; Fisher's exact test; Wilcoxon rank sum test.

Kingdom (9.3%) and Poland (2.9%). This geographic distribution was balanced between treatment arms (P = .8).

3.4. Outcomes

Outcomes are summarized in Table 2.

3.4.1. Comprehension (primary)

Participants in the AI-assisted group achieved a mean comprehension score of 88.9% (SE = 0.72%, 95% CI: 87.5%−90.4%), while those in the human-generated group scored 89.0% (SE = 0.72%, 95% CI: 87.5%−90.4%). The estimated difference was −0.03 percentage points (95% CI: −1.9 to 2.0 percentage points) (Table 2).

For the noninferiority assessment, the upper bound of the 95% CI (2.0 percentage points) did not exceed the +10 percentage-point margin, showing that AI-assisted summaries are noninferior to human-generated summaries in terms of comprehension. Superiority was not shown as the CI included zero. Model variance was mainly due to within-participant correlation.

3.4.2. Sensitivity analyses

The ITT population included 465 participants (233 assigned to AI-assisted summaries, 232 to human-generated summaries). Missing outcome data requiring imputation affected 6 participants (1.3%), balanced between treatment groups. In the ITT analysis, the mean proportion of correct responses was 89% (SE = 0.71%) in the AI-assisted group and 89% (SE = 0.72%) in the human-generated group. The estimated mean difference was 0.03 percentage points (95% CI: −1.45 to 1.49 percentage points), which is well within the 10-percentage point noninferiority margin.

Results were consistent between the per-protocol and ITT analyses. The per-protocol analysis (n = 453) showed a mean difference of −0.03 percentage points (95% CI: −1.9 to 2.0 percentage points), while the ITT analysis showed a mean difference of 0.03 percentage points (95% CI: −1.45 to 1.49 percentage points). Both analyses demonstrated noninferiority of AI-assisted summaries, with lower confidence bounds (−1.9 and −1.45 percentage points, respectively) well above the −10-percentage point margin.

3.4.3. Secondary outcomes

3.4.3.1. Readability. The Flesch-Kincaid Grade Level showed no statistically significant difference between AI-assisted (8.20 ± 1.20) and human-generated summaries (8.38 ± 0.72), with a mean difference of −0.18 (95% CI: −2.08 to 1.72, P = .722). The upper bound of the two-sided 95% CI for the difference of 1.72 exceeded the prespecified noninferiority margin of 1.0 grade level (P = .058 for noninferiority), so formal noninferiority was not demonstrated. However, this result should be

Table 2. Outcome summary *(SE, SD)*

Outcome	AI-assisted	Human generated	Difference (95% CI)	Noninferior[a]	*P* value
Comprehension (%)	88.9 (SE 0.72)	89.0 (SE 0.72)	−0.03 (−1.9 to 2.0)	Yes	NS
Readability measures					
Flesch-Kincaid Grade Level[b]	8.20 (SD 1.20)	8.38 (SD 0.72)	−0.18 (−2.08 to 1.72)	No[c]	0.722
Flesch Reading Ease[d]	63.33 (SD 5.51)	50.00 (SD 3.61)	13.33 (8.16 to 18.50)	–	0.008[g]
Gunning Fog Index	9.40 (SD 1.23)	10.43 (SD 0.99)	−1.03 (−3.98 to 1.91)	–	0.270
Automated Readability Index	9.75 (SD 1.54)	8.22 (SD 0.86)	1.53 (−0.46 to 3.52)	–	0.081
Coleman-Liau Index	10.66 (SD 1.02)	11.56 (SD 0.97)	−0.90 (−1.45 to −0.35)	–	0.020[f]
SMOG Index	8.39 (SD 0.89)	7.83 (SD 0.63)	0.57 (−1.47 to 2.61)	–	0.355
Linsear Write Formula	7.57 (SD 1.35)	9.75 (SD 1.69)	−2.18 (−6.48 to 2.12)	–	0.198
Quality and safety					
High quality rating (%)	100 (3/3)	100 (3/3)	0	Yes	–
Safety criteria met (%)	100 (3/3)	100 (3/3)	0	Yes	–
Perceived trustworthiness					
Mean score (1−5 scale)	3.98 (SE 0.085)	3.91 (SE 0.086)	0.068 (−0.043 to 0.179)	Yes	0.230
Authorship perception					
Correct identification (%)	56.3	34.7	21.6	N/A	<0.001[e]

AI, artificial intelligence; N/A, not applicable; –, not tested for noninferiority; NS, not significant; SMOG, Simple Measure of Gobbledygook.

[a] Prespecified noninferiority margins: Comprehension −10%; Flesch-Kincaid 1.0 grade level; Quality/Safety −10%; Trustworthiness −0.5 points.

[b] Primary readability outcome.

[c] One-sided 95% CI upper bound (1.72) exceeded the margin of 1.0 (P = .058); observed difference favored AI.

[d] Higher scores = easier readability; all other readability metrics: lower = easier.

[e] P-value for human group deviation from chance (33.3%).

[f] $P < .05$.

[g] $P < .01$.

interpreted in the context that the observed difference favored AI-assisted summaries (−0.18 grade levels), and the failure to demonstrate noninferiority was due to imprecision from the small number of document pairs ($n = 3$) rather than evidence of inferiority. Among the other readability measures, AI-assisted summaries showed significantly better readability on two metrics, ie, Flesch Reading Ease (mean difference 13.33, 95% CI: 8.16 to 18.50, $P = .008$) and Coleman-Liau Index (mean difference −0.90, 95% CI: −1.45 to −0.35, $P = .020$). Five of seven readability metrics numerically favored AI-assisted summaries. Both AI and human summaries achieved appropriate readability levels for patient education materials (eighth-ninth grade level) (Table 2).

3.4.3.2. Quality and safety. Two independent reviewers assessed the quality and safety of all six summaries. Agreement on quality grades was 83.3% (5/6 summaries). Cohen's κ for quality assessment was 0.00 due to a lack of variance, with ratings concentrated at grade 3. As this fell below the prespecified threshold of 0.7, prevalence- and bias-adjusted κ was calculated as a sensitivity analysis to address the well-documented kappa paradox that occurs when prevalence is extreme, yielding 0.67, indicating substantial agreement. This deviation from the prespecified analysis plan was necessary because the kappa paradox rendered the standard kappa statistic uninterpretable in this context. The observed raw agreement of 83.3% (5/6 summaries) supports the conclusion of substantial agreement. A third reviewer arbitrated the single quality disagreement. After arbitration, all six summaries (three AI-assisted and three human-generated) received high-quality ratings (grade 3). Both AI-assisted and human-generated summaries achieved 100% high-quality ratings (3/3 in each arm, with no difference). All four core safety criteria were met in 100% of summaries in both arms, with no difference observed (Table 2). Both outcomes met the prespecified 10% noninferiority margin. Minor errors were rare, with 1 omission noted in each arm and 1 currency issue in the human arm. No critical safety issues were identified in either group.

3.4.3.3. Perceived trustworthiness. Participants rated AI-assisted summaries with a mean trustworthiness score of 3.98 (SE = 0.085, 95% CI: 3.71−4.25), while human-generated summaries received 3.91 (SE = 0.086, 95% CI: 3.64−4.18) on the 5-point scale (Table 2). The contrast analysis (Human — AI) yielded −0.068 (SE = 0.056, $P = .230$), indicating that AI-assisted summaries scored 0.068 points higher than human-generated summaries. The 95% confidence interval for the difference (AI — Human) was −0.043 to 0.179 points. For noninferiority assessment with a prespecified margin of 0.5 points, the upper bound of the 95% CI (0.179 points) did not exceed the +0.5-point margin, demonstrating noninferiority. Therefore, AI-assisted summaries were shown to be noninferior to human-generated summaries in terms of perceived trustworthiness. Although the point estimates numerically favored AI-assisted summaries, superiority was not established as the CI included zero ($P = .230$).

3.4.3.4. Authorship perception. Participants had difficulty discriminating between authorship methods. The AI-assisted group correctly identified the authors of their summaries in 56.3% of cases, while the human-generated group achieved 34.7% accuracy. For human-generated summaries, 34.7% accuracy is close to the chance level (33.3%) for a three-option choice. We observed a bias toward assuming AI involvement (≈55−56%) regardless of actual authorship (Table 2). This bias was especially pronounced for human-generated summaries, where 55.4% were misattributed to AI assistance, compared to only 34.7% that were correctly identified as human-authored. The distribution of guesses did not differ significantly between groups ($\chi^2 = 1.31$, $P = .521$), implying that the summaries themselves, rather than preconceptions, influenced attribution patterns.

3.4.3.5. Subgroup analyses. Prespecified subgroup analyses revealed a statistically significant interaction between treatment and age ($P = .023$). Participants aged 65 years and older showed significantly better comprehension with AI-assisted summaries (difference = 14.8%, 95% CI: 3.5% to 26.0%, $P = .010$), while treatment effects in younger age groups were minimal and nonsignificant. No treatment-by-education interaction was observed ($P = .256$) (Table 3). In prespecified adjusted analyses including age, gender, education, and ethnicity as covariates, the treatment effect remained consistent with the primary analysis (difference = −0.10 percentage points, 95% CI: −2.07 to 1.86 percentage points), confirming noninferiority after accounting for demographic characteristics.

4. Discussion

Across three Cochrane intervention topics, AI-assisted PLSs were noninferior in comprehension, with confidence intervals within the prespecified 10% margin. Interestingly, absolute comprehension was high in both arms, demonstrating noninferiority against a strong human-written benchmark. Readability was comparable, with five out of seven indices numerically favoring AI-assisted summaries, although noninferiority for the primary Flesch-Kincaid Grade outcome was not achieved because the one-sided upper bound slightly exceeded the 1.0 grade-level margin. Checks on quality and safety revealed no critical issues, and perceived trustworthiness was similar.

Our findings support evidence that LLMs can generate clear health content that matches human-written materials

Table 3. Prespecified subgroup analyses for Comprehension

Subgroup	*n*	AI-assisted mean (SE)	Human-generated mean (SE)	Difference (%)	95% CI	*P* value	Interaction P
Age Group							**0.023***
18–24	94	90.1 (1.7)	87.4 (1.5)	2.6	(−1.8, 7.0)	0.234	
25–34	205	88.2 (1.1)	89.6 (1.1)	−1.4	(−4.3, 1.6)	0.354	
35–44	68	88.1 (1.8)	90.8 (1.9)	−2.7	(−7.8, 2.4)	0.296	
45–54	46	87.3 (2.1)	90.6 (2.3)	−3.3	(−9.5, 2.9)	0.297	
55–64	26	91.4 (2.7)	86.7 (3.1)	4.7	(−2.7, 12.0)	0.215	
≥65	14	97.1 (4.0)	82.4 (4.0)	14.8	(3.5, 26.0)	0.010	
Education				**NS**			**0.256***

* $p < 0.05$. Difference = AI-assisted minus human-generated. Gender and ethnicity were not prespecified as subgroup analyses but were included as covariates in prespecified adjusted analyses (results reported in text).

in readability and perceived informativeness when using a human-in-the-loop approach. In a blinded, randomized noninferiority trial of PLSs, Ágústsdóttir et al [17] found that ChatGPT-4o summaries scored higher on informativeness, detail, and clarity, and that assessors often could not distinguish between authorship and preferred the AI versions. Our research expands on this by directly testing understanding in the public sample with structured questions aligned with the PLS template and by incorporating human-in-the-loop checks focused on accuracy, the balance of benefits and harms, and the presentation of uncertainty. Collectively, these trials suggest that AI-powered workflows can achieve or surpass communication standards that have been difficult to attain at scale using solely manual methods.

Three key observations warrant emphasis. First, achieving comprehension comparable to or better than matched readability demonstrates that a standardized AI-assisted process/workflow can produce public-facing summaries without compromising understanding. The process/workflow features likely play a role: constrained prompts, fixed certainty language, absolute numbers, side-by-side presentation of benefits and harms, and expert verification steps. Second, our finding that readers often presumed AI involvement without any negative impact on trust suggests that transparent disclosure is advisable, even though trust did not decline when readers assumed AI involvement. Fourth, a prespecified exploratory subgroup analysis revealed that participants aged ≥65 years showed higher comprehension with AI-assisted summaries (difference +14.8 percentage points, 95% CI: 3.5% to 26.0%), though this was based on only 14 participants and warrants cautious interpretation (see limitations).

Readability metrics were prespecified as secondary outcomes in the protocol and remain important for implementation, equity, and scalability. The Flesch-Kincaid noninferiority check failed to meet its margin in an underpowered comparison of three document pairs and is sensitive to sentence-splitting and word-length artifacts. Both arms sat within the recommended ranges for patient materials, and other indices, such as Flesch Reading Ease and Coleman-Liau, favored the AI-assisted texts. Taken together with the primary comprehension finding, we view readability as practically equivalent across arms, with a slight tendency toward simpler prose in the AI-assisted summaries.

Pretrial verification and a blinded quality and safety review found no critical concerns but identified and corrected specific numeric errors (eg, reversed serious adverse event rates and side-effect frequencies) before randomization, and all summaries then met the prespecified safety criteria. This demonstrates that human verification is essential and that accuracy risks from generative models can be mitigated when AI output is constrained and audited within a human-in-the-loop process. Embedding structured checks mapped to the QUEST domains both facilitated a balanced presentation, appropriate use of uncertainty language, and the avoidance of overreach beyond the source and enabled detection and correction of these issues in advance of participant exposure.

Participants often attributed authorship to AI regardless of the true source, with a systematic 55%–56% baseline assumption of AI involvement that may reflect increasing public awareness and expectations about AI in health information. Participants demonstrated limited ability to distinguish between authorship methods, with only 56.3% correctly identifying AI-assisted summaries and 34.7% accuracy for human-generated summaries (approximately chance for a three-option question). Despite human-generated Cochrane summaries adhering to established guidelines, 55.4% were misattributed to AI assistance. This challenges assumptions about a distinctive 'AI writing style. Ágústsdóttir et al's trial similarly reported low accuracy in detecting authorship. Interestingly, this bias toward assuming AI involvement did not diminish trust ratings, and AI-assisted summaries received slightly higher scores. The comparable comprehension scores and trust ratings, coupled with poor authorship discrimination, indicate that well-executed AI-assisted summaries are largely indistinguishable from those produced by humans. This indicates that public expectations for health communication may be evolving in response to AI capabilities, and that disclosing

AI involvement is both an ethical requirement and a practical necessity at present (particularly given the numeric inaccuracies we identified and corrected during pretrial verification), as readers already assume AI participation without it damaging trust or comprehension. These findings support consistent disclosure standards, provenance metadata, and brief statements outlining the role of AI and human review. Such practices can maintain transparency while avoiding unnecessary skepticism if AI use is discovered rather than explicitly declared.

Our study's strengths include preregistration, a randomized design with both per-protocol and intention-to-treat analyses, direct testing of comprehension aligned with the PLS template, and explicit safety and quality adjudication with high agreement after prevalence- and bias-adjusted κ adjustment. The study also involved public and patient input to improve accessibility features and utilized a reproducible prompt with fixed certainty wording and numeric framing.

Our study also has limitations. We tested only three review topics and a single AI system, which limits generalizability across clinical areas and AI model groups. Rapid evolution in LLM architectures means our findings may not directly transfer to other models (eg, Claude, Gemini) or future ChatGPT versions. However, the human-in-the-loop verification framework (structured prompts, mandatory fact-checking, balanced presentation of benefits and harms, and expert review) should be transferable across models. The key finding is that even when AI-assisted summaries achieve noninferiority, empirical verification with each specific model remains essential, as demonstrated by the numerical errors we identified and corrected during pretrial verification. Readability analyses were based on a small number of document pairs and using formulae that do not fully account for layout, typography, or cognitive load. We did not systematically quantify the time or cognitive effort required for the human verification process, which represents an important limitation. While the AI-generated drafts provided a foundation that required targeted corrections rather than complete rewriting, the burden of human oversight should not be underestimated. Future research should include formal time-motion studies comparing AI-assisted workflows with traditional manual writing, as well as process evaluations documenting error types and correction frequencies to inform scalable implementation. Participants were generally highly educated, reflecting common online panel demographics; however, this limitation restricts conclusions to populations with lower literacy levels.

In addition, our sample showed unexpected geographic concentration, with 72% of participants from South Africa, likely reflecting Prolific's available participant pool during the recruitment period. While this geographic clustering does not threaten internal validity due to successful randomization across arms, it may limit the generalizability of findings to other populations, particularly those from different cultural and linguistic backgrounds, where English may be processed differently. As with any online research platform, there is an emerging concern that participants may use AI assistance to complete surveys [26], which could affect response validity. However, our use of attention checks, exclusion of implausibly fast completion times, and evaluation of straight-line responding helps mitigate this risk. The finding for older adults was based on a small subgroup ($n = 14$), representing only 3% of the sample, and although the age interaction was prespecified, it was exploratory, lacked adjustment for multiple comparisons, and requires replication in adequately powered samples. We assessed immediate comprehension rather than behavioral intentions or decision quality in real-world settings.

For organizations producing PLSs at scale, these findings support piloting an AI-assisted workflow with human oversight. Key design elements include standardized prompts, required formats for absolute risk, harmonized certainty language, side-by-side presentation of benefits and harms, and structured expert verification. Provenance labeling and short disclosures about AI use should be adopted to meet transparency expectations without undermining trust. Implementation should be accompanied by monitoring of QUEST-aligned indicators, error-reporting pathways, and periodic audits against source reviews. Adoption should follow the Responsible use of AI in evidence SynthEsis (RAISE) guidance for responsible AI in evidence synthesis [27], including transparent provenance, documented human oversight, routine safety checks, and tool selection and evaluation against published criteria.

Future trials should broaden the scope of clinical topics, incorporate non-English languages and translated PLSs, and evaluate multiple models and prompting strategies. Outcomes should extend beyond immediate comprehension to include retention, decisional conflict, and behavioral change, with a particular focus on groups at risk of exclusion, such as those with limited literacy. Formal cost-effectiveness trials evaluating the time and resource implications of AI-assisted vs traditional manual PLS production are needed, alongside process evaluations that systematically document the time required for human verification, the types and frequencies of errors requiring correction, and the cognitive effort involved in oversight. Such evaluations should also consider living-review settings where PLSs need regular updates. The finding regarding age, indicating that AI summaries might especially benefit older adults, warrants targeted investigation. This should involve sufficiently powered age groups, usability testing tailored for older users, and examination of specific design features that enhance comprehension within this demographic.

5. Conclusion

In a randomized noninferiority trial with public participants, AI-assisted PLSs were noninferior to human-

generated Cochrane summaries on comprehension, with comparable readability, quality, safety, and trustworthiness. Human verification identified and corrected numerical errors in two of three AI-generated summaries, including reversed serious adverse event rates and incorrect side-effect frequencies, underscoring the need for human oversight for safety. Readers struggled to distinguish authorship, with a systematic bias toward assuming AI involvement. Taken together with independent evidence that ChatGPT-4o can produce Cochrane PLSs at least as effective as human-written versions, these results support adopting a human-in-the-loop AI workflow for PLS production, requiring clear guardrails, transparent AI-use disclosures, human verification of all claims, and ongoing evaluation focused on comprehension, safety, and equity.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the lead author (DDe) used ChatGPT (OpenAI) for two purposes: (1) grammar and language editing of the manuscript text, and (2) generation of R code for statistical analyses. For manuscript editing, DD used ChatGPT for grammar and language refinement, then formally reviewed all content for accuracy and made necessary edits. For statistical analyses, DDe generated R code (version 4.3.0) with the assistance of ChatGPT and executed it in RStudio. The code was iteratively refined through systematic error-checking, with all results validated by examining outputs for consistency with the study design and cross-verified against the prespecified statistical analysis plan. The authors take full responsibility for the accuracy and integrity of all content in this publication.

CRediT authorship contribution statement

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

Declaration of competing interest

D.D. declares that he holds several publicly funded grants, including those for evidence synthesis and randomized trials, but he has no conflicts of interest to declare. C.J. holds several publicly funded grants but has no conflicts of interest to declare. S.G. is employed in the pharmaceutical sector. J.T. 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.

Acknowledgments

We thank all those who participated in the HIET-1 trial.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2025.112102.

References

[1] Nussbaumer-Streit B, Booth A, Garritty C, Hamel C, Munn Z, Tricco AC, et al. Overview of evidence synthesis types and modes. J Clin Epidemiol 2025;187:111970. https://doi.org/10.1016/j.jclinepi.2025.111970.

[2] Sharifan A. Analysis of Cochrane systematic reviews: a comprehensive study of impact and influence from 1998 to 2024. Cochrane Evid Synth Methods 2024;2:e70010. https://doi.org/10.1002/cesm.70010.

[3] National Center for Science and Engineering (NCES). Alexandria, VA: Publications Output: U.S. Trends and International Comparisons; 2023.

[4] Curcic D. Number of academic papers published per year − WordsRated 2023. Available at: https://wordsrated.com/number-of-academic-papers-published-per-year/. Accessed September 22, 2025.
[5] Karačić J, Dondio P, Buljan I, Hren D, Marušić A. Languages for different health information readers: multitrait-multimethod content analysis of Cochrane systematic reviews textual summary formats. BMC Med Res Methodol 2019;19:75. https://doi.org/10.1186/s12874-019-0716-x.
[6] Gierisch JM, Hughes JM, Williams JW, Gordon AM, Goldstein KM. Qualitative exploration of engaging patients as Advisors in a Program of evidence synthesis. Med Care 2019;57:S246−52. https://doi.org/10.1097/MLR.0000000000001174.
[7] Cochrane. Plain Language Summary Pilot Project: Final evaluation report. London, UK: Cochrane; 2021.
[8] The impact of systematic review automation tools on methodological quality and time taken to complete systematic review tasks: case Study. JMIR Med Educ 2021;7:e24418. https://doi.org/10.2196/24418.
[9] Clark J, Glasziou P, Del Mar C, 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] Van Veen D, Van Uden C, Blankemeier L, Delbrouck J-B, Aali A, Bluethgen C, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 2024;30:1134−42. https://doi.org/10.1038/s41591-024-02855-5.
[11] Asgari E, Montaña-Brown N, Dubois M, Khalil S, Balloch J, Yeung JA, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. Npj Digit Med 2025;8:274. https://doi.org/10.1038/s41746-025-01670-7.
[12] Aljamaan F, Temsah M-H, Altamimi I, Al-Eyadhy A, Jamal A, Alhasan K, et al. Reference hallucination Score for medical artificial intelligence chatbots: development and usability Study. JMIR Med Inform 2024;12:e54345. https://doi.org/10.2196/54345.
[13] Nazer LH, Zatarah R, Waldrip S, Ke JXC, Moukheiber M, Khanna AK, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS Digit Health 2023;2:e0000278. https://doi.org/10.1371/journal.pdig.0000278.
[14] Bakken S. AI in health: keeping the human in the loop. J Am Med Inform Assoc 2023;30:1225−6. https://doi.org/10.1093/jamia/ocad091.
[15] Cohen IG, Babic B, Gerke S, Xia Q, Evgeniou T, Wertenbroch K. How AI can learn from the law: putting humans in the loop only on appeal. Npj Digit Med 2023;6:160. https://doi.org/10.1038/s41746-023-00906-8.
[16] 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.
[17] Ágústsdóttir DH, Rosenberg J, Baker JJ. ChatGPT-4o compared with human researchers in writing plain-language summaries for cochrane reviews: a blinded, randomized non-inferiority controlled trial. Cochrane Evid Synth Methods 2025;3:e70037. https://doi.org/10.1002/cesm.70037.
[18] Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183:589−96. https://doi.org/10.1001/jamainternmed.2023.1838.
[19] Dennstädt F, Zink J, Putora PM, Hastings J, Cihoric N. Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain. Syst Rev 2024;13:158. https://doi.org/10.1186/s13643-024-02575-4.
[20] Devane D, Pope J, Byrne P, Forde E, Woloshin S, Culloty E, et al. Comparison of AI-assisted and human-generated plain language summaries for Cochrane reviews: protocol for a randomized trial (HIET-1). J Clin Epidemiol 2025;185:111894. https://doi.org/10.1016/j.jclinepi.2025.111894.
[21] Hopewell S, Chan A-W, Collins GS, Hróbjartsson A, Moher D, Schulz KF, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ 2025;389:e081123. https://doi.org/10.1136/bmj-2024-081123.
[22] Minozzi S, Rosa GRML, Salis F, Camposeragna A, Saulle R, Leggio L, et al. Combined pharmacological and psychosocial interventions for alcohol use disorder. Cochrane Database Syst Rev 2025;3:CD015673.
[23] Ferraro MC, Urquhart DM, Ferreira GE, Wewege MA, Shaheed CA, Traeger AC, et al. Antidepressants for low back pain and spine-related leg pain. Cochrane Database Syst Rev 2025;3:CD001703.
[24] Steen JT, Wouden JC, Methley AM, Smaling HJA, Vink AC, Bruinsma MS. Music-based therapeutic interventions for people with dementia. Cochrane Database Syst Rev 2025;3:CD003477.
[25] Rooney MK, Santiago G, Perni S, Horowitz DP, McCall AR, Einstein AJ, et al. Readability of patient education materials from high-impact medical journals: a 20-Year analysis. J Patient Exp 2021;8:2374373521998847. https://doi.org/10.1177/2374373521998847.
[26] Zhang S, Xu J, Alvero A. Generative AI Meets Open-Ended Survey responses: research participant use of AI and homogenization. Sociological Methods Res 2025;54:1197−242. https://doi.org/10.1177/00491241251327130.
[27] Thomas J, Flemyng E, Noel-Storr A, et al. Responsible AI in Evidence Synthesis (RAISE): guidance and recommendations (version 2). Washington DC: Center for Open Science; 2025.