Artificial intelligence-generated plain-language summaries from Systematic Reviews for improving general public healthcare participation using the best evidence

Experience generating AI-powered plain language summaries from Cochrane SR abstracts. Readability, content, and the views of the authors

Background

Plain-language summaries (PLSs) make technical information more accessible and understandable to the general public.

They overcome the barrier imposed by technical language to facilitate consumers' and patients' health literacy and active participation in their healthcare.

We aimed to evaluate the feasibility of generating PLSs with large language models (LLM).

Methods

- We used the official Cochrane Guidance "Template and guidance for writing a Cochrane Plain Language Summary: Nicole Pitcher, Denise Mitchell, and Carolyn Hughes Version 1 January 2022." to structure Plain Language Summaries (PLSs)

- We instructed ChatGPT4o to generate PLSs from each Cochrane abstract of 800 words using this Cochrane Guidance
- 100 AI-generated PLSs were distributed among Cochrane Authors from SRs published in 2022, before the widespread use of AI
- All authors were enquired by email about several ítems.

We assesed in the 100 PLSs:

-**Readability** was based on the Flesch Reading Ease score:

1 (Extremely easy) to 7 (Extremely difficult) (see Box)

We also tested:

-**Text score** as provided by Grammarly (free version) (0-100) higher=better quality

-**Use of Rare words** (%) lower=better

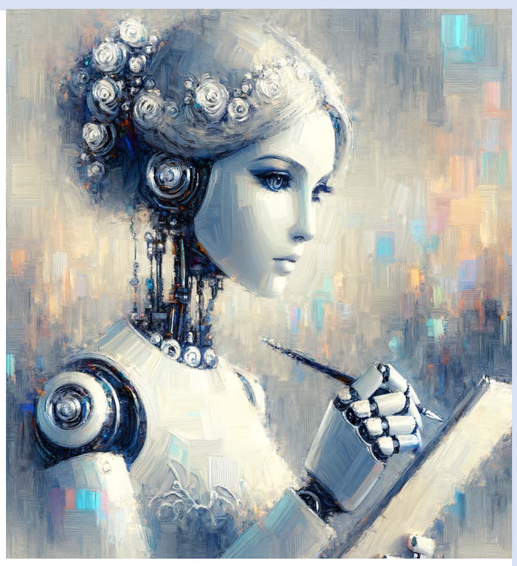
Corresponding School Level	Reading Difficulty
1	5th grade
2	6th grade
3	7th grade
4	8th and 9th grade
5	10th to 12th grade
6	College
7	College graduate and Professional

Additionally, we asked authors:

•Were the dimensions of the PICO questions described correctly? (Yes/No)

•Were the Results and Key Messages described correctly? (Yes/No)

•Would you use this PLS to disseminate your work among laypersons with minimal edits? (Yes/No)



Results

Table 1 shows the performance differences

Additionally, 10 Authors replied:

- 6 considered that the PICO questions were right
- 5 made criticisms on the key messages
- 2 proposed new versions of their PLS from the one generated by AI
- 8 would use them for dissemination with minimal changes

Average readability according to authors: 6 (College)

Example of AI-PLS:

Title: Does screening for lung cancer using low-dose computed tomography (LDCT) reduce deaths from lung cancer?

Key Messages:

•Screening for lung cancer using LDCT may reduce deaths from lung cancer by approximately 21% in high-risk individuals, such as those over 40 with a history of heavy smoking.
•LDCT screening may also slightly reduce deaths from all causes by 5%.
•LDCT screening leads to more invasive tests, higher rates of false-positive results, and increased recall rates compared to chest X-ray (CXR) screening.
•There is a possibility of overdiagnosis, with an estimated 18% of lung cancers detected by LDCT potentially not needing treatment.

What is lung cancer? Lung cancer is the leading cause of cancer-related deaths worldwide. Early detection is crucial for improving survival rates, but screening programs for lung cancer have not been widely implemented in many countries. Previous evidence suggested limited effectiveness of screening with chest X-rays (CXR) or sputum cytology, but there is growing evidence supporting LDCT as a more effective screening tool.

Why is it important to screen for lung cancer? Lung cancer often presents at an advanced stage, where treatment options are limited and survival rates are low. Screening high-risk populations, such as long-term smokers, could help detect lung cancer earlier, leading to more effective treatments and improved survival outcomes.

What did we want to find out? We aimed to determine whether screening for lung cancer using LDCT reduces deaths from lung cancer, and to evaluate the potential harms associated with LDCT screening, such as false-positive results, overdiagnosis, and the need for invasive procedures.

What did we do? We reviewed 11 randomized controlled trials (RCTs) involving 94,445 participants, primarily conducted in Europe and the USA. Participants were aged 40 years or older, with most trials requiring a history of heavy smoking. We assessed the certainty of the evidence using GRADE and performed meta-analyses to combine the results of the included studies.

What did we find?

Lung Cancer-Related Mortality: Screening with LDCT reduced deaths from lung cancer by 21% compared to control groups that either had no screening or were screened with CXR (RR 0.79, 95% CI 0.72 to 0.87). This was based on moderate-certainty evidence from 91,122 participants across 8 trials.

•**All-Cause Mortality:** LDCT screening also led to a slight reduction in deaths from all causes, including lung cancer (RR 0.95, 95% CI 0.91 to 0.99).
•**Invasive Tests:** Participants in the LDCT group underwent more invasive tests compared to those in the control groups (RR 2.60, 95% CI 2.41 to 2.80).
•**False Positives and Overdiagnosis:** LDCT screening had higher false-positive rates and recall rates compared to CXR, though the evidence was of low certainty. Overdiagnosis was estimated at 18%, but the wide confidence interval (0 to 36%) indicates uncertainty in this estimate.
•**Health-Related Quality of Life (HRQoL):** Anxiety levels were slightly lower in the LDCT group compared to the control group (SMD -0.43, 95% CI -0.59 to -0.27), based on low-certainty evidence. There was not enough data to determine the impact of LDCT screening on smoking behavior.

What are the limitations of the evidence? The certainty of the evidence varied, with some outcomes being supported by low-certainty evidence due to risks of bias and variability in study results. Additionally, the studies included in the review had different designs and protocols, which could affect the consistency of the findings.

How up-to-date is this evidence? The evidence is current up to July 2021.

Authors' Conclusions: The current evidence supports the use of LDCT for lung cancer screening in high-risk populations, such as older adults with a significant history of smoking, as it likely reduces lung cancer-related mortality. However, the potential harms, including overdiagnosis and increased invasive testing, must be carefully considered. More research is needed to refine screening strategies, including participant selection and screening frequency, and to explore the role of LDCT screening in non-smokers.

Table

Comparison of performance between 100 AI-PLS, Au-PLS, and Cochrane Abstracts

Measure	AI PLS Vs Au PLS			AI PLS Vs Abstract			Au PLS Vs Abstract		
	TS Diff	ReadDiff	RareWDiff	TS Diff	ReadDiff	RareWDiff	TS Diff	ReadDiff	RareWDiff
Average score	9.41	0.5	2.08	5.67	14.98	-2.05	-3.74	14.48	-4.13
T-test (paired)	0.000	0.689	0.000	0.000	0.000	0.000	0.000	0.000	0.000

*Note: TS Diff = **Text Score** Difference, ReadDiff = **Readability** Difference, RareWDiff = **Rare Words** Difference.*

Further expert evaluation is necessary to address the linguistic limitations comprehensively. Need to test PLSs generated from full-texts

Conclusions:

- AI PLS readability was not inferior and text scores were higher compared with both Au-PLS and Abstract. Content is generally acceptable.
- The initial AI text is a solid starting point for a Systematic Review Plain-Language Summary, but it requires human editing to improve its clarity and comprehensiveness.
- There are occasional gaps in how populations (P of the PICO) are captured or described; sometimes, the key messages may need more accuracy.
- These summaries can also be adapted to other dissemination materials, such as newsletters and tweets, enhancing their utility.

Thanks!: Ms Carmela Tuñón

Bardach Ariel, Glujovsky Demián, Agustín Ciapponi

abardach@iecs.org.ar www.iecs.org.ar

Cochrane Argentina

IECS INSTITUTO DE EFECTIVIDAD CLINICA Y SANITARIA