Plain language summaries created with Artificial intelligence – Is it saving or wasting time?

Kuennemann, Kathi[1], Martin, Sandra[1], Oeztuerk, Seyma[1], Iwobi, Azuka[1], Kuepper, Edith[1], van Dijk, Maarten[1], Esmaeili, Habib[1]

[1]Staburo GmbH, Munich, Germany

BACKGROUND

- **Plain language summaries (PLS)** are currently a requirement to accompany the summary of clinical trial results submissions according to the European Union Clinical Trials Regulation (EU CTR) 536/2014 Annex V. They aim to contribute to more transparency for people interested in learning about clinical study results, especially for those without a medical background.
- **Artificial intelligence (AI) tools** are fast evolving and play an increasingly important role in many fields, including healthcare and medicine. In medical writing, AI could become a powerful tool to increase speed and efficiency in creating outputs, however legitimate concerns arise as to whether it could overtake medical writers in the future. A recent publication describes the creation of 50,000 PLS solely with AI tools (D). **But can AI generate text with the same quality as written by an (experienced) Medical Writer (MW), especially regarding correct interpretation of study results and requirements of lay language?**

GOALS

- The main objective of this study is to investigate if a PLS created by using an AI tool followed by a review from a MW can **save the MW a significant amount of time** to finalize the document.

- The secondary objective is to investigate if a PLS created by using an AI tool has a **similar quality in terms of correctness and understandability** compared to a PLS created by a MW.

Primary endpoint:
- The time used to create a PLS by using an AI tool followed by a MW review, and by a MW.

Secondary endpoint:
- Correctness of a PLS created by an AI tool
- Understandability of a PLS created by an AI tool and reviewed by a MW, or a PLS created by a MW

METHODS

Selected studies, ChatGPT 4.0 prompts, PLS template: Documents from 5 publicly available clinical studies were selected (including study synopsis (*source document*), MW created PLS (*reference document*))

Prompt example:
#You are a medical writer, expert in **communicating complex clinical concepts in plain language.**
Please explain **in 6th grade reading level,** and **in 3rd person** what was the aim of this study, including an explanation of the treated disease and the drug that was used in this study. Please do not use metaphors.
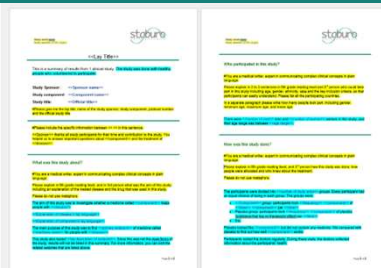
METHODS

Figure 1: PLS template with defined ChatGPT prompts for each PLS section

Two documents:
A. PLS created with ChatGPT 4.0 and reviewed by a MW (*test document*) (study synopsis as *source document*)
B. PLS created by a MW (*reference document*)

Variables to capture objectives:
- The **time** to create each document
- **Correctness** of the AI created document by collecting all issues that were found during the MW review
- **Understandability by using a questionnaire**

A questionnaire was designed to capture the understandability. It consisted of 6 questions about the content of each section of the PLS. Eligible participants to test the understandability were adults and ideally had no medical background. Participants had to self evaluate on a scale from 1 (*content unclear, very complicated*) to 4 (*very well understandable*) as well as write free text answers. Each participant received documents A and B from different studies in the order A-B or B-A. The participants did not know which of their documents was AI created or MW created.

STATISTICAL ANALYSIS

The answer to each question was assessed by researchers and rated as 1 (correct), 0.5 (partially correct) and 0 (wrong). The **percentage of the correct answers** for each PLS section as well as the **overall percentage of understandability** was summarized and reported graphically. The overall score of understandabilities (0-6) was compared between AI and MW created PLS by an **ANOVA** model including treatment (AI or MW), period (1 or 2) and sequence (A-B or B-A) as fixed effects as well as subject specific random effect. The **sample size** was not considered based on a statistical method, such as the power calculation. A sample size of 20-30 participants (equivalent to 40-60 answers) was considered to be sufficient for this exploratory analysis. We received the evaluable answers from N=24 subjects.

RESULTS

Average time used to create each document
- The time to create the AI PLS draft document using ChatGPT 4.0, the time for the following review by a MW, as well as the total time to create the final **document A** were captured (*Figure 2*).

RESULTS

Average time continued
- The average time to create an AI PLS draft document (n=5) was 24.8 minutes (min), and for the following review by a MW was 36.6 min. Hence, the average time to create the final document A was 61.4 min.
- The average reference time to finalize a PLS created by a MW (**document B**) was 4 working days (32 hours).

Understandability of document A compared to document B
- Overall, more than 92% of the free text answers were correct for both documents (*Figure 3*). A similar results was obtained from the self evaluation. No statistically significant difference was found in terms of understandability between document A and B suggesting a similar quality ($p>0.8$).

Figure 2: Times to create document A for each clinical study

Summary of issues regarding the correctness
- Repetition of introductory sentences for the study for almost each prompt
- Inclusion of unnecessary information ("There were no women in this study")
- Problems with reading the adverse event data (missing placebo group, present data not found in source document, missing of percentages)
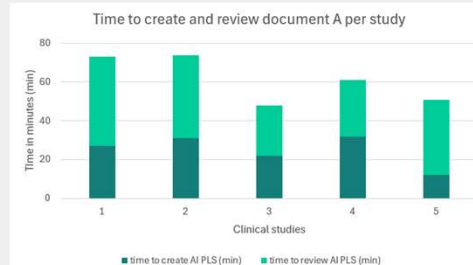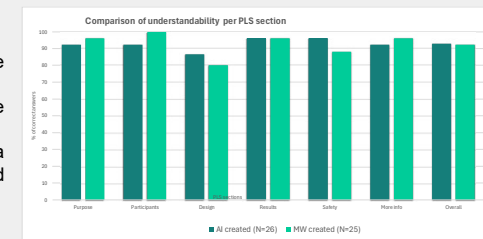- Changes in the order of words in the sentence

Figure 3: Percentage of correct answers per PLS section (A and B)

CONCLUSIONS

Time and correctness
- ChatGPT 4.0, the AI tool used for this project, could help MWs to create PLSs in a significantly shorter time compared to a MW alone. However, the AI tool was not able to replace a MW's work completely.
- There are some issues with the AI created PLS. All these issues can be removed with the MW review of the draft document.

Understandability
- Understandability was similar between the AI created and the MW created document showing that a PLS created by an AI tool and reviewed by a MW can reach the same quality standards as a MW created PLS

Limitations
- Extracting information from the pdf source document did not always work, for the whole pdf or specific sections
- ChatGPT 4.0 imposed a usage limit or rate limiting to each session depending on the amount of data to be processed and to ensure fair access to all users (obligatory breaks of an hour or more).

REFERENCES

A.Barnes, S. Patrick; Lay Summaries of Clinical Study Results: An Overview; Pharmaceutical Medicine (2019) 33:261–268
B.M Delahunty; Artificial intelligence – will we be replaced by robots? September 2019 Medical Writing | Volume 28 Number 3
C.L. Cobb, N. Haycock; Narratives for a clinical study report: The evolution of automation and artificial intelligence; Medical Writing. 2023;32(3):28–31
D.TrialAssure AI Creates Plain Language Summaries For 50,000 Drug Studies; Press Release; TrialAssure Oct 18, 2023