A Multistakeholder Assessment of AI- and Human-Generated PLS

Andy Shepherd,[a] Amy Shaberman,[b] Catherine Richards Golini,[c] Karen Rockell,[d] Jacqui Oliver,[a] Joanne Walker,[e] Trishna Bharadia[f]

[a]Envision Pharma Group, Wilmslow, UK, [b]Boehringer Ingelheim Pharmaceuticals Inc., Ridgefield, CT, USA, [c]Karger Publishers, Basel, Switzerland, [d]UK Organ Donation & Transplantation Research Network, UK, [e]Becaris Publishing, Chesham, Bucks, UK, [f]The Spark Global, Buckinghamshire, UK; Patient Author; Centre for Pharmaceutical Medicine Research, King's College London, UK

POSTER #8

Objective

- To compare and assess 1000-word, text-only plain language summaries (PLS), as created by different generative Artificial Intelligence (genAI) platforms and medical writers.

Research design and methods

- PLS were created for two different published articles using four genAI platforms and seven medical writers or healthcare professionals (HCPs) self-described as having experience or training in writing for a broad audience. To ensure consistency, both the genAI and human writers received exactly the same guidance and instructions to generate the PLS, with no special prompting:
  - Asked to create a 1000-word summary of each scientific article suitable for a patient or lay audience.
  - Provided with two example scientific articles and their corresponding PLS.
  - Given a single page of guidance on writing PLS and specific style requirements.
- Prior to the multistakeholder assessment, all PLS were assessed for compliance with the brief, accuracy, and reading level, and all underwent linguistic analysis.[1,2]
- PLS were then spell checked and re-formatted for consistency before being sent for blind-evaluation to a panel of expert reviewers, which included medical writers, industry representatives, patient advocates, and publishers who all had extensive experience in writing and reviewing PLS.
- Reviewers were asked to rate each PLS using a Likert scale for understandability, ease of reading, content inclusion, and utility.
- Reviewers also gave an overall rating out of 10 and stated if they thought the PLS were generated using AI.

Results

- All PLS met the brief, producing 900–1200-word summaries that covered the minimum specified content requirements.
  - No issues with data mistakes, misinterpretations or inaccuracies were identified for any of the summaries.
  - An assessment of how well the summaries followed the general writing instructions is given in **Table 1**.

Table 1. Adherence to instructions and guidance

Excellent	
genAI: Claude 3-5 Sonnet	Closely followed the guidance with only minor issues, mostly related to brevity or repetition
Human: Specialist PLS medical writer	
Human: Medical writer following specialist PLS training	
genAI: JC3	
genAI: ChatGPT4o	
Human: Doctor who writes content for a medical app	
Good	
genAI: Claude 3-5 Sonnet	**Missing clear explanations of some technical terms**
Human: Specialist PLS medical writer	**Limited information on next steps/implications**
Moderate	
genAI: Claude 3-5 Sonnet	Inconsistent use of accessible language, tone
Human: Specialist PLS medical writer	Inconsistent use of accessible language
Human: Medical writer following specialist PLS training	Inconsistent use of accessible language

Conclusions

- **GenAI systems demonstrated marked improvements in creating PLS compared to our previous evaluation in Q4 2023.[3]**
- **Although PLS specialist medical writers achieved the highest overall scores, AI-generated PLS were generally well-received and scored consistently well.**
- **Language models without specialized medical writing training produced PLS that were comparable with those written by nonspecialist medical writers and HCPs, even without specialized prompting.**
- **Expert reviewers could not consistently differentiate between PLS created using genAI and those created by human writers.**

Results (continued)

- Readability scores and basic lexical complexity assessments showed little range, showing no correlation with human reviewer assessments of readability or ease of understanding.
- More detailed lexical analysis revealed some potential differences between the outputs from specialist medical writers, genAI outputs, and those from HCPs who write content for lay audiences; this will be investigated further across a broader data set of PLS to confirm.
- Overall, AI-generated PLS were rated higher on average than human-written PLS across most categories (**Table 2**).
- Ratings for human-written PLS were more variable, with both the lowest- and highest-scoring PLS overall developed by human writers, with specialist PLS medical writers achieving the highest overall scores.
- Reviewers were only 31% accurate in identifying whether PLS were created by AI or human writers.
  - Notably, reviewers thought that poor quality human generated content was likely to be AI-generated, indicating a potential reviewer bias.

Table 2. Multistakeholder expert panel ratings for AI-generated and human-generated text-only PLS

	Was the summary pitched at an appropriate level? (rated out of 5)*	Did you find this to be a useful summary? (rated out of 5)**	Did you understand the summary? (rated out of 5)**	Was the summary easy to read? (rated out of 5)**	Was the information you would like to see in a summary present? (rated out of 5)**	Rate the summary overall out of 10**	Do you think this summary was generated using AI?
Human overall (SD)	3.9 (0.9)	2.4 (1.3)	3.7 (1.1)	2.4 (1.4)	3.0 (1.1)	4.1 (2.5)	63%
genAI overall (SD)	2.9 (0.6)	4.1 (0.9)	4.6 (0.6)	4.4 (8.8)	3.5 (1.0)	7.4 (1.8)	19%
Human: Specialist PLS medical writer	3.0	4.0	4.8	5.0	4.7	9.3	0%
AI: Claude 3-5 Sonnet	2.9	4.5	4.7	4.7	4.1	8.3	8%
AI: JC3	3.5	4.0	4.5	4.0	4.2	8.2	0%
Human: Medical writer (general) + PLS training	3.2	4.2	4.8	4.5	3.3	7.3	0%
AI: ChatGPT-4o	2.9	4.4	4.8	4.8	3.5	7.8	25%
AI: Google Gemini 1-5	2.7	3.3	4.3	4.2	2.5	5.7	50%
Human: Doctor who writes content for a medical app	4.2	3.2	4.0	2.7	3.5	5.0	83%
Human: Doctor and medical writer	4.3	2.5	3.8	1.8	3.3	4.0	83%
Human: Doctor and writes web articles	3.7	2.5	3.8	2.4	2.8	4.4	92%
Human: Nurse and medical blogger	4.7	1.7	3.3	1.8	2.9	3.1	75%
Human: Pharmacist, MedEd writer	5.0	1.5	2.8	1.3	2.8	1.8	67%

*Rated 1 to 5 (1: Far too simplistic, 2: Slightly simplistic, 3: About right, 4: Slightly too complex, 5: Far too complex)
**Higher is better
PLS, plain language summary; SD, standard deviation
Heatmap shows unweighted average scores from red to green (worst to best)

Limitations

- Only nonspecialized genAI platforms using general language models were included in this comparison.
- The analysis presented here included only a limited sample size and a small panel of expert reviewers; further analysis based on additional PLS and data from a wider stakeholder survey is ongoing.

Future work

- Nine PLS were also selected for blinded evaluation by a wider stakeholder panel.
  - Surveys were opened in November 2024 and distributed via email and direct posting to various interest groups including HCPs, MedComms professionals, and patient groups.
  - Results of the wider stakeholder survey will be submitted for presentation at the ISMPP Annual Meeting, 2025.
- A larger sample of PLS will be analyzed to further investigate the potential differences between outputs from specialist PLS medical writers, genAI outputs, and those from HCPs.

Presented at the 2025 European Meeting of the International Society for Medical Publication Professionals (27–29 January 2025; London, United Kingdom)

References

1. https://readabilityformulas.com/
2. https://www.sketchengine.eu/
3. Shepherd A, et al. Curr Med Res Opin 2024;40(Suppl 2):S41–S57. doi: 10.1080/03007995.2024.2314805

Disclosures

In addition to the use of generative AI for the creation of the PLS assessed in this study, AI was also used to assist the collation of responses and for the initial data analysis.

Author disclosures are available upon request.

Acknowledgments

Editorial and creative support were provided by Leonard Wills and Jack Tempest of Envision Pharma Group. The authors would like to thank the medical professionals and writers who gave their time to write the PLS and everyone who took the time to complete the survey.