PLSing the crowd? Multistakeholder assessment of human- and AI-generated plain language summaries

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

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

POSTER #34

Objective

- To compare and assess 1000-word, text-only plain language summaries (PLS), as created by different generative AI (genAI) platforms, professional medical writers, and self-described medical writers

Methods

- PLS were created for two different published articles using four genAI platforms, five healthcare professionals (HCPs) self-described as having experience or training in writing for a broad audience, and two professional medical writers
- To ensure consistency, both the genAI and human writers received exactly the same guidance and instructions to generate the PLS, with no further 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
- PLS were spell-checked and reformatted for consistency before being sent for blind evaluation by a panel of expert reviewers.[1] The panel 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, utility, and whether they thought the summary was pitched at an appropriate level
  - Reviewers also gave an overall rating out of ten and stated if they thought the PLS was generated using AI
- Nine of the highest scoring PLS were then selected for blind evaluation by a wider stakeholder panel via survey using the same wording and questions as put to the expert panel
  - Surveys were distributed via email, LinkedIn, and direct posting to various interest groups including HCPs, MedComms professionals, and patient groups
  - Each survey required participants to evaluate either two AI-generated and one human-written PLS, or two human-written and one AI-generated PLS

Results

- Surveys were open from November 16, 2024 through February 17, 2025
- Overall, 660 PLS reviews were completed, including 69 from HCPs, 429 from people involved in MedComms or Publishing, and 162 "Other" (not HCP, MedComms, or Publishing)

The highest rated AI and human writer overall and averaged across all categories were Claude 3.5 Sonnet and a professional medical writer with specialist PLS training, respectively:

- Professional medical writers scored consistently higher than all self-described medical writers

Respondents were unable to accurately identify if a PLS was AI-generated or written by a person

Q. Do you think this PLS was written by AI?

Total responses: 660 (325 Yes vs 335 No), overall accuracy: 41.4%		
Reviewer	**PLS type**	**Correct responses, % (n/N)**
Overall	AI-generated Human-written	42.1 (165/392) 40.3 (108/268)
HCP	AI-generated Human-written	50.0 (21/42) 40.7 (11/27)
MedComms or Publishing	AI-generated Human-written	44.3 (108/244) 41.6 (77/185)
Other	AI-generated Human-written	34.0 (36/106) 35.7 (20/56)

- AI-generated PLS generally received higher overall rating scores than human-written PLS (7.13 vs 5.65, respectively, [P<0.0001])
- However, summaries believed to be AI-generated received a lower mean overall score of 5.71 vs 7.33 for those they believed to have been written by a human (P<0.0001)
  - These findings indicate a potential bias, with reviewers expecting the human-written PLS to be better than the AI-generated PLS, which will be further investigated
- Out of the human writers, professional medical writers with PLS training scored highest overall and across most categories
- JC3 was the AI model scored highest overall and across most categories by HCPs, while Claude 3.5 Sonnet was scored highest by MedComms/Publishing and Other reviewers
- AI summaries were consistently rated as easier to understand across all reviewer groups
- Human-written summaries were more often rated as being pitched at an appropriate level in terms of complexity
- Statistically significant differences were seen in how different stakeholder groups rated the PLS across different categories
  - These differences and the different weightings given to various categories vs overall scores indicate potential differences between what PLS writers, HCPs, and patients valued in a PLS and is the subject of further investigation

Limitations

- Only nonspecialized genAI platforms using general language models were used to generate PLS for this comparison, and better results may be obtained using specialized models
- The analysis presented here included multiple PLS based on two article types from a single therapy area; additional manuscript PLS are being included in ongoing research

Presented at the International Society for Medical Publication Professionals (ISMPP) 2025 Annual Meeting, Washington, DC, USA, May 12–14, 2025.

Conclusions

- **AI-generated PLS have improved significantly since our previous analysis[2]**
- **AI-generated summaries were generally well received, scoring higher on average than all the human-written summaries, except those written by professional medical writers**
- **While reviewers were unable to accurately identify if a PLS was AI-generated or written by a person, reviewers who scored a PLS lower overall were significantly more likely to think it was AI-generated**
- **Overall preferences and relative ratings for each writer or AI varied by stakeholder group**

Fig 1. In general, AI-generated summaries scored higher than human-written summaries overall

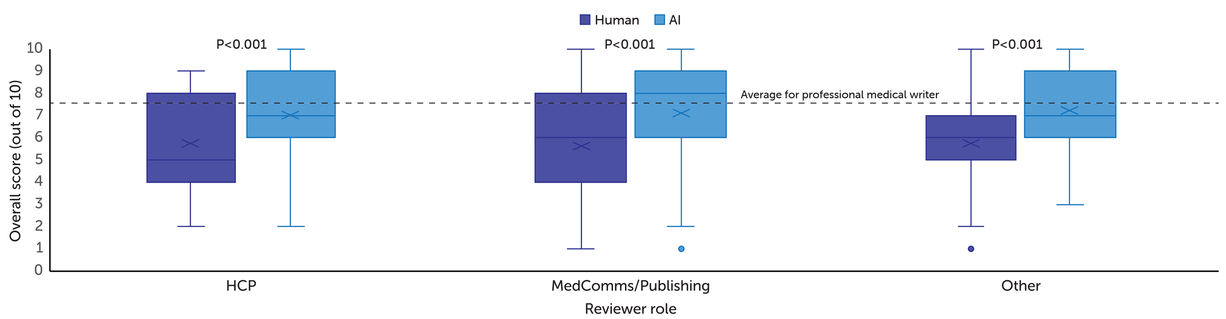


Fig 2. Reviewers who gave a PLS a lower overall score were significantly more likely to think it was AI-generated

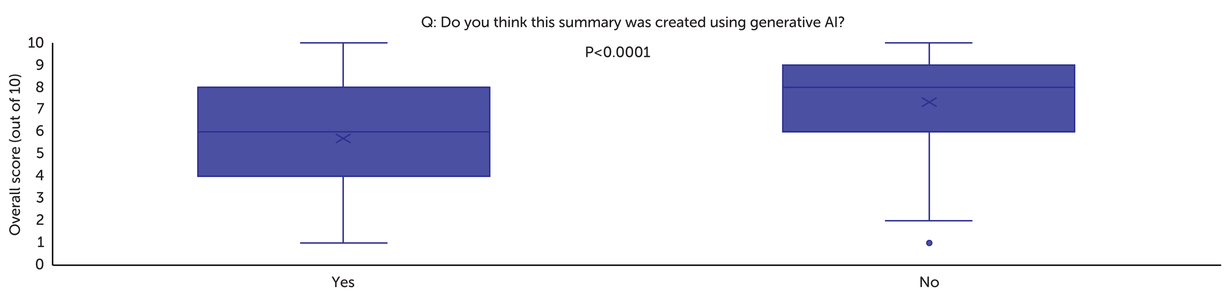
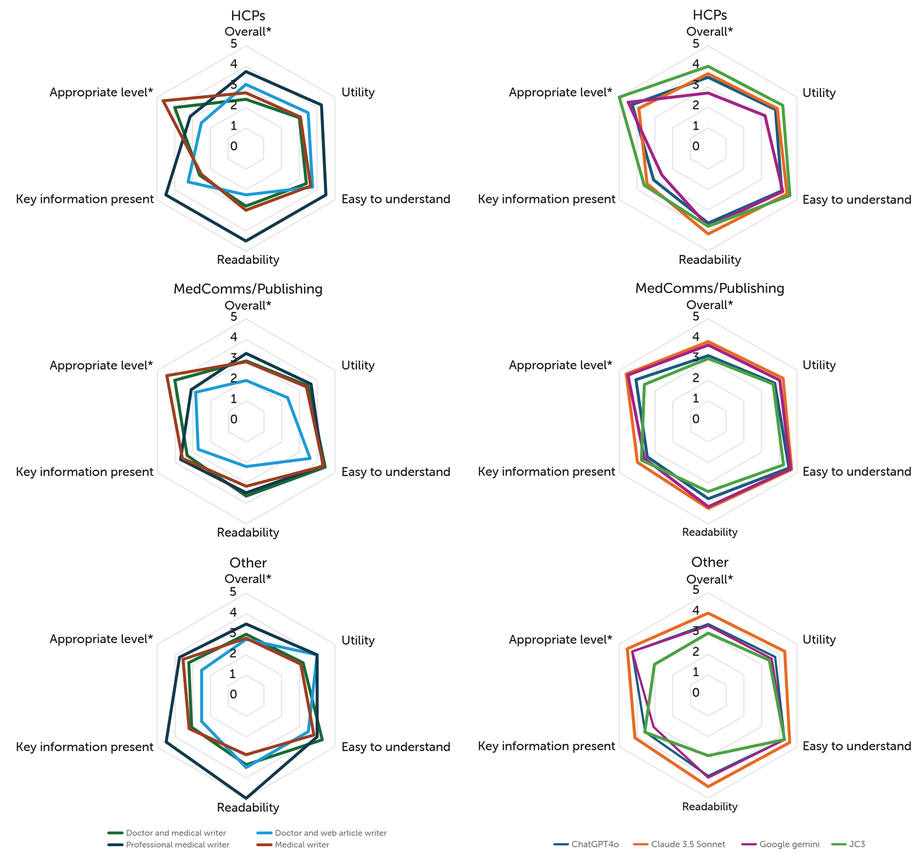


Fig 3. Overall preferences and relative ratings for each writer/AI varied by reviewer role



*Normalized score out of 5. Overall rating was out of 10. Appropriate level was rated out of 5 where 3 was the best score (lower indicating too simplistic and higher indicating too complex).

References

1. Shepherd A, et al. *Curr Med Res Opin.* 2024;40(Suppl 2):S41-S57. doi: 10.1080/03007995.2024.2314805
2. Shepherd A, et al. Presented at the ISMPP EU meeting 2025, London, UK.

Disclosures

Author disclosures are available upon request.

Acknowledgments

Editorial and Creative support was provided by Amanda Clerkin and Clare Goodyear 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.

Scan for PDF