the ability to be relied on as honest or truthful.

[JBI's critical appraisal tools](#)

Quantitative experimental research studies

The escalating complexity of medical literature necessitates tools to enhance readability for patients. This study aimed to evaluate the efficacy of ChatGPT-4 in simplifying neurology and neurosurgical abstracts and patient education materials (PEMs) while assessing content preservation using Latent Semantic Analysis (LSA).

Picton et al. total of 100 abstracts (25 each from Neurosurgery, Journal of Neurosurgery, Lancet Neurology, and JAMA Neurology) and 340 PEMs (66 from the American Association of Neurological Surgeons, 274 from the American Academy of Neurology) were transformed by a GPT-4.0 prompt requesting a 5th grade reading level. Flesch-Kincaid Grade Level (FKGL) and Flesch Reading Ease (FKRE) scores were used before/after transformation. Content fidelity was validated via LSA (ranging 0-1, 1 meaning identical topics) and by expert assessment (0-1) for a subset (n = 40). Pearson correlation coefficient compared assessments.

FKGL decreased from 12th to 5th grade for abstracts and 13th to 5th for PEMs ($p < 0.001$). FKRE scores showed similar improvement ($p < 0.001$). LSA confirmed high content similarity for abstracts (mean cosine similarity 0.746) and PEMs (mean 0.953). Expert assessment indicated a mean topic similarity of 0.775 for abstracts and 0.715 for PEMs. The Pearson coefficient between LSA and expert assessment of textual similarity was 0.598 for abstracts and -0.167 for PEMs. Segmented analysis of similarity correlations revealed a correlation of 0.48 ($p = 0.02$) below 450 words and a -0.20 ($p = 0.43$) correlation above 450 words.

GPT-4.0 markedly improved the readability of medical texts, predominantly maintaining content integrity as substantiated by LSA and expert evaluations. LSA emerged as a reliable tool for assessing content fidelity within moderate-length texts, but its utility diminished for longer documents, overestimating similarity. These findings support the potential of AI in combating low health literacy, however, the similarity scores indicate expert validation is crucial. Future research must strive to improve transformation precision and develop validation methodologies [1]

The study provides compelling evidence that GPT-4.0 can enhance the readability of medical literature while largely preserving content. However, it also highlights the limitations of automated tools in fully capturing content fidelity, particularly for longer texts. Future research should:

Explore methods to refine AI transformations for nuanced accuracy. Investigate the impact of simplified materials on patient comprehension and health outcomes. Develop hybrid workflows that integrate AI simplification with efficient expert review. This research sets a strong foundation for leveraging AI in patient education but underscores the need for continued development and expert oversight to ensure accuracy and trustworthiness.

[1]

Picton B, Andalib S, Spina A, Camp B, Solomon SS, Liang J, Chen PM, Chen JW, Hsu FP, Oh MY. Assessing AI Simplification of Medical Texts: Readability and Content Fidelity. Int J Med Inform. 2024

Dec 1;195:105743. doi: 10.1016/j.ijmedinf.2024.105743. Epub ahead of print. PMID: 39667051.