Assessing Readability and Stigma of ChatGPT's Opioid Use Disorder FAQs

Vitor M Campos, MD[1A] Daniel L Heringer, MD [2A], Gabriel P A Costa[3A], Akhil Anand, MD, FASAM[2A]
1 Hospital Azambuja; 2 Department of Psychiatry and Psychology, Cleveland Clinic; 3 Department of Psychiatry
A Nothing to disclose

Background and Introduction

Artificial intelligence (AI) chatbots such as ChatGPT are increasingly used for patient education, including topics related to opioid use disorder (OUD).

Given the heightened sensitivity of OUD, educational materials must balance depth of information, readability, and non-stigmatizing language. However, AI-generated content may inadvertently increase text complexity and include stigmatizing terms, potentially undermining its utility.

This study aimed to assess whether ChatGPT-generated responses to common OUD Frequently Asked Questions (FAQs):
1) achieve appropriate readability and understandability,
2) adhere to recommended non-stigmatizing language guidelines, and
3) provide sufficient yet accessible detail for patient education.

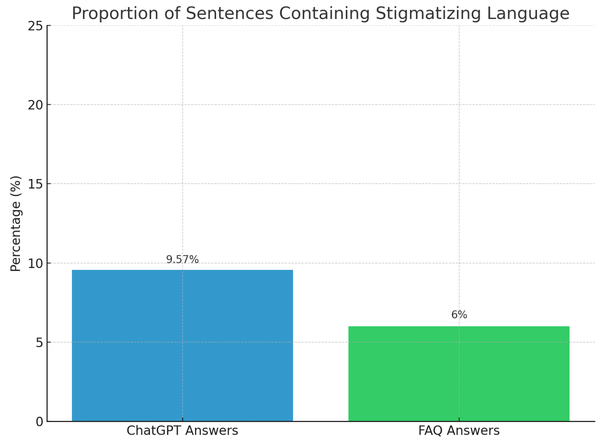
Methods

Using multiple search engines, we identified the top 40 webpages for OUD FAQs. We included only those pages that had a reputable affiliation (e.g., CDC), contained at least five OUD-related FAQs, and offered original, non-duplicated content. After removing webpages with repetitive questions, we ended up with 50 unique OUD FAQs. ChatGPT (model GPT-4o, 2024-11-13) generated answers for each FAQ.
We used online software (e.g., online-utility.org) to compare each answer's character count, word count, sentence count, and lexical density. We also assessed three readability metrics:

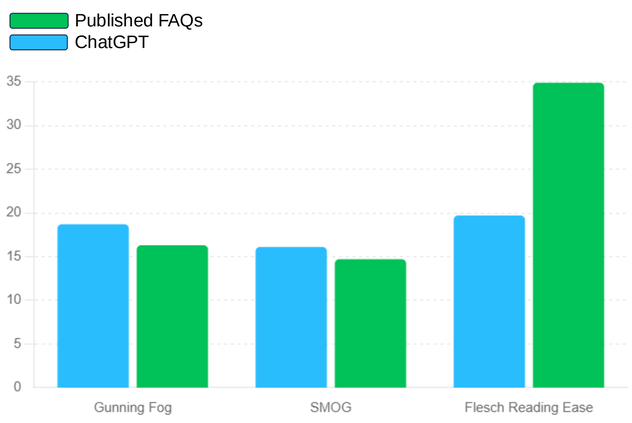
Gunning Fog Index: Estimates the educational level needed to understand a text by focusing on sentence length and the use of complex words.
SMOG: Evaluates text difficulty by counting polysyllabic words to estimate reading grade level.
Flesch Reading Ease: Ranks how easy or difficult a text is to read, using a scale where higher scores mean the text is simpler to comprehend

To compare these readability metrics statistically, we performed a two-tailed Student's t-test (significance set at $p < 0.05$). We examined stigmatizing language using Voyant Tools and the National Institute on Drug Abuse (NIDA) "Words Matter" guidelines. Any sentence containing ≥1 stigmatizing term was labeled as stigmatizing. We then calculated the proportion of stigmatizing sentences in both the published FAQs and the ChatGPT responses. To determine whether these proportions differed significantly, we used a two-tailed z-test for proportions (significance set at $p < 0.05$).

Stigmatizing language comparison



Readability comparison



Results

ChatGPT's answers were significantly longer than existing FAQ responses (mean 1,435 vs. 379 characters).

Readability indices indicated ChatGPT-generated text was more complex (Gunning Fog: 18.7 vs.16.3; SMOG: 16.1 vs.14.7; $p< 0.001$) and had a lower Flesch Reading Ease score, where higher scores mean easier to read (19.7 vs. 34.9; $p< 0.001$).

Stigmatizing language appeared in 9.57% of ChatGPT-generated sentences (49 of 512) and 6.0% of published FAQ sentences (14 of 233). The difference was not statistically significant (z = -1.3944, p = 0.16452).

Conclusion

While ChatGPT can rapidly generate OUD FAQs with comprehensive detail, its outputs are more difficult to read and sometimes include terms that could potentially stigmatize people with substance use disorders.

Although it offers detailed information, refinements are needed to improve readability and reduce stigma in AI-generated patient education materials.

Limitations include potential AI involvement in original FAQs, a relatively small set of 50 questions that may limit generalizability, and the use of a single GPT model version.

Future research with larger, more diverse question sets and updated AI models is warranted to refine AI-generated patient education materials for sensitive conditions like OUD.

References

National Institute on Drug Abuse. (n.d.). *Words matter: Preferred language for talking about addiction.* Retrieved January 22, 2025, from https://nida.nih.gov/research-topics/addiction-science/words-matter-preferred-language-talking-about-addiction
Kelly, J. F., Wakeman, S. E., & Saitz, R. (2015). Stop talking 'dirty': Clinicians, language, and quality of care for the leading cause of preventable death in the United States. *American Journal of Medicine, 128*(1), 8–9. https://doi.org/10.1016/j.amjmed.2014.07.043
Soliman, P., Amaefuna, I., Gully, B. J., & Haass-Koffler, C. L. (2024). Evaluating the readability of online patient-facing resources for alcohol use disorder. *Alcohol, 115*, 1–4. https://doi.org/10.1016/j.alcohol.2023.08.012
Bellanda, V. C. F., Santos, M. L. D., Ferraz, D. A., Jorge, R., & Melo, G. B. (2024). Applications of ChatGPT in the diagnosis, management, education, and research of retinal diseases: A scoping review. *International Journal of Retina and Vitreous, 10*(1), 79. https://doi.org/10.1186/s40942-024-00595-9
Samaan, J. S., Yeo, Y. H., Rajeev, N., Hawley, L., Abel, S., Ng, W. H., et al. (2023). Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. *Obesity Surgery, 33*(6), 1790–1796. https://doi.org/10.1007/s11695-023-06413-2