DANISH MEDICAL JOURNAL

Brief Research Report

ChatGPT versus physician-derived answers to drug-related questions

Ole K. L. Helgestad[1], Astrid J. Hjelholt[1], Søren V. Vestergaard[2], Samuel Azuz[1], Eva A. Sædder[2, 3] & Thure F. Overvad[1]

1) Department of Clinical Pharmacology, Aalborg University Hospital, 2) Department of Clinical Pharmacology, Aarhus University Hospital, 3) Department of Biomedicine, Health, Aarhus University, Denmark

Dan Med J 2025;72(1):A05240360. doi: 10.61409/A05240360

ABSTRACT

INTRODUCTION. Large language models have recently gained interest within the medical community. Their clinical impact is currently being investigated, with potential application in pharmaceutical counselling, which has yet to be assessed.

METHODS. We performed a retrospective investigation of ChatGPT 3.5 and 4.0 in response to 49 consecutive inquiries encountered in the joint pharmaceutical counselling service of the Central and North Denmark regions. Answers were rated by comparing them with the answers generated by physicians.

RESULTS. ChatGPT 3.5 and 4.0 provided answers rated better or equal in 39 (80%) and 48 (98%) cases, respectively, compared to the pharmaceutical counselling service. References did not accompany answers from ChatGPT, and ChatGPT did not elaborate on what would be considered most clinically relevant when providing multiple answers.

CONCLUSIONS. In drug-related questions, ChatGPT (4.0) provided answers of a reasonably high quality. The lack of references and an occasionally limited clinical interpretation makes it less useful as a primary source of information.

FUNDING. None.

TRIAL REGISTRATION. Not relevant.

Clinical pharmacology aims to advance and apply rational pharmacotherapy to benefit patients, healthcare professionals and society at large [1]. To support clinicians, pharmaceutical counselling is available in every Danish region (e.g., "Lægemiddelrådgivningen", "Lægemiddelinformationen" and "MedicinInfo"). The joint pharmaceutical counselling service of the Central and North Denmark regions is enacted by gathering information from books, online sources (e.g., databases, clinical guidelines and literature searches), and occasionally other medical experts.

The emergence of large language models may potentially provide a valuable resource in pharmaceutical counselling if answers are of sufficient quality. c3.5 and 4.0 analyse the content and context of a question and produce output by predicting the most likely next word or sequence of words based on 175 billion and 1.5 trillion parameters, respectively [2-4]. The models are not accountable for their answers or currently considered appropriate for medical guidance. Even experts within the field lack a complete understanding of how ChatGPT works. Still, its answers are based on publicly available online information, licensed from third parties and human trainers, excluding unwarranted information (hate speech, adult content, etc.) [5, 6]. To explore the potential of ChatGPT in pharmaceutical counselling, we conducted a brief investigation assessing how ChatGPT 3.5 and 4.0 responded to questions encountered in the pharmaceutical counselling service of the Central and

North Denmark regions, using our existing physician-derived answers as the gold standard.

Methods

The pharmaceutical counselling service of the Central and North Denmark regions serves a public healthcare sector caring for approximately 1.9 million citizens. Clinical inquiries are answered by physicians in training after guidance and final approval by a senior physician. Inquiries and answers are stored in anonymised form in an online database.

A total of 50 consecutive inquiries during the autumn of 2023 were converted into questions containing minimal clinical context, translated into English and entered into ChatGPT 3.5 and 4.0 (see Supplementary material for questions).

Responses from ChatGPT were compared to those from the counselling service and evaluated using the following six-point Likert scale:

Dangerous: Recommending or overlooking something that is contraindicated or harmful

Worse: Overlooking a clinically meaningful interaction or recommending insufficient treatment

Satisfactory: Providing all relevant information but imprecise or lacking minor details

Equal: Providing roughly the same information

Slightly better: Offering slightly more relevant information

Much better: Providing additional significant and relevant information.

Physicians in clinical pharmacology specialty training (OKLH, AJH, SVV and TFO) assessed each response individually before reaching a joint consensus that was approved by clinical pharmacologists (SA and EAS). When ChatGPT provided supplementary information or answers running counter to the answers provided by the counselling service, we performed an additional literature search for evaluation. No pre-study sample size calculations were performed.

A χ^2 test was used to compare the proportion of responses from ChatGPT 3.5 and 4.0 that were rated ≥ satisfactory compared with the responses from the counselling service. The statistical test was performed using STATA 18 (StataCorp, Texas, USA) and a two-sided $p < 0.05$ was considered statistically significant.

Trial registration: not relevant.

Results

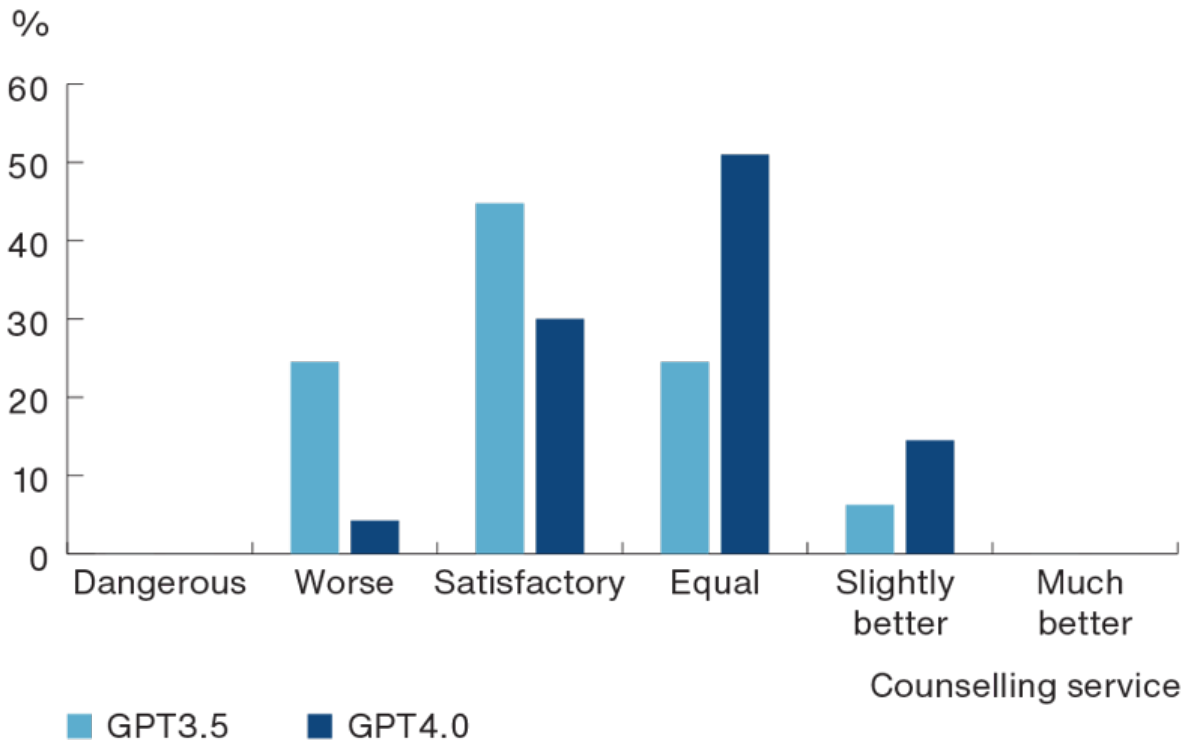
The distribution of the 50 questions by category is shown in **Table 1**. One answer was excluded from the analysis because of a drug mix-up in the question put to ChatGPT.

TABLE 1 Distribution of the questions into categories.

Category	n
Drug interaction	14
Side effects	10
Choice of therapy	8
Pregnancy	7
Miscellaneous	3
Pharmacokinetics	3
Therapeutic drug monitoring	1
Breastfeeding	1
Allergy	1
Pharmacogenetics	1
Poisoning	1

Figure 1 demonstrates the assessment of ChatGPT 3.5 and 4.0 compared with the counselling service evaluated using the Likert scale. No answers provided by either GPT were considered dangerous, nor were any answers considered much better than those provided by the counselling service. In general, GPT 4.0 outperformed 3.5, with 48 (98%) answers considered satisfactory or better, compared to 38 (80%) for GPT 3.5, p = 0.004.

FIGURE 1 The assessment of ChatGPT 3.5 and 4.0 compared with the counselling service, evaluated using the Likert scale.



Neither model provided any references along with its answers, nor did they elaborate on what was most clinically relevant when providing several explanations.

Selected examples

In one case, ChatGPT reported a potentially clinically relevant drug interaction that was not identified by our common drug interaction tools. In another example, ChatGPT provided an additional plausible pharmacodynamic explanation for a suspected side effect. ChatGPT was rated inferior to the counselling service when providing only general considerations not sufficiently helpful for actual clinical decision-making, e.g., "proceedings should be effected with caution" or "doing so is a complex decision that requires special attention".

Discussion

To the best of our knowledge, this is the first evaluation of any large language model regarding clinical pharmacology. In response to 49 consecutive questions, ChatGPT 4.0 and 3.5 provided answers that were rated as better than or equal to satisfactory in 48 (98%) and 39 (80%) cases, respectively, compared to the responses from the pharmaceutical counselling service of the Central and North Denmark regions.

Previous reports have described that ChatGPT passed the United States medical license examination, giving answers of superior quality and empathy compared to physician responses to medical questions from a public social media forum [7, 8]. Our understanding of medicines is far from complete, and translating clinical pharmacological data and evidence into clinical decisions is challenging. Therefore, we find the high quality of

responses (ChatGPT 4.0 in particular) and the speed with which the information could be obtained promising. In a few cases (seven for ChatGPT 4.0 and three for version 3.5), answers were considered more valuable or accurate than those provided by the counselling service. However, in two and 12 cases for ChatGPT 4.0 and 3.5, respectively, inaccurate information was provided, and clinical interpretation was absent, making all considerations seem equally important when the response included multiple explanations. These features make the models less useful without prerequisite expert knowledge, and we find the lack of references particularly worrisome as it impedes direct fact-checking. These limitations might already be outdated, as the newest version of ChatGPT (4o) includes the possibility to specify that answers be based on scientific evidence and accompanied by references (e.g., Scholar GPT, Consensus).

The present study has some limitations; answers were not blinded, which might have led to bias in the assessment (e.g., if assessors were prejudiced in favour of ChatGPT, this might have caused a better rating of ChatGPT, and vice versa). The evaluation was not performed by the clinicians for whom the drug counselling is aimed, and our assessment might differ from a clinician's assessment. Adding more context to the questions might have improved the answers provided by ChatGPT [9].

Conclusions

ChatGPT (4.0) provided answers of reasonably high quality compared to the pharmaceutical counselling service of the Central and North Denmark regions. The most recent version of ChatGPT (4.0) has resolved some of our concerns and provides references along with its answers. However, we believe that additional analyses of substantially larger datasets are required before large language models should be used in daily clinical practice.

Disclosure: ChatGPT 3.5 was used for proofreading, excluding the results section.

Correspondence *Ole K. L. Helgestad*. E-mail: okhelgestad@gmail.com

Accepted 25 September 2024

Published 12 December 2024

Conflicts of interest none. Disclosure forms provided by the authors are available with the article at ugeskriftet.dk/dmj

References can be found with the article at ugeskriftet.dk/dmj

Cite this as Dan Med J 2025;72(1):A05240360

doi 10.61409/A05240360

https://content.ugeskriftet.dk/sites/default/files/2024-09/a05240360-supplementary.pdf

REFERENCES

1. Danish Society of Clinical Pharmacology. Articles of association for the Danish Society of Clinical Pharmacology. Updated Sep 25 2002. https://kliniskfarmakologi.dk (3 Mar 2024)
2. Ruby M. How ChatGPT works: the model behind the bot. https://towardsdatascience.com/how-chatgpt-works-the-models-behind-the-bot-1ce5fca96286 (3 Mar 2024)
3. Emmanuel C. GPT-3.5 and GPT-4 comparison: exploring the developments in AI-Language Models. https://medium.com/@chudeemmanuel3/gpt-3-5-and-gpt-4-comparison-47d837de2226 (3 Mar 2024)
4. OpenAI. Introducing ChatGPT. https://openai.com/blog/chatgpt (3 Mar 2024)

5. OpenAI. How ChatGPT and our language models are developed. https:// help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed (10 Apr 2024)
6. Bowman SR. Eight things to know about Large Language Models. arXiv. 2023; arXiv.2304.0061212v1 [cs.CL]. https://doi.org/10.48550/arXiv.2304.00612
7. Kung TH, Cheatham M, Medenilla A et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198
8. Ayers JW, Poliak A, Dredze M et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183(6):589-96. https://doi.org/10.1001/jamainternmed.2023.1838
9. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res. 2023;25:e50638. https://doi.org/10.2196/50638

