THE JOURNAL OF TELEMEDICINE

Heydari et al., Comparing AI models, Summer 2025 2(3), 55-65

ORIGINAL ARTICLE

Comparative Evaluation of ChatGPT-4o, Claude3.5, Gemini1.5 Pro, and Copilot for Determining Oral Medication Dosages

1. Morteza Heydari*, Mohammadreza Razavizadeh, Saman Sameri, Kaveh Eslami

Received 01/10/2025
Accepted for publication 18/10/2025
Published 29/10/2025

* ***Correspondence to****: Morteza Heydari, Student Research Committee, Faculty of Pharmacy, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran. Email:* Mor547teza@gmail.com

About the authors:

Morteza Heydari; PharmD Candidate, Student Research Committee, Faculty of Pharmacy, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran.

Mohammadreza Razavizadeh; PhD in Pharmaceutic, Department of Pharmaceutics, Faculty of Pharmacy, Tehran University of Medical Sciences, Tehran, Iran.

Saman Sameri; PharmD Candidate, Student Research Committee, Faculty of Pharmacy, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran.

Kaveh Eslami; PhD in Clinical Pharmacy, Associate Professor, Department of Clinical Pharmacy, Faculty of Pharmacy, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran.

This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction, provided the original author(s) and source are credited.

ABSTRACT

Delivering the most effective treatment to patients in the shortest possible time remains one of the most pressing challenges in modern healthcare. Large language models (LLMs) are widely accessible and have shown remarkable potential across domains, including achieving passing scores on the USMLE (United States Medical Licensing Examination), reducing physician visits, and lowering healthcare costs. This study aims to assess the capabilities, limitations, and practical considerations of integrating LLMs alongside pharmacists, with a focus on oral medication dosage prescriptions across different age groups. Questions were organized into seven domains, each comprising three Questions accompanied by clinical case scenarios, and prompts were designed using a zero-shot approach. Responses were evaluated against UpToDate using five criteria: response rate, accuracy, completeness, clarity, and safety. While none of the models had direct access to UpToDate, GPT-4o achieved the highest performance, correctly answering 100% of case-based questions. Copilot achieved 71.43% overall accuracy and 85.71% on case-based questions, but ranked lowest in completeness and clarity. Gemini 1.5Pro demonstrated the lowest response rate, while Copilot and Claude3.5 SonnetV2 generated unsafe outputs. Overall, the findings underscore the importance of evaluating context-dependent effectiveness before the broader adoption of large language models in clinical practice.

Keywords: Generative Artificial Intelligence, Artificial Intelligence, Large Language Models, Drug Information Services, Electronic Prescribing

INTRODUCTION

One of the most pressing challenges in modern healthcare is delivering the most effective treatment to patients in the shortest possible time. In this regard, pharmacists, due to their accessibility and extensive specialized knowledge, are regarded as reliable sources of drug-related information for both patients and physicians. They play a crucial role in addressing specialized needs, optimizing treatments, and ensuring patient safety (1,2).

Providing drug information is one of pharmacists' primary responsibilities. This task is carried out with objectives such as supporting specific medication use practices and improving therapeutic outcomes, which require staying up to date with the latest pharmaceutical information. Therefore, familiarity with systematic approaches is essential for pharmacists (3).

Among the key concerns in drug information provision is determining appropriate oral Medication dosages, which requires careful consideration of the patient's clinical status and medical history. However, one of the major threats to patient safety stems from errors made by healthcare providers, which may result from factors such as a lack of attention to critical details, including pre-existing conditions or patient weight. These errors are intensified by increasing workloads in medical settings (3–6).

With the growing demand for healthcare services, pharmacists' workloads have expanded, limiting their available time. This constraint may hinder their ability to conduct systematic reviews of databases such as UpToDate, thereby restricting their access to the latest and most optimized pharmaceutical information (1–3,7).

LLMs, such as ChatGPT, are emerging technologies resulting from significant advancements in natural language processing (NLP), which involves the processing, understanding, and generation of human language (2,7).

During the development process of these language models, they have been trained on billions of words from various sources. Utilizing deep learning networks, these models have learned the relationships among these words. When asked a question, they analyze the connections between the words in the query and search their training data to identify suitable patterns, generating a response accordingly. Since these responses arise from patterns in human-written text, they appear to be human-generated answers (2,8).

These models are easily accessible and have demonstrated their capabilities across various fields, including achieving the required score on the USMLE exam, identifying potential drug targets, providing personalized education, reducing the need for physician visits, and, consequently, lowering costs. These capabilities could open the door to a smarter, safer future, where language models serve as assistants to humans, facilitating various tasks (2).

However, despite their impressive performance in previous studies, these models have also exhibited serious shortcomings, such as suggesting insulin doses up to 10 times the permissible limit. Such risks raise concerns regarding their applicability in critical domains (7,9–12).

Given the vast scope of large language models and the contradictory findings about their performance, using them without a clear understanding of their capabilities could pose serious risks to patient safety. Given the limited prior research in this area, the present study aims to assess the capabilities, limitations, and practical considerations of integrating these models with pharmacists. Specifically, the performance of four LLMs was evaluated in responding to related queries for oral Medication dosage prescriptions across different age Groups.

METHODS

This descriptive-analytical study was conducted on four large language models (LLMs), selected based on international benchmark performance (GPT-4o and Claude 3.5 Sonnet V2), strong institutional backing (Gemini 1.5 Pro by Google), and broad accessibility (Copilot by Microsoft).

Sampling was performed using the resource equation method to ensure a reliable estimation of error with an optimal number of questions. In this method, the error degree of freedom (E) must lie between 10 and 20 to provide a satisfactory estimate of variance. According to the formula E = N – K, where N is the number of selected questions and K is the number of categories, this study included 7 categories with 3 questions in each (E = 21 – 7 = 14) (2).

The questions were derived from seven clinical domains, including Pediatrics, Adults, Geriatrics, Pediatrics with renal impairment, Pediatrics with hepatic impairment, Adults with renal impairment, and Adults with hepatic impairment.

These domains were selected to reflect routine clinical considerations, such as age, weight, and comorbid conditions. In each category, three commonly prescribed medications were selected based on global prescription trends. For each drug, one direct question and one case-based scenario were designed, yielding a total of 42 questions (21 direct and 21 case-based) per model.

All questions were written in English and applied using the Zero-Shot technique, where no prior examples are provided to the model. This method enables unbiased evaluation of a model's performance and response-generation capabilities, particularly in terms of completeness and reasoning.

Response Evaluation

All responses (excluding those where the model refused to answer) were assessed by a clinical pharmacist across five key dimensions: accuracy & Response Rate, completeness, clarity, and safety. In cases of ambiguity, a second clinical pharmacist was consulted. The reference standard for evaluation was UpToDate.

Accuracy was scored as 0 (incorrect) or 1 (correct), and Response Rate was also defined as the percentage of correct answers out of the total 42 questions (including unanswered ones) to offer a holistic view of performance.

Completeness was evaluated based on the inclusion of relevant details, including dosage, contraindications, common/rare side effects, and clinical considerations.

Clarity was assessed by the extent to which the response was understandable and whether technical terms were explained appropriately. Both completeness and clarity were scored as shown in Table 1.

Safety of responses was categorized into three levels: Safe (clinically acceptable and harmless), Unsafe (containing errors that could mislead but not cause immediate harm), and Hazardous (recommendations that could pose a direct threat to patient safety).

TABLE I. SCORING LEVELS IN EVALUATION

Score	*Completeness*	*Clarity*
1	incomplete	difficult to understand
3	moderately complete	moderately clear
5	Fully complete	completely clear

Statistical Analyses

All data were charted and converted into percentages for comparative analysis. The primary goal was to compare overall model performance in terms of accuracy, completeness, clarity, and safety across both direct and case-based questions.

The Shapiro-Wilk test was used to assess the normality of the data distribution. Because the data did not follow a normal distribution, the Kruskal-Wallis test was used to detect significant differences among models.

Where significant differences were observed, Dunnett's post-hoc test was used for pairwise comparisons. This test identifies which specific models differ significantly from one another.

The Mann-Whitney U test (Rank Sum Test) was used to compare performance between direct and case-based questions within each model, given the non-normal distribution and violations of the assumptions required for parametric testing.

Assessment of comparisons allows for ANOVA when three or more groups are involved and the data are normally distributed; however, if the data are nonnormally distributed, the Kruskal-Wallis test serves as the nonparametric alternative. Following this test, a pairwise comparison analysis would be performed using Dunnett's test to identify groups with significant differences.

All statistical analyses were performed using SPSS version 22, and significance was set at $P < 0.05$. Only the questions to which the models responded were included in the final statistical analysis.

Ethical Statement

Ethical considerations were addressed by obtaining ethical approval from the Ethics Committee of Ahvaz Jundishapur University of Medical Sciences (IR.AJUMS.REC.1404.239). No human participants were involved in this study; therefore, informed consent was not required.

RESULTS

None of the evaluated language models had direct access to the UpToDate database. However, GPT-4o, with a response rate and accuracy of 92.86%, provided significantly better responses ($P < 0.05$) compared to Gemini 1.5 Pro and Copilot. Notable differences in completeness were also observed among the models, except between Gemini 1.5 Pro and Claude 3.5 Sonnet V2 ($P < 0.05$). In terms of clarity, GPT-4o significantly outperformed all other models ($P < 0.001$), while no differences were found among the remaining LLMs (Figure 1).

Although Gemini 1.5 Pro had the lowest response rate (50%), it achieved 75.00% accuracy. Conversely, Copilot demonstrated the lowest completeness, placing it last among the evaluated models ($P < 0.05$). In terms of clarity, although Copilot had the lowest percentage, the difference was only statistically significant compared to GPT-4o.

Regarding response safety, the majority of answers were deemed safe. However, Copilot generated one unsafe and one hazardous response, while Claude 3.5 Sonnet V2 produced one unsafe response.

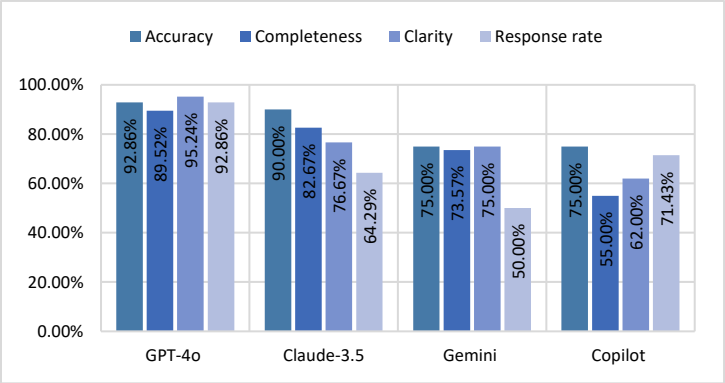

FIGURE I. PERFORMANCE IN ALL QUESTIONS

Direct Questions category

In responding to direct questions, GPT-4o achieved the highest response rate with an accuracy of 85.71%. While there were no statistically significant differences in completeness among the other models, GPT-4o's responses were significantly more complete ($P < 0.05$). In terms of clarity, GPT-4o showed a significant difference only in comparison to Copilot. The lowest response rate and accuracy in this category were recorded for Gemini 1.5 Pro (Figure 2).

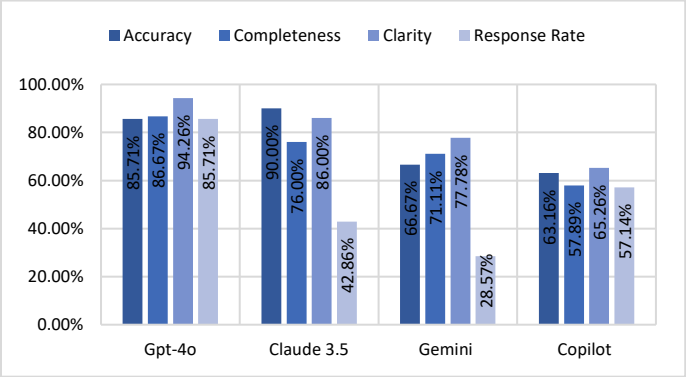

FIGURE II. PERFORMANCE IN DIRECT QUESTIONS

Case-Based Questions category

In case-based scenarios, GPT-4o answered all questions accurately and demonstrated greater clarity than all other models ($P < 0.001$). However, there were no significant differences in the remaining metrics across the models. (Figure 3).

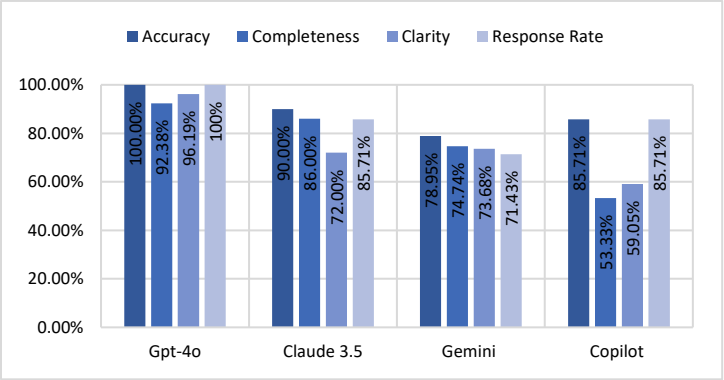

FIGURE III. PERFORMANCE IN CASE-BASED QUESTIONS

Comparison Between Direct and Case-Based Questions

A general increase in response rate was observed across all LLMs when comparing performance on case-based versus direct questions. The most substantial improvements were seen in Gemini 1.5 Pro and Claude 3.5 Sonnet V2. However, a statistically significant decrease in clarity was noted in Gemini 1.5 Pro ($P < 0.05$) when comparing these two question types (Table 2).

TABLE III. PERCENTAGE OF CHANGES IN THE PERFORMANCE OF LLMS IN CLINICAL CASES COMPARED TO DIRECT QUESTIONS

Platforms	Changes of Accuracy	Changes of completeness	Changes of clarity	Change of Response rate
Gpt-4o	16.67%	6.59%	2.05%	16.67%
Claude	0.00%	13.16%	-16.28%	99.86%
Gemini	18.42%	5.10%	-5.27%	150.02%
Copilot	35.70%	-4.04%	-9.56%	50.00%

The categorized analysis is summarized and displayed in Figure 4.

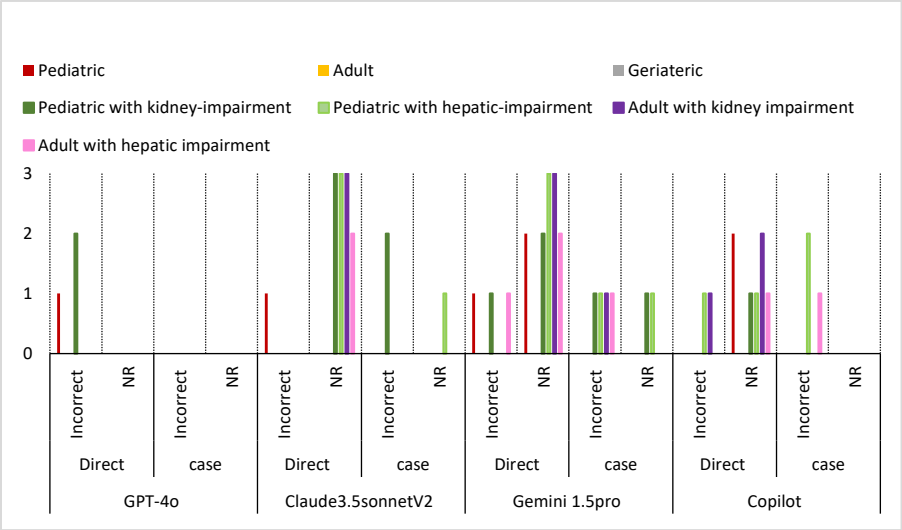

FIGURE IV. NO RESPONSE/INCORRECT QUESTION CATEGORISE

DISCUSSION

In scenarios where other LLMs either refused to provide answers or generated incorrect responses, GPT-4o demonstrated notable proficiency in handling both straightforward and complex queries, whether presented in direct or case-based formats. Compared to other models, it consistently delivered more appropriate answers. Notably, even for questions that were declined by some models, all LLMs still provided some form of user guidance. When faced with more specialized clinical case scenarios, GPT-4o maintained its performance, whereas other LLMs, despite higher response rates, often produced fewer clear answers.

An analysis of the questions that LLMs either failed to answer or answered incorrectly revealed that complex inquiries requiring multi-factorial reasoning, particularly those involving hepatic or renal conditions, posed significant challenges. Many such questions were inadequately addressed by Gemini 1.5 Pro or Copilot. However, GPT-4o was uniquely capable of providing suitable responses to several questions related to hepatic or renal failure. In instances where GPT-4o offered a weaker response, the remaining models typically failed to provide an appropriate answer as well (Figure 4).

Copilot and Gemini 1.5 Pro exhibited weaker overall performance. While Copilot showed a higher response rate than Gemini 1.5 Pro, it lagged significantly in response completeness and even generated one unsafe response concerning a pediatric patient with hepatic impairment, as well as one hazardous recommendation in a pediatric domain. Although Gemini 1.5 Pro was selected for its strong backing from Google, it surprisingly had the lowest overall response rate. Both models showed some improvement on case-based questions, with higher accuracy, yet they continued to trail other LLMs.

GPT-4o, with strong performance, suggests it may serve as a gateway to a safer and more intelligent future in healthcare applications. However, it is critical to interpret these results

in the context of existing limitations and prior research. Meanwhile, Claude 3.5 Sonnet V2 also demonstrated promising results, sometimes approaching the performance of GPT-4o. Nevertheless, its unsafe response to a pediatric hepatic impairment case and the limited number of studies conducted on this model highlight the need for further investigation.

Previous research on various ChatGPT versions has also provided key insights. In a study by Morth et al. (9), ChatGPT suggested insulin dosages up to 10 times higher than appropriate. This study involved real clinical cases with questions repeated at different levels and over time. A later study by Nuland et al. (10) found that although ChatGPT occasionally outperformed pharmacists, it was still not considered reliable for clinical deployment. This study used Dutch-language multiple-choice questions derived from a clinical database and evaluated by pharmacists. Another study by the same author (12), which introduced variations such as personas, data sources, and languages, reported poor performance by ChatGPT. Furthermore, in the study by Grossman et al. (11), the free version of ChatGPT made a critical error in converting intrathecal baclofen to its oral form, underestimating the dose by a factor of 1,000. That study used complex English-language questions drawn from a pharmacy college database and found overall suboptimal performance.

Collectively, these past studies (9–12), along with our own, suggest that the quality of training data across language models is variable. Moreover, access to medical information and the ability to leverage it differ across model versions and can be influenced by prompt design. For instance, in our study, only three out of 25 unanswered questions remained unresolved by the end of testing. Case-based formats significantly reduced the number of unanswered questions, possibly due to more detailed prompts or user recognition as a qualified healthcare professional, both of which may prompt more informative responses through the model's interpretation of specialized medical terminology.

The design and outcome of language model evaluations vary based on question scope, source and language, evaluation method, and scoring criteria. Studies using real hospital cases (which are less likely to be part of LLM training corpora) have identified specific limitations in model responses. These factors likely influenced our findings as well, consistent with prior work such as that by Morth et al. (9).

Our study focused exclusively on seven domains related to oral prescription drug dosage, limiting the scope to prescribed medications. This narrow focus introduces uncertainty regarding the prior exposure of models to similar questions during training. Additionally, our evaluation was limited to four LLMs, none of which were specifically trained for clinical healthcare tasks. During our study period, rapid advances in model development, including the release of new generations such as ChatGPT, DeepSeek, and the Grok family, further complicated longitudinal comparisons, as these newer models have since drawn significant academic attention.

GPT-4o demonstrated exceptional capability in responding to drug information queries. However, given the limitations of our study and prior research, further investigation into the ChatGPT family, as well as alternative models such as Claude, DeepSeek, and Grok, is warranted before clinical deployment. Such research would provide a comprehensive understanding of their strengths and limitations, aiding the identification of models most suitable for use in medical and healthcare environments.

CONCLUSION

Training specialized language models using secure datasets restricted to healthcare professionals, enhancing pharmacist education, and developing tailored usage guidelines are essential steps toward the responsible integration of LLMs into pharmacy practice. As our findings, consistent with earlier studies, underscore, these systems should not be used unsupervised. Instead, they should function as support tools under expert oversight, forming part of a broader strategy for safe and effective healthcare delivery.

Acknowledgments

This study is part of the results from a student research project numbered 04s25 at Ahvaz Jundishapur University of Medical Sciences. The authors would like to thank the staff of the participating university departments for their help in validating the data.

Declaration of the Use of Artificial Intelligence Tools

During the preparation of this manuscript, the authors used ChatGPT to assist with language translation and improve the text's readability. All content generated by the AI tool was carefully reviewed, edited, and verified by the authors.

Contributorship Statement

Conceptualization: MH; Data curation: MH, SS; Formal analysis: MR; Methodology: M R; Writing original draft: MH, SS; Supervision: KE; Data scoring: MR, KE. All authors approved the final manuscript.

All authors reviewed, commented on, and approved the final manuscript, as well as taking responsibility for its content.

Funding Statement

This research did not receive any specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declaration Of Conflicting Interests

The authors declared there are no conflicts of interest regarding the research, authorship, and publication of this article.

Data Availability Statements

The data will be made available by the corresponding author upon reasonable request.

REFERENCES

1. Flôres DDRV, Augusto De Toni Sartori A, Antunes JB, Nunes Pinto A, Pletsch J, Da Silva Dal Pizzol T. Drug information center: challenges of the research process to answer enquiries in hospital pharmaceutical practices. European Journal of Hospital Pharmacy. 2018;25(5): 262–266. DOI: https://doi.org/10.1136/ejhpharm-2017-001417
2. Huang X, Estau D, Liu X, Yu Y, Qin J, Li Z. Evaluating the performance of ChatGPT in clinical pharmacy: A comparative study of ChatGPT and clinical pharmacists. British Journal of Clinical Pharmacology. 2024;90(1): 232–238. DOI: https://doi.org/10.1111/bcp.15896
3. Ghaibi S, Ipema H, Gabay M. Pharmacist's Role in Providing Drug Information. American Journal of Health-System Pharmacy. 2015;72(7): 573–577. DOI: https://doi.org/10.2146/sp150002
4. Radha Krishnan RP, Hung EH, Ashford M, Edillo CE, Gardner C, Hatrick HB, et al. Evaluating the capability of ChatGPT in predicting drug–drug interactions: Real-world evidence using hospitalized patient data. British Journal of Clinical Pharmacology. 2024;90(12): 3361–3366. DOI: https://doi.org/10.1111/bcp.16275
5. Hoyle JD, Davis AT, Putman KK, Trytko JA, Fales WD. Medication Dosing Errors in Pediatric Patients Treated by Emergency Medical Services. Prehospital Emergency Care. 2012;16(1): 59–66. DOI: https://doi.org/10.3109/10903127.2011.614043
6. Peeriga R. Individualization of Drug Dosage. In: Manubolu K, Peeriga R, Chandrasekhar KB (eds) A Short Guide to Clinical Pharmacokinetics. Singapore: Springer Nature Singapore; 2024. p. 97–120. DOI: https://doi.org/10.1007/978-981-97-4283-7_6
7. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Internal Medicine. 2023;183(6): 589. DOI: https://doi.org/10.1001/jamainternmed.2023.1838
8. Giray L. Prompt Engineering with ChatGPT: A Guide for Academic Writers. Annals of Biomedical Engineering. 2023;51(12): 2629–2633. DOI: https://doi.org/10.1007/s10439-023-03272-4
9. Morath B, Chiriac U, Jaszkowski E, Deiß C, Nürnberg H, Hörth K, et al. Performance and risks of ChatGPT used in drug information: an exploratory real-world analysis. European Journal of Hospital Pharmacy. 2024;31(6): 491–497. DOI: https://doi.org/10.1136/ejhpharm-2023-003750
10. Van Nuland M, Erdogan A, Açar C, Contrucci R, Hilbrants S, Maanach L, et al. Performance of ChatGPT on Factual Knowledge Questions Regarding Clinical Pharmacy. The Journal of Clinical Pharmacology. 2024;64(9): 1095–1100. DOI: https://doi.org/10.1002/jcph.2443
11. Grossman S, Zerilli T, Nathan JP. Appropriateness of ChatGPT as a resource for medication-related questions. British Journal of Clinical Pharmacology. 2024; bcp.16212. DOI: https://doi.org/10.1111/bcp.16212
12. Van Nuland M, Lobbezoo AFH, Van De Garde EMW, Herbrink M, Van Heijl I, Bognàr T, et al. Assessing accuracy of ChatGPT in response to questions from day to day pharmaceutical care in hospitals. Exploratory Research in Clinical and Social Pharmacy. 2024;15: 100464. DOI: https://doi.org/10.1016/j.rcsop.2024.100464