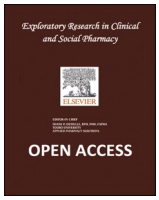
Exploratory Research in Clinical and Social Pharmacy 15 (2024) 100464

Contents lists available at ScienceDirect

Exploratory Research in Clinical and Social Pharmacy

journal homepage: www.elsevier.com/locate/rcsop



Assessing accuracy of ChatGPT in response to questions from day to day pharmaceutical care in hospitals

Merel van Nuland [a,1], Anne-Fleur H. Lobbezoo [a,b,1], Ewoudt M.W. van de Garde [b,c], Maikel Herbrink [e], Inger van Heijl [a], Tim Bognàr [d], Jeroen P.A. Houwen [d], Marloes Dekens [b], Demi Wannet [e], Toine Egberts [c,d], Paul D. van der Linden [a,f,*]

[a] Department of Clinical Pharmacy, Tergooi Medical Center, Hilversum, the Netherlands
[b] Department of Pharmacy, St. Antonius Hospital, Utrecht, Nieuwegein, the Netherlands
[c] Division of Pharmacoepidemiology and Clinical Pharmacology, Department of Pharmaceutical Sciences, Faculty of Science, Utrecht Institute for Pharmaceutical Sciences (UIPS), Utrecht University, Utrecht, the Netherlands
[d] Department of Clinical Pharmacy, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands
[e] Department of Clinical Pharmacy, Meander Medical Center, Amersfoort, the Netherlands
[f] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands

ARTICLE INFO

Keywords:
ChatGPT
Language model
Clinical pharmacy
Drug information
Accuracy

ABSTRACT

Background: The advent of Large Language Models (LLMs) such as ChatGPT introduces opportunities within the medical field. Nonetheless, use of LLM poses a risk when healthcare practitioners and patients present clinical questions to these programs without a comprehensive understanding of its suitability for clinical contexts.
Objective: The objective of this study was to assess ChatGPT's ability to generate appropriate responses to clinical questions that hospital pharmacists could encounter during routine patient care.
Methods: Thirty questions from 10 different domains within clinical pharmacy were collected during routine care. Questions were presented to ChatGPT in a standardized format, including patients' age, sex, drug name, dose, and indication. Subsequently, relevant information regarding specific cases were provided, and the prompt was concluded with the query "what would a hospital pharmacist do?". The impact on accuracy was assessed for each domain by modifying personification to "what would you do?", presenting the question in Dutch, and regenerating the primary question. All responses were independently evaluated by two senior hospital pharmacists, focusing on the availability of an advice, accuracy and concordance.
Results: In 77% of questions, ChatGPT provided an advice in response to the question. For these responses, accuracy and concordance were determined. Accuracy was correct and complete for 26% of responses, correct but incomplete for 22% of responses, partially correct and partially incorrect for 30% of responses and completely incorrect for 22% of responses. The reproducibility was poor, with merely 10% of responses remaining consistent upon regeneration of the primary question.
Conclusions: While concordance of responses was excellent, the accuracy and reproducibility were poor. With the described method, ChatGPT should not be used to address questions encountered by hospital pharmacists during their shifts. However, it is important to acknowledge the limitations of our methodology, including potential biases, which may have influenced the findings.

1. Introduction

The widespread availability of artificial intelligence (AI) has led to a substantial increase in their adoption across various industries, including medical care.[1] The potential of AI resides in its ability to analyze and learn from extensive databases. Recently, Chat Generative Pre-trained Transformer (ChatGPT) has been largely accepted by the wider public. ChatGPT is a Large Language Model (LLM) developed by OpenAI, which is a type of AI that is able to generate human-like responses to questions.[2]

* Corresponding author at: Laan van Tergooi 2, 1212 VG, Hilversum, the Netherlands.
E-mail address: P.D.vanderlinden-4@umcutrecht.nl (P.D. van der Linden).
[1] *These authors have contributed equally and thus share first authorship: Merel van Nuland & Anne-Fleur H. Lobbezoo.*

https://doi.org/10.1016/j.rcsop.2024.100464
Received 12 April 2024; Received in revised form 26 May 2024; Accepted 6 June 2024
Available online 13 June 2024

LLMs, like ChatGPT, introduce new opportunities within the medical field, like improvement of personalized healthcare.[3] On the contrary, it poses a risk when healthcare providers or patients ask questions without a comprehensive understanding of its suitability in such contexts.[4] With the increasing publicity of ChatGPT, its application in medical care – one way or another – seems inevitable and the stakes in healthcare are high. Therefore, it is important to obtain information on the appropriateness of using LLMs in clinical practice.

Various studies about the suitability of LLMs to answer clinical questions have been conducted. For example, a study by Kung et al. showed that ChatGPT achieved sufficient scores (about 60%) on questions from the United States Medical Licensing Examination (USMLE).[5] These results suggest that LLMs hold the potential to support clinical care, and may also be used within pharmaceutical care to enhance efficiency. Meanwhile, the number of studies that investigated the application of LLMs in clinical pharmacy remains limited. In the Netherlands, hospital pharmacists play a crucial role in ensuring patient safety by optimization of pharmacotherapy. Among the tasks of a hospital pharmacist, as stated by the Dutch society for hospital pharmacists (NVZA), are monitoring and guidance of medication, preparation of medication, monitoring the availability of medication, laboratory testing including pharmacogenetics, providing information about medication to doctors and nurses, and providing education. They do so by participating in patient discussions by various medical specialties, by being consulted by physicians and nurses, and intervening in the pharmacotherapy of a patient when deemed necessary. Guidance ranges from providing general information about drug-drug interactions and dosing advice in renal dysfunction to highly individualized advice that necessitates an in-depth search in literature.

The objective of this study was to assess ChatGPT's ability to generate appropriate responses to drug-related clinical questions that a hospital pharmacist could encounter in day to day patient care.

2. Methods

2.1. Setting

This study was conducted by a group of hospital pharmacists and hospital pharmacy residents from the Netherlands in the region Utrecht. Two hospital pharmacy residents from each of the 4 different hospitals participated in data collection. The coordinating hospital was Tergooi Medical Center (Hilversum), and participating hospitals were St. Antonius hospital (Utrecht/Nieuwegein), Meander Medical Center (Amersfoort) and the University Medical Center Utrecht (Utrecht). These hospitals are secondary and tertiary care centers with over 2200 beds together. Three senior hospital pharmacists formed the expert panel for evaluating the accuracy of ChatGPT in answering questions.

2.2. Questions

Questions were gathered during regular clinical pharmacy service hours. For the purpose of this study, ten different domains were defined in which pharmacists have expertise and in which questions emerge from routine patient care. These were: dose advice (over-under dosing), drug-drug interactions, contra-indications, renal dysfunction & dosing advice, therapeutic drug monitoring (TDM), pharmacogenetics, toxicology, compatibilities, manipulation of drug formulations and extravasations. Hospital pharmacy residents collected 3 questions per domain from routine patient care. These questions were checked for consistency and forwarded to a hospital pharmacy residents from a different hospital. This second resident presented the question to ChatGPT 3.5. A new dialogue was started in ChatGPT for each question, except when regenerating a response. Questions were presented in a standardized format in English and presented with a health care provider personification, being 'what would a hospital pharmacist do?'. In all questions, the following fixed set of variables was provided: sex, age, drug name, drug dose and treatment indication. Additional variables, such as target plasma concentration for TDM, were included when deemed necessary. Table 1 presents the template utilized for the standardization of the questions. Furthermore, an example is provided. Based on the number of additional variables included in each question, all questions were classified into three categories: 'simple', 'moderate', and 'complex'. Specifically, a question was appointed as 'simple' when it included $\leq$1 additional variable, 'moderate' when it included 2 additional variables, and 'complex' when it involved $\geq$3 additional variables.

2.3. Input variations

Each of the 30 questions, spanning over 10 different domains, was presented to ChatGPT. The effect of modifying a question on the response by ChatGPT was examined by introducing additional information elements or 'variations' for 1 question per domain, resulting in a total of $n = 10$ additional questions per variation. First, personification was changed from 'what would a hospital pharmacists do' to 'what would you do'. Second, questions were presented in Dutch. Last, responses to the primary questions were regenerated immediately after the initial response by using the regeneration button. The same question within each domain was utilized to assess the effect of the variations. Table 1 presents the template utilized for the standardization of the input variations. Furthermore, examples were provided.

2.4. Performance assessment

All responses by ChatGPT underwent an independent evaluation by two senior hospital pharmacists. This evaluation encompassed responses to the primary question ($n = 30$) and variations in personification and language, as well as for the regenerated responses. The assessment focused on three aspects: the availability of advice, accuracy and concordance, as outlined in Table 2. Initially, it was determined whether ChatGPT provided relevant advice in response to the case presented. If ChatGPT refrained from answering or stated that not enough

Table 1
Template for questions to be presented to ChatGPT, including an example.

Primary question	
Format	A [age]-year old [sex] is treated with [drug][dose]mg for [treatment indication]. [Relevant information for the case]. What would a hospital pharmacist recommend regarding the [drug] dose?
Example	A 66-year old female is treated with edoxaban 60 mg once daily for prophylaxis of stroke and systemic embolism second to atrial fibrillation. Her body weight is 50 kg. What would a hospital pharmacist recommend regarding the edoxaban dose?
Additional variables	Body weight.
Complexity	**Simple: ≤1 additional variable** Moderate: 2 additional variables Complex: ≥3 additional variables.
Input variations	
Modifying personification	A [age]-year old [sex] is treated with [drug][dose]mg for [treatment indication]. [Relevant information for the case]. What would you recommend regarding the [drug] dose?
Example	A 66-year old female is treated with edoxaban 60 mg once daily for prophylaxis of stroke and systemic embolism second to atrial fibrillation. Her body weight is 50 kg. What would you recommend regarding the edoxaban dose?
Question in Dutch	Een [leeftijd]-jarige [geslacht] wordt behandeld met [geneesmiddel][dosis]mg voor [behandelindicatie]. [Relevante aanvullende informatie voor de casus]. Wat zou een ziekenhuisapotheker adviseren met betrekking tot de dosering van [geneesmiddel]?
Example	Een 66-jarige vrouw wordt behandeld met edoxaban 60 mg voor atriumfibrilleren. Het lichaamsgewicht is 50 kg. Wat zou een ziekenhuisapotheker adviseren met betrekking tot de dosering van edoxaban?

Table 2
Scoring system for availability of advice, accuracy and concordance as used by two independent senior hospital pharmacists.

Scoring ChatGPT response	
Input	**Response**
Step 1. Availability advice	1. Advice is *present*. 2. Advice is *absent*. When *present* ➔ continue to step 2. When *absent* ➔ further scoring is not indicated.
Step 2. Accuracy of advice	1. Advice is *correct and complete*. The information is accurate and comprehensive; a board-certified hospital pharmacist would have nothing more to add when consulted. 2. Advice is *correct but incomplete*. All information is correct but incomplete; a board-certified hospital pharmacist would have to add more information when consulted. 3. Advice is *partially correct and partially incorrect.* 4. Advice is *completely incorrect*.
Step 3. Concordance of response	1. Response is *concordant*. The explanation affirms the answer. 2. Response is *discordant*. Any aspect of the explanation contradicts itself.

information was available to commit to an answer, advice was considered absent. If advice was present, a subsequent assessment for accuracy and concordance was conducted. Accuracy was classified across 4 levels, ranging from incomplete and incorrect to correct and complete (see Table 2). The reference sources permitted for use were predetermined, including the summary of product characteristics (SmPC) for individual drugs, the Medicines Information Centre of the Royal Dutch Pharmacists Association (KNMP Kennisbank), the Renal Drug Database, UpToDate, and Micromedex.[6–9] Alternative sources were only allowed to be consulted when information was not available within this designated set of sources. In cases where the two independent senior hospital pharmacists assigned different levels of accuracy, consensus was achieved through consultation with a third senior hospital pharmacist. Finally, responses were considered concordant if the explanation aligned with the answer, and discordant if any aspect of the explanation was contradictory.

To examine the impact of modifying personification and language in the primary question, the accuracy of response was compared between the primary and adjusted question. Additionally, to assess reproducibility, a comparison of accuracy was performed between the response to the primary question and its regenerated counterpart. Lastly, an evaluation was conducted to determine whether complexity of the question was correlated with accuracy of the response.

2.5. Statistical analysis

Descriptive analyses were used for data analysis. Furthermore, the Fisher-Freeman-Halton exact test was used to assess whether accuracy of responses was associated with complexity of the questions.

3. Results

3.1. Data

Responses to the primary question ($n = 30$) and to the modified questions (n = 30) were evaluated by two independent hospital pharmacists. There was consensus for 30 responses, and consensus was reached for the other 30 with a third senior hospital pharmacists. Fig. 1 shows all the individual questions in short, including the presence or absence of advice and the corresponding accuracy level.

3.2. Accuracy

In 23 of 30 questions (77%), ChatGPT provided an advice in response to the question. For these responses, accuracy and concordance were determined. Accuracy was correct and complete for 6 (26%) responses, correct but incomplete for 5 (22%) responses, partially correct and partially incorrect for 7 (30%) responses and completely incorrect for 5 (22%) of responses. The level of accuracy was seemingly evenly distributed across all domains. In comparison to the other domains, 'compatibility' scored fairly well with 2 out of 3 deemed complete and correct, while 'TDM' and 'extravasations' displayed lower scores. The accuracy per question is presented in Fig. 1.

3.3. Concordance

Concordance was determined for 23 of 30 questions (77%) for which an advice was generated. All responses (100%) were considered concordant by the expert panel.

3.4. Complexity

The complexity was calculated for all 30 primary questions. In total, 13 (43%) of the questions were deemed simple, 8 (27%) moderate and 9 (30%) complex. Simpler questions seemed to have a higher level of accuracy when advice was given. The exact p value calculated using the Fisher-Freeman-Halton exact test was 0.034, indicating that there is a significant association between accuracy of the responses and complexity of the questions. For simple questions, the level of accuracy was correct and complete for 4 out of 13 (31%), while this was 2 out of 8 (25%) for moderate questions and 0 out of 9 (0%) for complex questions. Furthermore, the level of accuracy was completely incorrect for 1 out of 8 (12.5%) simple questions, for 3 out of 8 (37.5%) moderate questions and for and 1 out of 7 (14%) complex questions. For questions to which no advise was generated, 5 (71%) were categorized as simple and 2 (29%) as complex. The complexity in relationship to accuracy is depicted in Fig. 2.

3.5. Variations and regeneration

3.5.1. Personification

Ten questions, each representing one domain, were presented to ChatGPT both with and without personification as a hospital pharmacist. Among these 10 questions, an advice was present for 6 (60%) of these. The accuracy varied: 1 out of 6 (17%) responses achieved correct and complete accuracy, 2 out of 6 (33%) had correct but incomplete accuracy, 2 out of 6 (33%) were partially correct and partially incorrect, and 1 out of 6 (17%) was completely incorrect. In 6 out of 10 (60%) questions, the accuracy remained consistent between the primary question and the adjusted question. Among the 4 questions with divergent grading, 75% (3 out of 4) lacked advice when personification was removed.

3.6. Language modification to Dutch

Ten questions, each representing one domain, were presented to ChatGPT in Dutch and in English. An advice was present for 7 of 10 questions (70%). The accuracy was correct and complete for 1 out of 7 (14%), correct but incomplete accuracy for 2 out of 7 (29%), partially correct and partially incorrect for 3 out of 7 (42%) and completely incorrect for 1 out of 7 (14%) of responses. In 3 out of 10 (30%) questions, the accuracy remained consistent when the question was asked in Dutch compared to English. Among the 7 questions with divergent grading, no advice was given in response to 3 (43%) questions.

3.7. Regenerated responses

Ten questions, each representing one domain, were regenerated. For 6 out of 10 questions (60%) an advice was present. The accuracy was correct and complete for 2 out of 6 (33%), correct but incomplete

Domain (total n=30)	Question in short	Advice present?	Accuracy
Dose advice	What is the recommended dose of edoxaban in a patient with body weight 50 kg?	●	●
	What is the recommended dose of paracetamol for chronic use in an 80-year old patient?	●	●
	What would be the advice regarding treatment in a patient using metoprolol tartrate 100 mg OD for atrial fibrillation?	●	NA
Drug-drug interactions	How to handle pain management in a patient using oxycodone and rifampicin concomitantly?	●	●
	Is it safe when ibuprofen treatment is initiated by a surgeon in a patient using methotrexate 25 mg once a week?	●	NA
	How to handle the QTc interaction between fluconazole and escitalopram?	●	●
Contra-indications	What is the recommended dose of nadroparin for a bedridden patient with a BMI of 45 kg/m²?	●	●
	Is pantoprazole considered adequate for treatment of acid reflux in a patient with liver cirrhosis (Child-Pugh score C)?	●	●
	Can cefazolin be administered safely to a patient with confirmed amoxicillin allergy?	●	●
Renal dysfunction	What is the recommended dose of metformin in a patient with eGFR 19 ml/min?	●	●
	What is the recommended dose of apixaban for atrial fibrillation during peritoneal dialysis?	●	NA
	What is the recommended dose of rivaroxaban for atrial fibrillation in a patient with eGFR 36 ml/min?	●	●
Therapeutic drug monitoring (TDM)	How to continue treatment in a patient using vancomycin 2500 mg per 24 hours (continuous infusion) with a steady-state plasma concentration of 33.4 mg/L?	●	●
	How to continue treatment in a patient using gentamicin 5 mg/kg every 36 hours with a plasma trough concentration of 0.9 mg/L?	●	●
	How to continue treatment in a patient using tacrolimus 22 mg TID with a plasma trough concentration of 132.6 µg/L?	●	●
Pharmacogenetics	What is the recommended dose of clopidogrel for treatment of a cerebrovascular event in a patient with CYP2C19 *2/*2 genotype?	●	●
	What is the recommended dose of capecitabine in a patient with DPYD *1/*13 genotype?	●	●
	What is the recommended dose of azathioprine in a patient with TPMT *1/*2 genotype?	●	●
Toxicology	What treatment is recommended for a patient who ingested 7000 mg paracetamol 2 hours prior to presentation at the ER with a plasma level of 279 mg/L?	●	●
	What is considered the severity of an intoxication with clozapine 100 mg ingested 1 hour prior to presentation at the ER?	●	NA
	What toxidrome is evident in a patient presenting with tinnitus, metabolic acidosis, high blood pressure, tachypnea, general confusion, hyperthermia and bradyphrenia?	●	●
Compatibility	Can benzylpenicillin be dissolved and diluted in glucose 5%?	●	●
	Can vancomycin be safely administered over the same Y-site as furosemide?	●	●
	Can gentamicin be added to an infusion bag containing cefuroxime and metronidazole?	●	NA
Manipulation of drug formulations	Can macitentan 10 mg OD and riociguat 2.5 mg TID be administered through an enteral feeding tube?	●	NA
	Can levodopa/benserazide capsules or tablets, with controlled and immediate release formulations, be administered through an enteral feeding tube?	●	●
	Can darifenacin 7.5 mg OD be administered through an enteral feeding tube?	●	NA
Extravasation	How to manage a grade 2/3 extravasation with 150 mL parenteral nutrition?	●	●
	How to manage a grade 3 extravasation with 250 mL infliximab with a concentration of 1.48 mg/mL dissolved in sodium chloride 0.9%?	●	●
	How to manage a grade 1 extravasation with 5 mL sodium chloride 0.9% in a 2-year old?	●	●

Advice ● *Advice present* ● *Advice not available*

Accuracy ● *Correct and complete* ● *Correct but incomplete* ● *Partially correct and partially incorrect* ● *Completely incorrect*

Fig. 1. Per domain all 30 questions asked to ChatGPT in short, whether advice is present or not and the accuracy of the response. Color codes are given below the Table. Abbreviations: CYP = cytochrome P450, DPYD = dihydropyrimidine dehydrogenase, eGFR = estimated glomerular filtration rate, ER = emergency room, OD = once daily, TID = three times daily, TPMT = thiopurine methyltransferase.

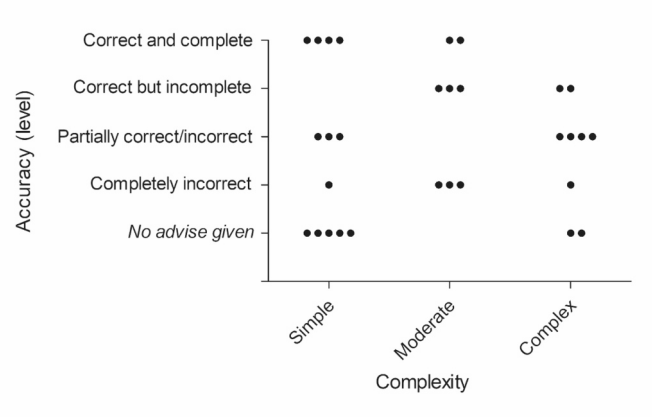


Fig. 2. Complexity of questions versus accuracy. The possible association between accuracy of the responses and complexity of the questions was calculated solely for responses where advice was given. *P*-value: 0.034.

accuracy for 2 out of 6 (33%), partially correct and partially incorrect for 3 out of 6 (33%) and completely incorrect for 0 out of 6 (0%) of responses. In 1 out of 10 (10%) questions, the availability of advice and level of accuracy remained the same after regenerating the primary question. The 1 question for which the level of accuracy remained the same, accuracy was correct and complete. Among the 9 questions with divergent grading, no advice was given in response to 3 (33%) questions.

4. Discussion

Key finding in this study are the limited accuracy of responses, with only 23% being correct and complete and 22% being completely incorrect. The responses that were correct and complete could potentially be used directly for clinical practice, while completely incorrect advice may result in patient harm if instructions were followed. While the accuracy and reproducibility were poor, the concordance was excellent, meaning that ChatGPT is able to provide a consistent response without contradictions.

In this study, we faced several challenges, including where ChatGPT declined to respond and a substantial number of incomplete and incorrect answers. The findings of our study indicate that, in its current form and with this methodology, ChatGPT should not be used to address the questions encountered by hospital pharmacists during their shifts. The accuracy is below the level for safe and high-quality clinical care that is expected from a hospital pharmacist. While other studies have demonstrated variable levels of accuracy in responses by ChatGPT to drug-related questions, the majority underlines our findings; ChatGPT demonstrated correct responses in 26–71% of drug-related questions within clinical pharmacy.[10–12] Additionally, ChatGPT showed excellent results in drug counselling. However, it exhibited limitations in advanced reasoning and handling complex instructions, as observed in tasks such as medication reviews, patient education and the identification and causality assessment of adverse drug reactions (ADRs).[13] Roosan et al. reported a 100% accuracy rate in addressing patient cases of varying complexity, concluding that ChatGPT has the potential to enhance patient safety.[14] It is noteworthy that patient cases in that study were obtained from publicly available materials such as the internet, pharmacy textbooks, and pharmacy school sources, which were accessible by ChatGPT during processing of these questions. Furthermore, patient cases were limited to identification of drug interactions, recommendations on alternative treatment and management plans, lacking patient-specific considerations. Given the discrepancy with other studies, we believe this may have influenced the results. Therefore, we assert that, in its current use and form, ChatGPT (version 3.5) poses a safety risk rather than enhancing patient safety. Additionally, the role of pharmacists is crucial in this context, as the manner in which questions are asked and interpreted requires skill and training. By providing comprehensive training and education, healthcare systems can empower users to utilize AI tools more effectively, ultimately improving patient care and outcomes.

The accuracy we report in this study is evenly distributed across all domains. It seems that questions regarding Y-site compatibility score fairly well, while questions addressing TDM or extravasation of drugs received the lowest scores. This may be attributed to the availability of information in training data. Compatibility data is freely available, while advice following TDM or an extravasation of a drug is mainly retrieved from Dutch protocols that are established by local healthcare authorities, healthcare societies or hospitals. Furthermore, we were surprised by the poor levels of accuracy to questions regarding drug-drug interactions and contra-indications, as such information is freely available in the product information and clinical decision support software such as UpToDate. For these domains, we see that the ChatGPT, on the whole, is able to provide background information, but is unable to translate this to a practical advice for clinicians and patients.

Our study demonstrates a poor reproducibility, with merely 1 of 10 (10%) responses maintaining consistency upon regeneration of the primary question. Poor to very poor reproducibility by ChatGPT was previously highlighted in literature[10] and is of major concern for its use in healthcare.

To ensure valid and reliable results, it is important that outcomes are reproducible, leading to consistent results and conclusions. Enhancing reproducibility can be achieved through the refinement of language models, such as those developed by OpenAI. Additionally, authors play a crucial role to help improve reproducibility by providing data and code alongside their submitted papers.[15]

It is anticipated that language models, including ChatGPT, may eventually integrate into healthcare practices in the future. With the rapid development of these models, their accuracy will improve over time. However, language models can only be as good as the training dataset,[16] thus it is important for datasets to undergo evaluation by experts, and ideally, local guidelines should be incorporated. We believe that pharmacists hold this expertise and that they should take a leading role in integration of language models in pharmaceutical care by doing research, by educating clinicians and by reviewing training datasets. This also means that pharmacists should develop familiarity in using such models.

Our study has some strengths. The questions included in this research are a good representation of questions that are typically posed to hospital pharmacists, covering all relevant domains within pharmaceutical care. Moreover, these questions were gathered through a multicenter approach, including primary, secondary and tertiary care hospitals, thereby enabling the extrapolation of data. Additionally, all the questions in the study are disclosed, providing complete transparency regarding our dataset. Finally, we introduced a straightforward and clear definition of complexity of cases, based on the number of variables.

However, our study is not without limitations. First, there is potential bias due to the assessment methodology. As the use of LLMs is a relatively new field within clinical pharmacy, establishing a standardized set of assessment criteria would be beneficial. Furthermore, training pharmacists in the use of LLMs and involving a prompt engineer could further enhance the validity of the study. Second, even though all relevant domains are covered for a representative 'real-world' subset of questions, we included only 3 questions per domain in this study. Additionally, it is important to note that in the Netherlands, clinical practice relies on national or local guidelines, such as for TDM or extravasations of drugs. ChatGPT does not have access to these guidelines, facing limitations in generating responses aligned with the practices of hospital pharmacists in the Netherlands. A potential for improvement lies in future training of ChatGPT with local guidelines and practice.

5. Conclusion

While our study demonstrated ChatGPT's excellent concordance in responses, the poor accuracy and reproducibility raise concerns regarding its current suitability for addressing day-to-day pharmaceutical care questions in hospitals. The findings of our study indicate that, with the described methodology, ChatGPT should not be used to address the questions encountered by hospital pharmacists during their shifts as the limited performance of ChatGPT poses a safety risk rather than enhancing patient safety. However, it is important to acknowledge the limitations of our methodology, including potential biases, which may have influenced the findings.

Moving forward, further research is needed to fully assess ChatGPT potential in healthcare settings. Despite the current limitations, the promising aspects of ChatGPT suggest that with refinement and proper integration, it could contribute significantly to patient care. Therefore, future studies should address methodological shortcomings and explore the optimal utilisation of ChatGPT in enhancing healthcare practices.

CRediT authorship contribution statement

Merel van Nuland: Writing – original draft, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Anne-Fleur H. Lobbezoo:** Writing – original draft, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Ewoudt M.W. van de Garde:** Writing – review & editing, Supervision, Methodology, Formal analysis, Conceptualization. **Maikel Herbrink:** Formal analysis, Data curation, Conceptualization. **Inger van Heijl:** Formal analysis, Data curation, Conceptualization. **Tim Bognàr:** Formal analysis, Data curation, Conceptualization. **Jeroen P.A. Houwen:** Formal analysis, Data curation, Conceptualization. **Marloes Dekens:** Formal analysis, Data curation, Conceptualization. **Demi Wannet:** Formal analysis, Data curation, Conceptualization. **Toine Egberts:** Writing – review & editing, Supervision, Methodology, Formal analysis, Conceptualization. **Paul D. van der Linden:** Writing – review & editing, Supervision, Methodology, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Plana D, Shung DL, Grimshaw AA, Saraf A, Sung JJY, Kann BH. Randomized clinical trials of machine learning interventions in health care: a systematic review. *JAMA Netw Open*. 2022 Sep 1;5(9), e2233946.
2.. Open AI. Introducing ChatGPT [Internet]. Available from: https://openai.com/blog/chatgpt/.
3. 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 Jun 1;183(6):589–596.
4. Sharma M, Savage C, Nair M, Larsson I, Svedberg P, Nygren JM. Artificial intelligence applications in health care practice: scoping review. *J Med Internet Res*. 2022;24(10):1–17.
5. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digital Health*. 2023 Feb;2(2):e0000198.
6.. Wolters Kluwer. https://www.uptodate.com/login; 1992.
7.. Merative US. https://www.micromedexsolutions.com/home/dispatch; 1973.
8.. De Koninklijke Nederlandse Maatschappij ter bevordering der Pharmacie (KNMP). KNMP Kennisbank [Internet]. [cited 2023 May 26]. Available from: www.kennisbank.knmp.nl.
9.. UK Renal Pharmacy Group (UKRPG). The Renal Drug Database [Internet]. [cited 2023 May 26]. Available from: https://renaldrugdatabase.com/.
10. Morath B, Chiriac U, Jaszkowski E, et al. Performance and risks of ChatGPT used in drug information: an exploratory real-world analysis. *Eur J Hosp Pharm*. 2023 Jun 1.
11. Al-Dujaili Z, Omari S, Pillai J, Al Faraj A. Assessing the accuracy and consistency of ChatGPT in clinical pharmacy management: a preliminary analysis with clinical pharmacy experts worldwide. *Res Soc Adm Pharm*. 2023 Dec;19(12):1590–1594.
12. Fournier A, Fallet C, Sadeghipour F, Perrottet N. Assessing the applicability and appropriateness of ChatGPT in answering clinical pharmacy questions. *Ann Pharm Fr*. 2023 Nov 20.
13. 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. *Br J Clin Pharmacol*. November 2022;2023:232–238.
14. Roosan D, Padua P, Khan R, Khan H, Verzosa C, Wu Y. Effectiveness of ChatGPT in clinical pharmacy and the role of artificial intelligence in medication therapy management. *J Am Pharm Assoc*. 2003, 2023 Dec 2.
15. Pineau J, Vincent-Lamarre P, Sinha K, et al. Improving reproducibility in machine learning research. *J Mach Learn Res*. 2021;22(1):1–20.
16. Harrison CJ, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction to natural language processing. *BMC Med Res Methodol*. 2021;21(1):1–18.