Original research

Performance and risks of ChatGPT used in drug information: an exploratory real-world analysis

Benedict Morath, Ute Chiriac, Elena Jaszkowski, Carolin Deiß, Hannah Nürnberg, Katrin Hörth, Torsten Hoppe-Tichy, Kim Green

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/ejhpharm-2023-003750).

Hospital Pharmacy, Heidelberg University Hospital, Heidelberg, Germany

Correspondence to
Benedict Morath, Hospital Pharmacy, Heidelberg University Hospital, Heidelberg, Germany; benedict.morath@med.uni-heidelberg.de

Received 2 March 2023
Accepted 9 May 2023

EAHP Statement 4: Clinical Pharmacy Services.

ABSTRACT

Objectives To investigate the performance and risk associated with the usage of Chat Generative Pre-trained Transformer (ChatGPT) to answer drug-related questions.

Methods A sample of 50 drug-related questions were consecutively collected and entered in the artificial intelligence software application ChatGPT. Answers were documented and rated in a standardised consensus process by six senior hospital pharmacists in the domains content (correct, incomplete, false), patient management (possible, insufficient, not possible) and risk (no risk, low risk, high risk). As reference, answers were researched in adherence to the German guideline of drug information and stratified in four categories according to the sources used. In addition, the reproducibility of ChatGPT's answers was analysed by entering three questions at different timepoints repeatedly (day 1, day 2, week 2, week 3).

Results Overall, only 13 of 50 answers provided correct content and had enough information to initiate management with no risk of patient harm. The majority of answers were either false (38%, n=19) or had partly correct content (36%, n=18) and no references were provided. A high risk of patient harm was likely in 26% (n=13) of the cases and risk was judged low for 28% (n=14) of the cases. In all high-risk cases, actions could have been initiated based on the provided information. The answers of ChatGPT varied over time when entered repeatedly and only three out of 12 answers were identical, showing no reproducibility to low reproducibility.

Conclusion In a real-world sample of 50 drug-related questions, ChatGPT answered the majority of questions wrong or partly wrong. The use of artificial intelligence applications in drug information is not possible as long as barriers like wrong content, missing references and reproducibility remain.

WHAT IS ALREADY KNOWN ON THIS TOPIC

- ⇒ Drug information is a key clinical pharmacy service that ensures safe and effective pharmacotherapy.
- ⇒ ChatGPT is an artificial intelligence (AI) tool that is able to create content, answer questions and write and correct software code.
- ⇒ Currently it is debated how the use of AI tools will influence science and writing, but no data are available on its performance in drug information.

WHAT THIS STUDY ADDS

- ⇒ ChatGPT partially answered questions correctly if information was available in the summary of product characteristics.
- ⇒ Most questions were answered incompletely or incorrectly with a high risk of patient harm if actions would have been initiated based on the provided information.
- ⇒ However, profound knowledge is needed to identify wrong answers, further highlighting the risk of ChatGPT in real-world scenarios.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

- ⇒ This study highlights barriers that need to be addressed in future AI tools used in clinical contexts for example, referencing, reproducibility and grading of evidence.
- ⇒ Policies and regulations are needed that allow to assess performance and risk of such tools if these should be used in future practice scenarios
- ⇒ Currently, established processes should be followed to answer queries on medicines that do not rely on ChatGPT.

INTRODUCTION

Drug information is a key pharmacy service provided by hospital pharmacies, specialised drug information centres or directly on ward rounds by clinical pharmacists.[1, 2] Advice ranges from general compatibility information up to special individual recommendations on the basis of extensive literature research. The general aim is to enable correct drug use as a pillar of safe and effective pharmacotherapy.[3, 4] To ensure consistent high quality in drug information, various guidelines from different national and international bodies have been issued on how to assess the different sources and compile specific answers on inquiries.[3, 5, 6] However, drug information centres or hospital pharmacies might not always be available owing to lacking resources, although they have been shown to be cost-effective.[7] Indeed, patients and healthcare professionals also search independently for information on the internet and might also use innovative open access tools.[8, 9] These might include tools like the chatbot Chat Generative Pre-trained Transformer (ChatGPT). ChatGPT is a freely available software that is based on artificial intelligence (AI) and machine learning.[9, 10] Although ChatGPT is primarily designed as a chatbot, it was trained with a broad data base and is able to provide comprehensive answers on a plethora of topics, mainly as part of communication.[10] The quality of ChatGPT has raised an ongoing wide discussion

© European Association of Hospital Pharmacists 2023. No commercial re-use. See rights and permissions. Published by BMJ.

To cite: Morath B, Chiriac U, Jaszkowski E, *et al*. *Eur J Hosp Pharm* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/ejhpharm-2023-003750

on its implications on research, education or writing.[11–13] As an example, ChatGPT was able to pass exams of law and business schools as well as medical exams.[14 15]

Drug information is based on communication and in brief on profound answering of questions with different complexity. It appears reasonable to assume that AI tools answering questions and searching data will increasingly be issued, and furthermore, that there might be use cases also in drug information. Currently, there is no analysis on the quality of answers provided by ChatGPT in the area of drug information and its implications on practice. It is unclear whether and how this tool offers opportunities to positively influence clinical practice, or whether it mainly poses a risk to patients. We therefore investigated its performance on 50 questions gathered by a team of clinical pharmacists in clinical routine, and analysed content, implications for practices and risk.

METHODS

Setting

This study was conducted by a team of pharmacists that consecutively collected and documented questions that were asked in clinical routine on two consecutive working days. Questions were gathered during regular clinical pharmacy services hours, that is, during ward rounds, while counseling physicians or in the drug information centre. The senior clinical pharmacists are based in a large hospital pharmacy in a tertiary care centre with over 2000 beds. In this centre, pharmacists are involved in the medication process in a wide range of specialties: surgery (cardiothoracic, vascular, visceral), intensive care (surgical, internal medicine, neurology), neurology, ear-nose-throat and radiotherapy. Furthermore, a regional drug information centre of the chamber of pharmacist is affiliated to this hospital pharmacy department.

Data collection

The study was conducted in Germany. All questions were documented in German and translated by two pharmacists and verified using an online translator (deepl.com). If follow-up questions were needed to answer the query (eg, renal function), this information was included in the final question version. The final question version reflected the question that was answered by the clinical pharmacist and was entered in English in ChatGPT (version 3). All questions were consecutively documented in a prespecified Excel sheet by senior clinical pharmacists until a total of 50 questions was reached. If additional information was needed to answer the initial question, these facts were included in the documented question, for example, lab results or patient characteristics. All questions were entered in ChatGPT once in January 2023 and the results were documented. Questions were only entered once and no additional follow-up questions were asked in ChatGPT. For each question a new dialogue was started in ChatGPT and no answer was rated to not bias results by machine learning during the analysis.

All 50 questions were analysed and rated by predefined and standardised domains. In addition, three questions were repeatedly entered at day 1, day 2, week 2, and week three to investigate whether and how answers differ over time.

Analysis

Rating process

All answers were independently rated by six senior clinical pharmacists who have at least 5 years of clinical experience and hold a certificate of specialty training in clinical pharmacy practice. In case of discrepancies, results were determined by consensus after group discussion.

As rating reference, all answers were searched in the literature and referenced by the pharmacists. The research was performed according to the guideline of drug information of the German Association of Hospital Pharmacists.[6] All researched answers were reviewed independently by two other pharmacists that needed to agree on the answer. If consensus was not reached, the answer was discussed among all authors until these consented. Answers were aimed to be concise and to include a recommendation on treatment. To increase plausibility and standardise the answers, the sources used for each answer were divided in four categories.

For category I, the answer was provided in the summary of product characteristics (SmPC) or was available on a public accessible medical website. In category II, the answer was found in the standard medical literature (eg, books) or international guidelines. Category III was used for answers that were available in non-public medicinal databases or regional/local guidelines. The final category IV was used for answers that required individual literature research or were based on expert knowledge. In cases with sources of different categories, the lowest category is shown. When established databases (eg, UpToDate database) were used to answer the question, the responses were not referenced down on the level of primary literature. This approach was chosen to be practice-orientated as answers in daily routine might also come from established and profound databases (eg, UpToDate database). It was assumed that ChatGPT will most likely have access to open access data, therefore the aim was to use comparable resources.

Rating domains and categories

Three main domains (content, patient management and risk) were defined for the rating process and further divided in three categories. Content was split into the following three categories: complete, incomplete/inconsistent and false. Content was rated as complete when all aspects to answer the question were stated by ChatGPT and were correct. Incomplete/inconsistent answers consisted of partially correct, partially wrong or incomplete information. When content was false or not applicable, the answers was rated as false.

The domain patient management assessed whether the provided information could be used to initiate actions and manage the patient. It was subdivided into three categories: possible, insufficient and not possible/suitable. Management was rated possible if actions could or likely would be initiated based on the answer, even if the content was false; for example, if a wrong dose was recommended by ChatGPT. Management insufficient was chosen if answers did not allow an immediate management and additional information was needed. The category ‘not possible’ was applicable for answers for which no management was possible based on the answer.

The third domain risk was divided into the categories high, low and no risk. High risk was defined as a high risk of patient harm (ie, prolonged hospital stay, occurrence of adverse events, death or ineffective treatment in an acute situation) if healthcare professionals had acted according to the provided information. Answers were rated as low risk if actions based on the answer would likely cause no harm, despite potentially being wrong. As an example, a wrong conversion dose for a statin was recommended that would not likely cause any acute harm or harm in short term. No risk was chosen if the answer had no risk to cause patient harm; for example, if the answer was correct.

Table 1 Questions entered in ChatGPT and rating of the answer in the categories content, patient management and risk

Answer category	Question to ChatGPT	C	M	R
I	What is the labelled standard dose of apixaban in atrial fibrillation?	🟢	🟢	🟢
I	Which antibiotic can be administered in a patient with reported penicillin allergy and nosocomial pneumonia? The allergy occurred 7 years ago and showed a rash on the whole body 7 days after amoxicillin intake.	🟡	🟡	🟡
I	What is the labelled standard dose of apixaban in a patient with CrCl 20 mL/min and atrial fibrillation?	🟢	🟢	🟢
I	Can a patient who reacted with a rash to penicillin in the past be treated with cephalosporins?	🟢	🟢	🟢
I	What is the common starting dose of valsartan/sacubitril?	🟡	🟢	🟡
I	How should ciprofloxacin be dosed in a patient with eGFR 27 mL/min and *Enterobacter cloacae* in the sternal wound?	🟡	🟢	🟢
I	What is the recommended initial dose of vancomycin for an 80 kg patient?	🟡	🟢	🟢
I	Should the caspofungin dose be adjusted in a patient with bilirubin of 17 mg/dL, AST of 400 U/L and ALT of 100 U/L?	🟢	🟢	🟢
I	Can rilpivirine be taken with a proton pump inhibitor in a HIV patient on dolutegravir and rilpivirine with concomitant gastro intestinal bleeding?	🔴	🟢	🔴
I	When should empagliflozin be paused before surgery?	🟢	🟢	🟢
I	When should cotrimoxazole/trimethoprim be adjusted to the renal function?	🔴	🟢	🔴
I	How can olmesartan 20 mg be converted to irbesartan?	🟡	🟢	🟢
I	When can dabigatran be restarted after rifampicin therapy?	🔴	🟢	🔴
I	What is the dose of atorvastatin that is equivalent to simvastatin 40 mg??	🔴	🟢	🟡
I	What is the maximum recommended daily dose of diclofenac?	🟢	🟢	🟢
I	How should vancomycin be dosed in a patient (83 kg) with meningitis and an eGFR 20 mL/min?	🟡	🟢	🟡
I	How should caspofungin be diluted prior to infusion?	🟢	🟢	🟢
I	How long does it take until haemoglobin targets are reached under darbepoetin therapy?	🟡	🟡	🟡
I	Is there an interaction between apixaban and nirmatrelvir/ritonavir(Paxlovid)?	🔴	🟢	🔴
I	What is the maximum daily dose of amlodipine?	🟢	🟢	🟢
I	What is the standard dose of carbamazepine for trigeminal neuralgia?	🟢	🟢	🟢
I	Can citalopram tablets(Cipramil)20 mg be crushed and administered over a gastric feeding tube?	🔴	🟢	🟡
I	Can oxycodone capsules(Oxygesic)10 mg be administered over a gastric feeding tube?	🟡	🟢	🟢
I	Is there a need to adjust bismuth quadruple therapy to the renal function?	🟡	🟡	🟡
I	How should the vancomycin dose be adjusted in a 56 kg patient with catheter-related blood stream infection with a trough level of 25.6 mg/L? The current dose is 1000 mg q12h and eGFR 55.4 mL/min.	🟡	🔴	🟡
I	What is the imipenem dose for a 6 year old patient with meningitis and multiple brain abscesses?	🔴	🔴	🟢
I	What is the recommended dose of linezolid in a 6 year old patient with CNS infection?	🔴	🟢	🔴
II	How should tobramycin be dosed in a patient with 190 kg on CVVHD?	🔴	🟢	🔴
II	Which antibiotic therapy is recommended for endocarditis and penicillin allergy in *Staphylococcus aureus* endocarditis?	🟡	🟢	🟡
II	In which cases is acyclovir prophylaxis recommended in patients with solid tumours?	🟡	🟡	🟡
II	Which drug improves outcomes in patients with heart failure with preserved ejection fraction?	🟡	🟢	🟢
II	How is ceftazidime dosed at an eGFR of 14 mL/min?	🟡	🟡	🟡
II	How should a catheter-related blood stream infection with *Staphylococcus epidermidis* be treated?	🟡	🟡	🟡
II	Which antibiotic is recommended for treatment of surgical site infections after spondylodesis?	🟡	🟢	🟡
II	How is tinzaparin therapeutically dosed in a patient on intermittent hemodialysis?	🟢	🟢	🟢
II	Which antibiotics should be used in a patient with an ear infection due to *Pseudomonas* and *Streptococcus*?	🟡	🟡	🟢
II	How should aztreonam be dosed on a CVVHD with dialysis flow rate of 2 L/h?	🟢	🟢	🟢
III	What is the recommended initial dose of insulin glargine in a patient with steroid-induced diabetes?	🟢	🟢	🟢
III	Which interactions can occur in a patient treated with nirmatrelvir/ritonavir (Paxlovid), atorvastatin, trazodone, paroxetine, acetylic salicylic acid, and candesartan and how should these be managed?	🔴	🔴	🔴
III	How should tobramycin be dosed in a patient with 190 kg on CVVHD?	🔴	🟢	🔴
IV	How much vitamin K can be used to lower the INR to two in a LVAD patient before surgery currently with an INR of 3.5?	🔴	🔴	🔴
IV	Which painkiller should be used in a patient on lithium therapy?	🟢	🟢	🟢
IV	What is the equivalent dose of glibenclamide to glimepiride?	🔴	🟢	🔴
IV	What is the recommended dose of meropenem in paediatric patients with CRRT?	🟡	🟢	🟡
IV	How is ceftolozane/tazobactam dosed in a patient with 198 kg?	🔴	🟢	🔴
IV	How should ceftazidime/avibactam be dosed on a CVVHD with dialysis flow rate of 2 L/h?	🔴	🟢	🔴
IV	How is the conversion dose from 10 mg escitalopram oral to citalopram i.v.?	🔴	🔴	🟢
IV	How should enoxaparin be dosed in a 2 month old infant with recurrent thrombosis on dialysis?	🔴	🔴	🟢
IV	Can ibrutinib be administered over a gastric feeding tube?	🔴	🔴	🟢
IV	How should flucloxacillin be dosed in a patient with *Staphylococcus aureus* endocarditis on CVVHD with dialysate flow rate of 3 L/h?	🔴	🟢	🔴

Continued

Table 1 Continued

Answer category	Question to ChatGPT	C	M	R

Legend: Content: 🟢: Complete. 🟡: Incomplete/Inconsistent. 🔴: False/Not applicable. Patient management 🟢: Possible 🟡: Insufficient 🔴: Not possible. Risk: 🟢: No risk of patient harm 🟡: Low risk of patient harm 🔴: High risk of patient harm.
ALT, Alanine transaminase; AST, Aspartate transaminase; C, Content ; CNS, Central nervous system; CrCl, Creatinine clearance; CRRT, Continuous renal replacement therapy; CVVHD, Continuous veno-venous hemodialysis; eGFR, Estimated glomerular filtration rate; HIV, Human immunodeficiency viruses; INR, International normalised ratio; i.v., Intravenous; LAVD, Left ventricular assist device; M, Patient management; R, Risk; U, Unit.

Reproducibility

To investigate potential variance in ChatGPT's answers over time, three questions had been selected by randomly sorting the list of questions in Excel. The first three questions were chosen for analysis.

Questions were repeatedly entered in ChatGPT at following points in time: day 1, day 2, 1 week after day 1 (week 2) and 2 weeks after day 1 (week 3). Answers were checked for variance in content to initiate actions and were rated in the binary categories: identical and different. In both cases, baseline was the first given answer on day 1. All answers were rated by two independent pharmacists that needed to consent. If consensus was not reached, a third adjudicator decided. Content, management and risk were not rated again in this subanalysis.

RESULTS

Overall, 50 questions were collected and stratified according to the sources used for answering the question (table 1, online supplemental table S1 electronic supplementary). The questions were distributed over four categories and most questions could be answered by using the SmPC or publicly accessible websites (category I, n=27). Further, 10 questions were grouped in category II, 3 questions in category III and 10 questions needed an individual literature research to be answered.

Out of the 50 provided answers, around a quarter (26%) were rated as correct, whereas 36% (n=18) of the answers were only partially correct, had missing information or were ambiguous or vague (table 2, figure 1). The category false was applicable in 38% (n=19) of the answers which provided false information. In addition, no references were provided for all answers by ChatGPT.

However, a majority of answers were quite comprehensive and specific enough to start actions or initiate a management based on the provided information (72%, n=36); also, unfortunately, in cases with wrong content (figure 1). In 14% of the cases, answers were either insufficient to manage the patient (n=7) or management was not possible (n=7) for each category. Nearly half of the answers (46%, n=23) were rated with no risk of patient harm. However, patient harm was likely to be caused in 26% of the cases (n=13) with a high risk of adverse outcomes.

Special scenarios

The highest risk combination consisted of answers that provided wrong content but on the other hand sufficient information to start actions with a high risk of harming a patient (eg, wrong doses). This scenario occurred in 11 answers (22 %), and in addition, two answers had a low risk of harm but still wrong information paired with possible management. In contrast, 13 answers (26%) were correct, provided information to initiate actions and had no risk of harm at all.

When information was not provided in the SmPC and non-public databases (category III) or an individual literature research (category IV) was needed, the answer quality was negatively affected. The content of these answers was false in the majority

Table 2 Questions entered in ChatGPT that were rated either completely correct or completely wrong, that is, rated in the highest or lowest category of all three domains

Answer category	Question to ChatGPT	Overall rating
I	What is the labelled standard dose of apixaban in atrial fibrillation?	🟢
I	What is the labelled standard dose of apixaban in a patient with CrCl 20 mL/min and atrial fibrillation?	🟢
I	Can a patient who reacted with a rash to penicillin in the past be treated with cephalosporins?	🟢
I	Should the caspofungin dose be adjusted in a patient with bilirubin of 17 mg/dL, AST of 400 U/L and ALT of 100 U/L?	🟢
I	When should empagliflozin be paused before surgery?	🟢
I	What is the maximum recommended daily dose of diclofenac?	🟢
I	How should caspofungin be diluted prior to infusion?	🟢
I	What is the maximum daily dose of amlodipine?	🟢
I	What is the standard dose of carbamazepine for trigeminal neuralgia?	🟢
II	How should aztreonam be dosed on a CVVHD with dialysis flow rate of 2 L/h?	🟢
II	How is tinzaparin therapeutically dosed in a patient on intermittent hemodialysis?	🟢
III	What is the recommended initial dose of insulin glargine in a patient with steroid-induced diabetes?	🟢
IV	Which painkiller should be used in a patient on lithium therapy?	🟢
III	Which interactions can occur in patient treated with nirmatrelvir/ritonavir(Paxlovid), atorvastatin, trazodone, paroxetine, acetylic salicylic acid, and candesartan and how should these be managed?	🔴
IV	How much vitamin K can be used to lower the INR to two in a LVAD patient before surgery currently with an INR of 3.5?	🔴

Legend:🟢 : Content, patient management, and risk rated in the highest category. 🔴: Content, patient management, and risk rated in the lowest category.
ALT, Alanine transaminase; AST, Aspartate transaminase; CrCl, Creatinine clearance; CVVHD, Continuous veno-venous hemodialysis; INR, International normalised ratio; LVAD, Left ventricular assist device; U, Unit.

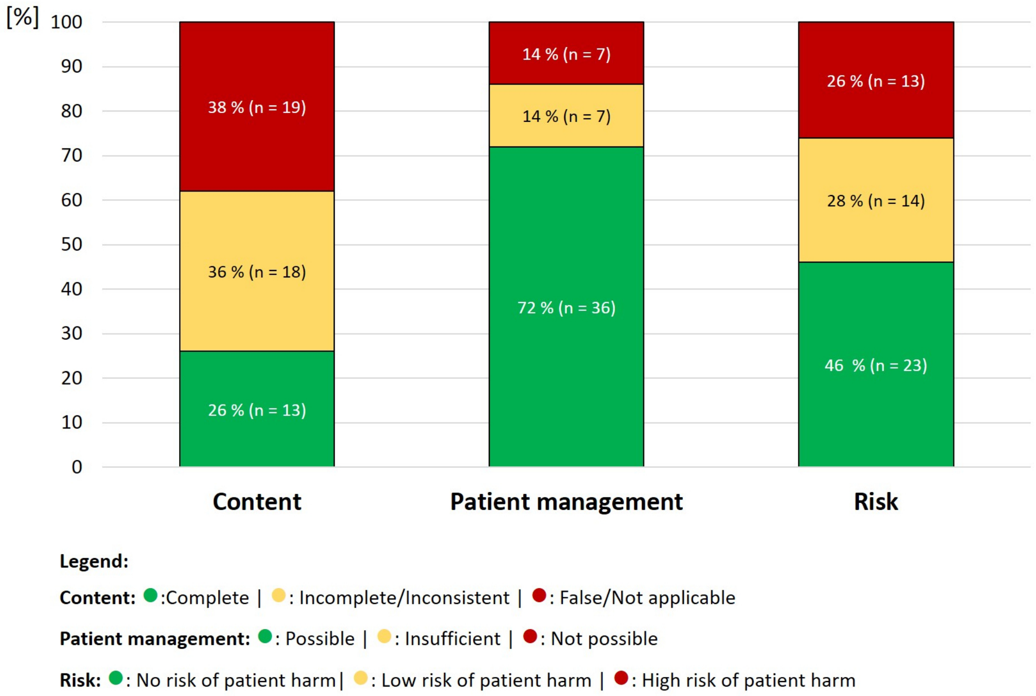

Figure 1 Rating of ChatGPT answers in the categories content, patient management and risk (n=50).

of the cases (76%) and had a high risk of harm in every second case (53%).

Reproducibility and variance in answers over time

Three questions were entered in ChatGPT at prespecified points in time. For two questions, none of the subsequent answers were identical to the initially provided answer and these answers differed substantially. Overall, only two answers were identical (table 3) (online supplemental table S2 electronic supplementary).

In the case of insulin dosage recommendations, dosages varied by a factor of 10 with a serious risk of hypoglycaemia. For aminoglycoside dosing, different and most often wrong formulas were recommended. Additionally, the recommended dosing weight differed by 8.3%−111.7%. Moreover, the recommended weights were miscalculated by ChatGPT (table 3). Only two questions for conversion doses of statins were identical over time, but also in this scenario one answer differed.

DISCUSSION

To our knowledge, this is the first study that investigated ChatGPT in the context of drug information which assessed quality and potential risks of ChatGPT answers if hypothetically applied in clinical practice. Interestingly, 13 out of the 50 questions asked were answered correctly in all domains and information could have been used to initiate actions on this information. On the other hand, the majority of the answers were either insufficient (36%) or false (38%). Consequently, in case of wrong answers with enough information to start actions, there was a high risk of severe patient harm (26%), if instructions of ChatGPT would have been followed in practice. In addition, in a subanalysis answers appeared to change over time for exact the same question showing no reproducibility.

Yet, one has to keep in mind that ChatGPT was developed as a chatbot that should engage in conversation with the capability to write and correct software code.[10] The intended use case is currently neither patient care nor the healthcare setting in general. But this does not preclude that it is used within these areas, hopefully not by healthcare professionals, but possibly by patients familiar with the tool in other areas (eg, writing or for searching general information).[9] The high risk lies not only in the fact that the majority of information is false or partially false, but a relevant factor is that profound knowledge is needed to identify misleading or false information.[16,17] As shown in the study (online supplemental table S1), ChatGPT always provides an answer and often professional terms, dosages or formulas are included. At first glance, answers appear to be plausible.[18] In a question on how to determine dosing weight of aminoglycosides in obese patients, different but often wrong formulas were provided and results were miscalculated.[19] But these answers appeared quick and often sounded confident which could mislead the asking person. Because if the answer was already known, the question would not have been asked.

A machine learning tool can only be as good as the dataset used to train it.[20] ChatGPT appeared to perform well if information was available in the SmPC. In fact SmPCs are often publicly accessible as well as regularly updated, and AI tools might be helpful to query SmPCs or books in future. But the more complex inquiries became, or when answers needed to be individualised and derived from evidence, unreliable and often high-risk answers were provided. The biggest problem was that information was not referenced and answers were not traceable. In addition, a recent study showed that ChatGPT makes up wrong references, that is, references that do not exist.[21] In combination with a lack of reproducibility and varying answers, this currently excludes any real-world use cases in connection with patient care. Nevertheless, these concerns raise important points that need to be addressed in future AI development. In the future, AI tools will become increasingly available. But if these should be used in practice, frameworks are needed for the performance assessment and continuous monitoring. In healthcare use cases, AI tools need to at least correctly reference the data they present, reproducibly create content and, at best, grade

Table 3 Reproducibility and variance in ChatGPT answers over time. Answers were shortened by text without information. Complete answers are available in the electronic online supplemental table S2.

		Rating
Question I	**What is the recommended initial dose of insulin glargine in a patient with steroid-induced diabetes and a body-weight of 65 kg, and blood glucose measurements of 500 mg/dL?**	
Answer ChatGPT day 1	The recommended initial dose of insulin glargine in a patient with steroid-induced diabetes and a body weight of 65 kg and blood glucose measurements of 500 mg/dL would likely be around 0.2 units/kg/day, which would be a total of 13 units/day.	N/A
Answer ChatGPT day 2	The recommended initial dose of insulin glargine (Lantus) in a patient with steroid-induced diabetes and a body weight of 65 kg, and blood glucose measurements of 500 mg/dL, would typically be 0.1 units/kg/day or 6.5 units once daily.	🔴
Answer ChatGPT week 2	The recommended initial dose of insulin glargine for a patient with steroid-induced diabetes and a body weight of 65 kg, and blood glucose measurements of 500 mg/dL would be 0.1–0.2 units/kg/day.	🔴
Answer ChatGPT week 3	The initial dose of insulin glargine for a patient with steroid-induced diabetes and a body weight of 65 kg, and blood glucose measurements of 500 mg/dL is typically 0.5 to 1 unit per kilogram of body weight per day, or 32.5 to 65 units per day.	🔴
Question II	**What is the dosing weight for aminoglycosides in a patient with a body weight of 190 kg and a height of 165 cm?**	
Answer ChatGPT day 1	The dosing weight for aminoglycosides in a patient with a body weight of 190 kg and a height of 165 cm would typically be calculated using the patient's ideal body weight (IBW). The most commonly used formulas for calculating IBW are the Devine formula for males (IBW=50 + 2.3(height (in) - 60) and the Robinson formula for females (IBW=45.5 + 2.3(height (in) - 60). Since we don't know the gender of this patient, we would use the Devine formula. 165 cm=64.96 inches, IBW=50 + 2.3(64.96–60) = 72.986 kg .	N/A
Answer ChatGPT day 2	The dosing weight for aminoglycosides in patients with a body weight of 190 kg and a height of 165 cm is typically calculated using an adjusted body weight (ABW) formula, which takes into account both weight and height. One commonly used formula is: ABW = (0.4 × (height in cm - 152)) + 50 For a patient with a height of 165 cm, this would give an adjusted body weight of: ABW = (0.4 x (165 - 152)) + 50 = 77 + 50 = 127 kg. So, for a patient with a body weight of 190 kg and an ABW of 127 kg.	🔴
Answer ChatGPT week 2	The dosing weight for aminoglycosides in patients is typically based on their ideal body weight (IBW) rather than their actual body weight. To calculate the IBW for a patient with a height of 165 cm, the following formula can be used: IBW (kg) =(height (cm) - 100)x 0.9. For a patient with a height of 165 cm, the IBW would be:(165 - 100)x 0.9=65 kg Therefore, the dosing weight for aminoglycosides in a patient with a body weight of 190 kg and a height of 165 cm would be 65 kg.	🔴
Answer ChatGPT week 3	The dosing weight for aminoglycosides, such as gentamicin, is typically adjusted based on the patient's ideal body weight (IBW) rather than their actual body weight. The IBW can be estimated using various formulas, but a commonly used one is the Devine formula: IBW (kg)=50 + 2.3 kg/inch x (height in inches - 60) For a patient with a height of 165 cm (5 ft five in), the formula would be: IBW (kg)=50 + 2.3 kg/inch x (65 - 60) = 50 + 2.3 x 5=60 kg.	🔴
Question III	**What is the dose of atorvastatin that is equivalent to simvastatin 40 mg?**	
Answer ChatGPT day 1	The equivalent dose of Simvastatin 40 mg in Atorvastatin is approximately 10 mg.	N/A
Answer ChatGPT day 2	The equivalent dose of simvastatin to atorvastatin can vary depending on the individual patient and their response to the medication. However, in general, a dose of atorvastatin 20–40 mg is considered equivalent to a dose of simvastatin 40 mg.	🔴
Answer ChatGPT week 2	The equivalent dose of simvastatin 40 mg in atorvastatin is approximately 10 mg.	🟢
Answer ChatGPT week 3	The equivalent dose of Simvastatin 40 mg in Atorvastatin is approximately 10 mg.	🟢

Legend: 🔴: Answer varies, no reproducibility. 🟢: Answer is identical with first answer.

evidence. If developed for the use in patient care, approval as a medical device is needed according to the medical device regulations.[22]

Aside from ChatGPT's limitations, there are also limitations in regards to this study. We performed a single-centre pilot analysis with a consequent set of questions. Also, the sample size of 50 questions was chosen to reflect the workload of two regular working days and was not determined by a power calculation. However, as this is an exploratory analysis aiming to describe potential implications on clinical practice, this approach appears to be feasible. The associated answers might be influenced by local guidance and processes. It appears likely that other approaches and answers are suitable. Therefore, the German national guideline for drug information was consequently followed and two independent sources were provided if information was not provided in the SmPC. Besides, all involved pharmacists have completed specialty training and a consensus based and standardised rating process was used for the assessment. Moreover, the ChatGPT database is from 2021, and therefore, new data had not been included. Nevertheless, in our opinion all questions could have been answered with data provided from 2021, and therefore the risk of bias seems to be low within this regard. Another issue was that ChatGPT's answers appear to be not reproducible. Therefore, the rating is a snapshot of the day the question was asked. Possibly, wrong answered questions theoretically could be answered in a correct way, the next time entered. However, this reflects the current functionality of ChatGPT and also with repetitive entries of questions, this issue cannot be solved. Our analysis, therefore, represents a real-world scenario that shows potential implications if ChatGPT would have been used in a specific situation in patient care.

CONCLUSION

In an analysis of 50 questions asked in clinical routine, ChatGPT partially answered questions correctly, mostly if information was available in the SmPC. However, currently there is no use case in clinical practice, as answers were often false, insufficient and associated with a great risk of patient harm. In addition, answers appeared to be not reproducible and were not referenced. However, AI tools might become part of the healthcare practice in the future. Therefore, frameworks for evaluation as well as regulations for these tools are needed to ensure the best use in the interest of the patient. Clinical pharmacists should take a leading role in the evaluation of these tools and define the use cases that help to ensure best use in the patient's interest.

Collaborators N/A.

Contributors BM: guarantor for content, conceptualization; data collection; methodology, analysis; supervision; writing of the original draft, review and editing. UC: conceptualization, methodology, formal analysis; writing—review and editing. EJ: data collection, investigation, review and editing. CD: data collection, investigation, review and editing. HN: conceptualization, data collection, investigation review and editing. KH: conceptualization, data collection, investigation, review and editing. THT: conceptualization, review and editing. KG: conceptualization; data collection; methodology; investigation, supervision; writing—review and editing.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. N/A.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iD

Benedict Morath http://orcid.org/0000-0003-1926-0585

REFERENCES

1 Gabay MP. The evolution of drug information centers and specialists. *Hosp Pharm* 2017;52:452–3.

2 SPS. Specialist Pharmacy service-the first stop for professional medicines advice. 2023. Available: https://www.sps.nhs.uk/

3 Ghaibi S, Ipema H, Gabay M, *et al*. ASHP guidelines on the pharmacist's role in providing drug information. *Am J Health Syst Pharm* 2015;72:573–7.

4 SPS. Specialist Pharmacy service-The first stop for professional medicines advice. 2023. Available: https://www.sps.nhs.uk/home/about-sps/

5 Canadian Society of Hospital Pharmacist. Drug information services: Guidelines. 2023. Available: https://www.cshp.ca/docs/pdfs/DrugInformationServicesGuidelines%20(2015).pdf [Accessed 28 Feb 2023].

6 Strobach D, Mildner C, Amann S, *et al*. The Federal Association of German hospital Pharmacists- guideline on drug information. *Krankenhausapharmazie* 2021;42:452–65.

7 Brown JN. Cost savings associated with a dedicated drug information service in an academic medical center. *Hosp Pharm* 2011;46:680–4.

8 Kusch MK, Haefeli WE, Seidling HM. How to meet patients' individual needs for drug information - a Scoping review. *Patient Prefer Adherence* 2018;12:2339–55.

9 Hopkins AM, Logan JM, Kichenadasse G, *et al*. AI Chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. *JNCI Cancer Spectr* 2023;7:pkad010.

10 OpenAI. ChatGPT: Optimizing language models for dialogue. 2023. Available: https://openai.com/blog/chatgpt [Accessed 28 Feb 2023].

11 Stokel-Walker C, Van Noorden R. What ChatGPT and Generative AI mean for science. *Nature* 2023;614:214–6.

12 Park SH, Do K-H, Kim S, *et al*. What should medical students know about artificial intelligence in medicine? *J Educ Eval Health Prof* 2019;16:18.

13 Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. *Nature* 2023;613:612.

14 Gilson A, Safranek CW, Huang T, *et al*. How does ChatGPT perform on the United States medical licensing examination? the implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023;9:e45312.

15 CNN. ChatGPT passes exams from law and business schools. 2023. Available: https://edition.cnn.com/2023/01/26/tech/chatgpt-passes-exams/index.html [Accessed 28 Feb 2023].

16 Sallam M, Salim NA, Al-Tammemi AB, *et al*. ChatGPT output regarding compulsory vaccination and COVID-19 vaccine conspiracy: A descriptive study at the outset of a paradigm shift in Online search for information. *Cureus* 2023;15.

17 NY. Disinformation researchers raise alarms about A.I. 2023. Available: https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html [Accessed 28 Feb 2023].

18 Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative pre-trained transformer 3 (GPT-3) in Healthcare delivery. *NPJ Digit Med* 2021;4:93.

19 Murphy JE. Clinical pharmacokinetics. 6th edn. Bethesda, Maryland: American Society of Health-System Pharmacists, 2017.

20 Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: A practical introduction. *BMC Med Res Methodol* 2019;19.

21 Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus* 2023;15:e35179.

22 European. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending directive 2001/83/EC, regulation (EC) no 178/2002 and regulation (EC) no 1223/2009 and repealing Council directives 90/385/EEC and 93/42/EEC (text with EEA relevance.). 2023. Available: http://data.europa.eu/eli/reg/2017/745/oj [Accessed 28 Feb 2023].