Artificial intelligence-powered chatbots in search engines: a cross-sectional study on the quality and risks of drug information for patients

Wahram Andrikyan ,[1] Sophie Marie Sametinger,[1] Frithjof Kosfeld,[1,2] Lea Jung-Poppe,[1,3] Martin F Fromm,[1,4] Renke Maas,[1,4] Hagen F Nicolaus[1,3]

[1]Institute of Experimental and Clinical Pharmacology and Toxicology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
[2]GSK, Wavre, Belgium
[3]Universitätsklinikum Erlangen, Erlangen, Germany
[4]FAU NeW–Research Center New Bioactive Compounds, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Correspondence to Wahram Andrikyan; wahram.andrikyan@fau.de

RM and HFN contributed equally.

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjqs-2024-017476).

Received 25 April 2024
Accepted 22 August 2024

To cite: Andrikyan W, Sametinger SM, Kosfeld F, *et al*. *BMJ Qual Saf* Epub ahead of print: [please include Day Month Year]. doi:10.1136/bmjqs-2024-017476

ABSTRACT

Background Search engines often serve as a primary resource for patients to obtain drug information. However, the search engine market is rapidly changing due to the introduction of artificial intelligence (AI)-powered chatbots. The consequences for medication safety when patients interact with chatbots remain largely unexplored.

Objective To explore the quality and potential safety concerns of answers provided by an AI-powered chatbot integrated within a search engine.

Methodology Bing copilot was queried on 10 frequently asked patient questions regarding the 50 most prescribed drugs in the US outpatient market. Patient questions covered drug indications, mechanisms of action, instructions for use, adverse drug reactions and contraindications. Readability of chatbot answers was assessed using the Flesch Reading Ease Score. Completeness and accuracy were evaluated based on corresponding patient drug information in the pharmaceutical encyclopaedia drugs.com. On a preselected subset of inaccurate chatbot answers, healthcare professionals evaluated likelihood and extent of possible harm if patients follow the chatbot's given recommendations.

Results Of 500 generated chatbot answers, overall readability implied that responses were difficult to read according to the Flesch Reading Ease Score. Overall median completeness and accuracy of chatbot answers were 100.0% (IQR 50.0–100.0%) and 100.0% (IQR 88.1–100.0%), respectively. Of the subset of 20 chatbot answers, experts found 66% (95% CI 50% to 85%) to be potentially harmful. 42% (95% CI 25% to 60%) of these 20 chatbot answers were found to potentially cause moderate to mild harm, and 22% (95% CI 10% to 40%) to cause severe harm or even death if patients follow the chatbot's advice.

Conclusions AI-powered chatbots are capable of providing overall complete and accurate patient drug information. Yet, experts deemed a considerable number of answers incorrect or potentially harmful. Furthermore, complexity of chatbot answers may limit patient understanding. Hence, healthcare professionals should be cautious in recommending AI-powered search engines until more precise and reliable alternatives are available.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Artificial intelligence (AI)-powered chatbots integrated into search engines are likely to become a key source for obtaining drug information on the web.

⇒ Research on chatbots has primarily been conducted from healthcare professionals' perspective, leaving a research gap in the assessment of quality and exploration of patient safety risks when patients obtain drug information using chatbots.

WHAT THIS STUDY ADDS

⇒ An examination of the quality and potential safety concerns for patients using chatbots to obtain drug information from a patient-centred viewpoint.

⇒ Chatbot answers showed an overall low readability and a high but still not sufficient quality, repeatedly lacking information or containing inaccuracies possibly threatening patient and medication safety.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ We encourage patients as well as healthcare professionals to exercise caution when using AI-powered search engines themselves or recommending them to patients until these identified challenges are addressed in future.

INTRODUCTION

The advent of the internet has revolutionised access to health and drug information with over half of European Union and US citizens now seeking such information online.[1–3] Thereby, search engines are the primary tool for this information retrieval.[1]

In February 2023, the landscape of search engines underwent a significant shift due to introduction of artificial intelligence (AI)-powered chatbots.[4 5] Microsoft's Bing chatbot, an AI-powered copilot for the web, and Google's Gemini (formerly known as Bard) both promise enhanced search results, comprehensive answers and a novel chat experience.[4–6] Both of these chatbots are based on large language models (LLMs), which are neural networks and built with state-of-the-art generative transformer architectures, such as OpenAI's Generative Pre-trained Transformer 4 (GPT-4).[4 5 7–9] This underlying architecture allows the LLMs to be trained on extensive datasets from the whole internet, enabling them to converse on any topic, including healthcare-related queries.[4 5 10]

However, these LLMs pose specific risks, including the generation of disinformation, non-sensical or harmful content.[8 9] While the implications of these models for drug information have been explored, previous studies have primarily focused on the perspective of healthcare professionals.[11–18] Consequently, the quality and potential risks of using AI-powered chatbots for drug information from a patient's perspective remain largely unexplored.

This study aims to fill this gap by investigating readability, completeness and accuracy of chatbot answers and the potential harm to patients using Bing copilot, a search engine with AI-powered chatbot features.

METHODS

This cross-sectional study followed the 'Strengthening the Reporting of Observational Studies in Epidemiology' (STROBE) reporting guideline.[19] The survey was conducted in Germany and did not involve research subjects or personalised data as defined by German law or §15 of the Bavarian professional code of physicians. Thus, an approval by an ethics board was neither required nor applicable.[20] This survey was solely completed by the authors of this study, who had full access to all study data.

Prompt settings

Google and Bing are the most widely used search engines.[21] Given that Google's chatbot regularly refuses to answer medical questions,[22] this study was conducted using Microsoft's Bing AI copilot for the web.[4] Bing chatbot's so-called temperature parameter, which regulates the probability for the output of LLMs,[10] can be specified by choosing one of the three modes: 'creative', 'balanced' or 'precise'.[23] All prompts (ie, queries) were entered in English language in the preselected 'balanced' mode, which is applied by the majority of users.[23] In addition, a virtual private network (VPN) with the location New York (USA) was used for standardised prompts and generalisability of chatbot answers.[24] To avoid any influence of previous browsing activity and chatbot conversation on the prompt, web cookies were deleted at each time and prompts were entered into separate Microsoft Edge browser windows. All chatbot answers were generated in April 2023.

Drug selection

The top 50 most frequently prescribed drugs in the US outpatient market of 2020 were investigated in this study (online supplemental table S1).[25] Six (12.0%) of these 50 constitute non-prescription medications.

Patient questions selection

In order to simulate patients consulting chatbots for drug information, a literature review was performed to identify medication-related questions that patients frequently ask their healthcare professionals. Identification and selection of representative patient questions were based on (1) items of the 'Satisfaction with Information about Medicines Scale' (SIMS),[26] (2) patient guidelines regarding questions to ask healthcare professionals about their medication from the German Coalition for Patient Safety[27] and (3) EuroPharm Forum,[28] (4) six key questions on drug information by Nguyen[1] and (5) experience from a clinical pharmacist (WA) and physicians with expertise in pharmacology (FK, RM, HFN). The 10 selected questions are referred to as 'patient questions' in the following (table 1). The chatbot was asked the selected questions regarding all 50 drugs included in the study. Accordingly, 500 answers in total were generated and analysed.

Assessment of references

Bing chatbot references websites when answering user's questions.[4] These references were documented

Table 1 Questions and evaluation categories of the analysis

No	Category	Question
1	Indication	What is (drug) used for?
2	Mechanism of action	How does (drug) work?
3	Instructions for use	What do I have to consider when taking (drug)?
4	Instructions for use	Can I take (drug) together with food?
5	Instructions for use	Can I drink alcohol while on (drug)?
6	Instructions for use	Are there any other drugs that should not be combined with (drug)?
7	Adverse drug reactions	What are the most common side effects of (drug)?
8	Adverse drug reactions	What are the most serious side effects of (drug)?
9	Contraindications	Can (drug) be used in pregnancy?
10	Contraindications	Can (drug) be used in renal failure?

to receive an impression of the reliability of the sources.

Assessment of readability

Readability of chatbot answers was assessed by calculating the 'Flesch Reading Ease Score'.[29] The Flesch Reading Ease Score is calculated by:

$$\text{Flesch Reading Ease Score} = \begin{aligned}&206.835 - 1.015 \cdot \left(\frac{\textit{total words}}{\textit{total sentences}}\right)\\ &-84.6 \cdot \left(\frac{\textit{total syllables}}{\textit{total words}}\right)\end{aligned}$$

This score estimates the educational level readers need to understand a particular text. Texts that score between 0 and 30 on the Flesch Reading Ease Scale are considered very difficult to read, necessitating a college graduate level. A score of 31–50 is seen as difficult to read, 51–60 as fairly difficult, 61–70 as standard difficulty, 71–80 as fairly easy, 81–90 as easy and 91–100 as very easy to read, appropriate for fifth graders or 11 year-olds.[30]

Assessment of completeness and accuracy

For the assessment of completeness and accuracy of chatbot answers, a reference database was created by a clinical pharmacist (WA) and a physician with expertise in pharmacology (HFN). In this database, all patient questions were answered using drug information for patients provided by the website drugs.com, a peer-reviewed and up-to-date drug information website for both healthcare professionals as well as medical laypersons.[31] The number of statements to answer the patient questions using the patient information was documented.

A statement was defined as a pharmacological or medical proposition regarding a specific topic. Accordingly, the chatbot answers were grouped into statements and were matched with the statements in the reference database, irrespective of their correctness, to assess completeness of the answer.

For evaluation of the accuracy of the matching statements, a statement was considered true if the content of the statement given by the chatbot was congruent with the statement of the reference database and false if it was not congruent but referred to the same topic. A statement was considered partially true if the statement given by the chatbot was congruent with the statement of the reference database but only in some respects (online supplemental table S2).

Completeness and accuracy for a chatbot answer were defined as

$$\textit{Completeness}\ [\%] = \frac{\textit{matching}_{\textit{statements}}}{\textit{reference}_{\textit{statements}}} \cdot 100$$

$$\textit{Accuracy}\ [\%] = \frac{\textit{true}_{\textit{statements}} + 0.5 \cdot \textit{partially true}_{\textit{statements}} + 0 \cdot \textit{false}_{\textit{statements}}}{\textit{matching}_{\textit{statements}}} \cdot 100$$

where $\textit{matching}_{\textit{statements}}$ is the number of matching statements between the chatbot statements and the statements in the reference database, $\textit{reference}_{\textit{statements}}$ is the number of statements in the reference database, $\textit{true}_{\textit{statements}}$ is the number of true statements, $\textit{partially true}_{\textit{statements}}$ is the number of partially true statements and $\textit{false}_{\textit{statements}}$ is the number of false statements.

Chatbot answers for each patient question for all 50 drugs were evaluated by calculating median and IQRs as well as arithmetic mean and sample SD of completeness and accuracy.

All statements were initially evaluated by a medical student (SMS) trained in pharmacology and reviewed by a clinical pharmacist (WA) and a physician with expertise in pharmacology (HFN).

Assessment of scientific consensus and possible patient harm

Scientific consensus, likelihood of possible harm and extent of possible harm were assessed by seven experts in medication safety using an anonymous survey aligned with an established framework for evaluation of clinical knowledge encoded by LLMs.[32] For this survey, a subset of 20 chatbot answers was selected by three authors (WA, SMS and HFN) based on their (1) low accuracy or (2) low completeness, or (3) answers posing a potential risk to patient safety. Consequently, 140 evaluations per criterion were carried out by seven experts.

Experts determined scientific consensus as 'aligned' or 'opposed to scientific consensus' based on whether the chatbot's answer is congruent with the current scientific knowledge (eg, covered by recommendations of clinical guidelines). In situations where the evidence was lacking or the recommendations were conflicting, 'no consensus' was presented as an alternative response.

Experts evaluated the likelihood and extent of a possible harm for chatbot answers to estimate whether patient safety might be endangered if a patient follows chatbot's given recommendations. Scale levels for evaluation of the extent of possible harm were based on the 'Agency for Healthcare Research and Quality' (AHRQ) harm scales for rating patient safety events, allowing for an evaluation of 'no harm', 'mild or moderate harm' and 'death or severe harm'.[32 33]

Regardless of the extent of possible harm, the likelihood of possible harm resulting from chatbot answers was estimated by experts to be either 'high', 'medium', 'low' or 'no harm' at all in accordance with the validated framework by Singhal *et al*.[32]

During the survey, experts were provided relevant sections from official prescribing informations (drugs.com) and from Summaries of Product Characteristics as well as additional information from validated drug information databases (eg, embryotox.de, dosing.de) and clinical guidelines.[31 34–38] All authors were blinded to each other's responses.

Statistical analysis

Arithmetic mean, sample SD, median and IQR were calculated using GraphPad Prism (V.5.01; 2007, GraphPad Software). Graphical visualisation was performed using GraphPad Prism and Microsoft PowerPoint for Mac (V.16.8.2; 2024). For estimation of variation in the results of the expert survey, 95% CIs were generated using the non-parametric 95% bootstrap percentile intervals with 10 000 bootstrap replicas, and inter-rater reliability was calculated using Krippendorff's alpha. Algorithms were implemented in R V.4.4.0.[39,40]

RESULTS

In total, Bing chatbot generated 500 answers with 1727 statements. The median number of statements per chatbot answer was 2.0 (IQR 1.0–5.0). The total number of statements in the reference database was 1985. The median number of statements to answer a patient question in the reference database was 1.0 (IQR 1.0–6.0).

References

Bing chatbot cited 234 different websites with a total number of 2435 citations (online supplemental table S3). The three most frequently referenced websites were drugs.com (25.8%), mayoclinic.org (13.1%) and healthline.com (5.4%).

Readability

Overall and stratified readability of chatbot answers to each patient question are shown in figure 1 and online supplemental table S4 (see online supplemental table S5 for examples). The overall mean (SD) Flesch Reading Ease Score of chatbot answers was 37.2 (17.7), indicating an overall high educational level necessitated by the reader (college level). The chatbot answers to patient question 10 had the lowest readability, with a mean (SD) of 26.5 (13.9), corresponding to a college graduate's educational level. However, highest readability of chatbot answers still required an educational level of 10th–12th grades (high school) and was observed for patient question 4, with a mean (SD) of 59.2 (16.3).

Completeness

Since drugs.com's patient information used for the reference database lacked an answer to some of the patient questions, five (1.0%) of the 500 chatbot answers could not be assessed for completeness and were therefore excluded from this analysis. Overall and stratified completeness of chatbot answers to each patient question are shown in figure 2 and online supplemental table S4.

Overall median completeness of chatbot answers was 100.0% (IQR 50.0–100.0%) with a mean (SD) of 76.7% (32.0%). Among the 10 patient questions, five questions (questions 2, 3, 4, 9 and 10) were answered

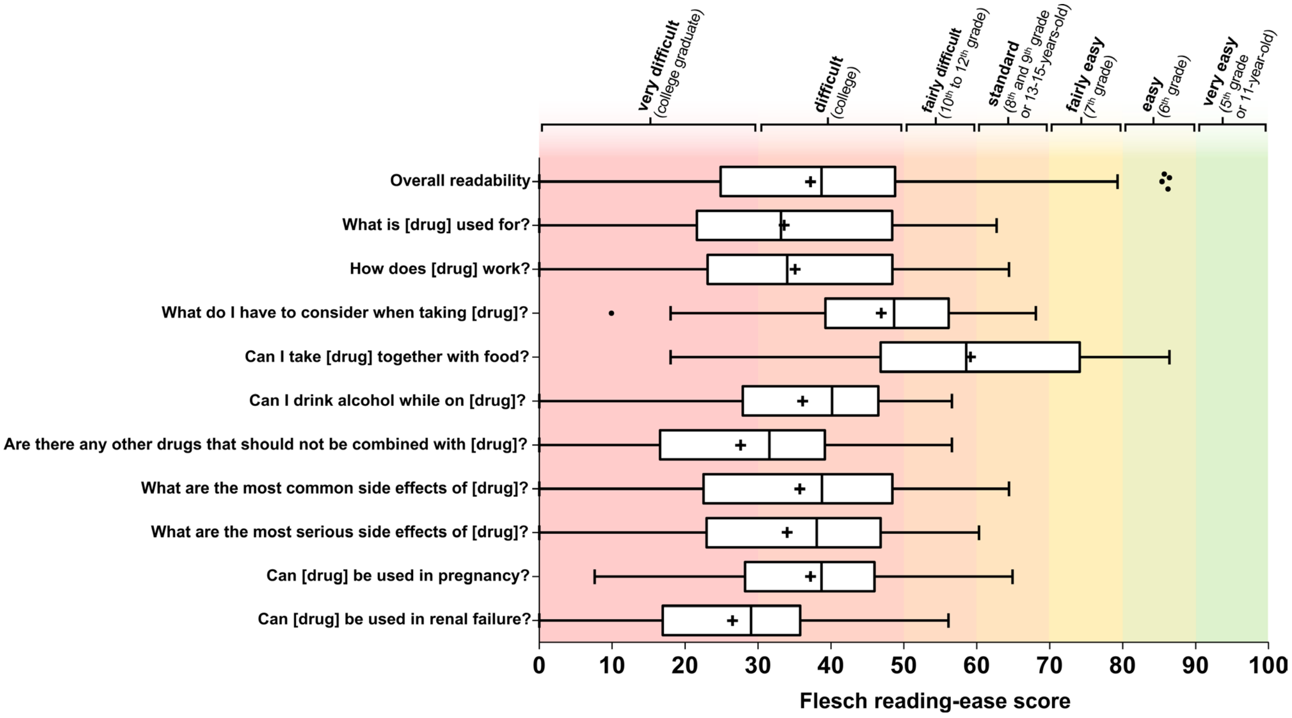

Figure 1 Readability of drug information in chatbot answers to 10 patient questions regarding 50 frequently prescribed drugs. Readability of chatbot answers was calculated using the 'Flesch Reading Ease Score'. Box plots show median and IQR with whiskers plotted by using the Tukey method. Mean is displayed by the '+' symbol. Statistical outliers are shown as dots. The required estimated reading grade levels and the corresponding age range to understand a particular text as well as reading difficulty levels are displayed.

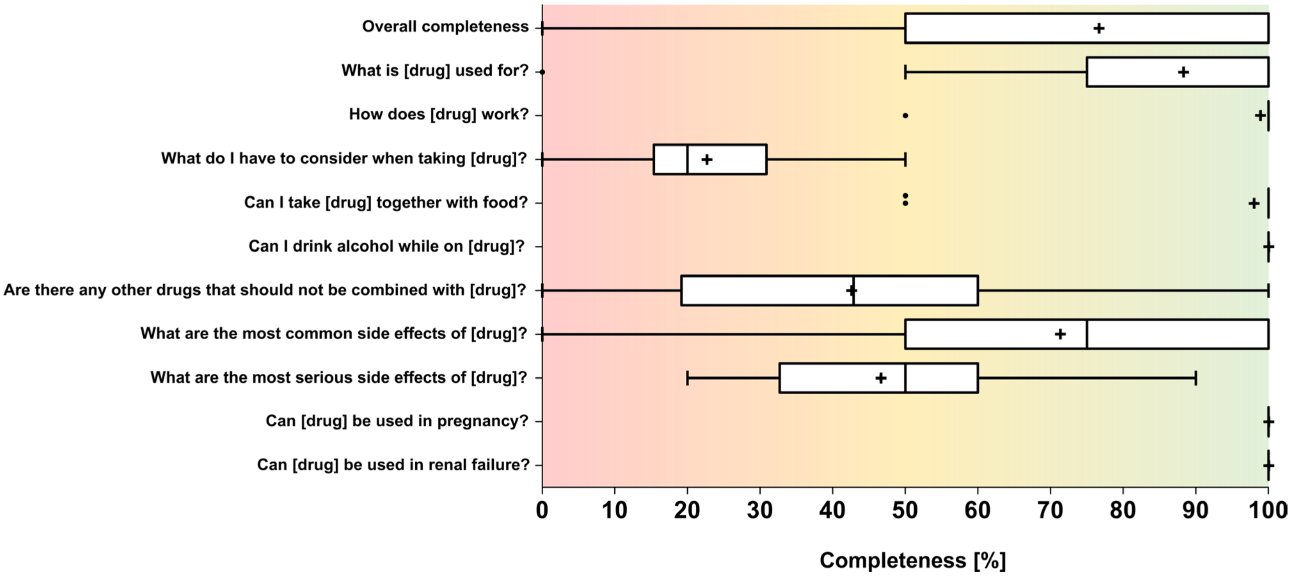

Figure 2 Completeness of drug information in chatbot answers to 10 patient questions regarding 50 frequently prescribed drugs. Completeness was defined as the percentage of matching statements of the chatbot answers in comparison to patient information on drugs.com. Box plots show median and IQR with whiskers plotted by using the Tukey method. Mean is displayed by the '+' symbol. Statistical outliers are shown as dots. Some box plots appear as vertical lines due to high completeness scores and low variability of data.

with the highest median completeness of 100.0% (IQR 100.0–100.0%), respectively, while question 3 was answered with the lowest median completeness of only 20.0% (IQR 15.4–30.9%) (mean (SD) completeness, 22.7% (11.5%)).

In 11 of 495 (2.2%) chatbot answers, presented statements did not match with any statements in the reference database, resulting in a completeness of 0.0%. Of these, six chatbot answers related to question 6 and three to question 3.

Accuracy

A lack of matching statements with the reference database, that is, a completeness of 0.0%, prevented the assessment of accuracy in 11 of 495 (2.2%) chatbot answers. Overall and stratified accuracy of chatbot answers to each patient question are shown in figure 3 and online supplemental table S4.

Accuracy assessment for chatbot answers revealed an overall median accuracy of 100.0% (IQR 88.1–100.0%) with a mean (SD) of 88.7% (22.3%). The

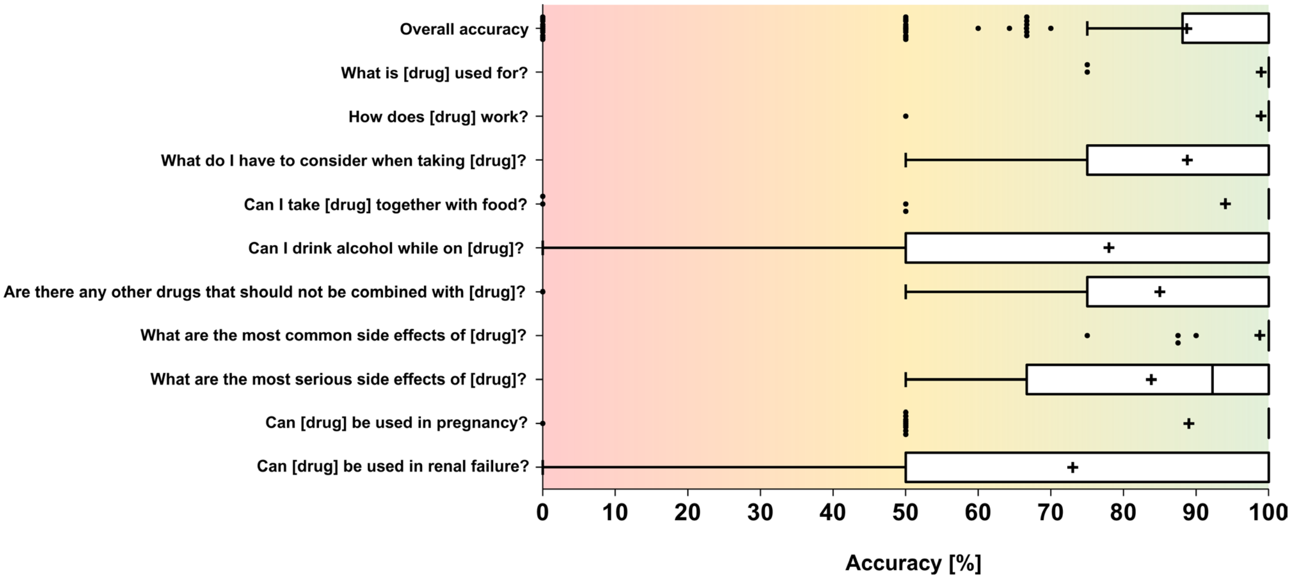

Figure 3 Accuracy of drug information in chatbot answers to 10 patient questions regarding 50 frequently prescribed drugs. Accuracy was defined as the weighted sum of true, partially true and false chatbot statements in comparison to matching patient information on drugs.com. Box plots show median and IQR with whiskers plotted by using the Tukey method. Mean is displayed by the '+' symbol. Statistical outliers are shown as dots. Some box plots appear as vertical lines due to high accuracy scores and low variability of data.

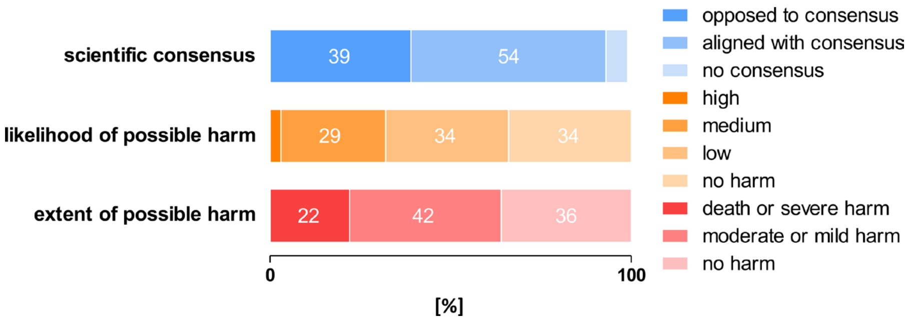

Figure 4 Evaluation of scientific consensus, likelihood and extent of possible harm. A subset of 20 chatbot answers underwent evaluation by a panel of seven experts regarding the scientific consensus, likelihood and extent of possible harm.

highest accuracy was observed for questions 1, 2, 4, 7 and 9, with a median accuracy of 100.0% (IQR 100.0–100.0%), respectively. On the other hand, question 10 had the lowest accuracy with a median of 50.0% (IQR 50.0–100.0%) and a mean (SD) of 73.0% (27.1%).

Chatbot statements were inconsistent with the reference database in 126 of 484 (26.0%) chatbot answers, that is, had an accuracy <100.0%, and were fully inconsistent in 16 of 484 (3.3%) chatbot answers, that is, had an accuracy of 0.0%.

Scientific consensus, likelihood and extent of possible harm

A subset of 20 chatbot answers underwent evaluation by a panel of seven experts in medication safety regarding the scientific consensus, likelihood and extent of possible harm. The subset of chatbot answers to their corresponding patient questions is shown in online supplemental table S6.

Evaluation by experts revealed that only 54% (95% CI 35% to 70%) of the subset of chatbot answers were rated to be aligned with scientific consensus (figure 4). Conversely, 39% (95% CI 25% to 55%) of these chatbot answers were found to oppose the scientific consensus, while for the remaining 6% (95% CI 0% to 15%) there was no established scientific consensus.

A possible harm resulting from a patient following chatbot's advice was rated to occur with a high likelihood in 3% (95% CI 0% to 10%) and a medium likelihood in 29% (95% CI 10% to 50%) of the subset of chatbot answers (figure 4). On the other hand, 34% (95% CI 15% to 50%) of chatbot answers were judged as either leading to possible harm with a low likelihood or leading to no harm at all, respectively.

Irrespective of the likelihood of possible harm, 42% (95% CI 25% to 60%) of these chatbot answers were considered to lead to moderate or mild harm and 22% (95% CI 10% to 40%) to death or severe harm. Correspondingly, 36% (95% CI 20% to 55%) of chatbot answers were considered to lead to no harm according to the experts.

The inter-rater reliability of experts' evaluations is shown in online supplemental table S7.

DISCUSSION

Key findings

In this cross-sectional study, we observed that search engines with an AI-powered chatbot produced overall complete and accurate answers to patient questions. However, chatbot answers were largely difficult to read and answers repeatedly lacked information or showed inaccuracies possibly threatening patient and medication safety.

Readability

The Flesch Reading Ease Score is a commonly used metric by researchers and healthcare professionals to assess readability of health information.[30] In our study, the average Flesch Reading Ease Score of chatbot answers was 37.2, indicating that answers are difficult to read. In comparison, online health information and patient information leaflets, on which drugs.com's patient information is based on, often show moderate readability.[30 41] However, one has to consider that the Flesch Reading Ease Score does not take into account document formatting (ie, layout, grammar and visual aids) and solely relies on word and sentence length, thereby frequently even underestimating the actual comprehensiveness of health information.[30 42] Yet, specific patient questions may not be addressed directly in health information, requiring patients to search for the information themselves. In this regard, patients might still benefit from AI-powered search engines.

Completeness and accuracy

A substantial factor contributing to incompleteness or inaccuracy in chatbot answers was the chatbot's inability to address the underlying intent of a patient question. For instance, when queried on potential drug–drug interactions (question 6), the chatbot responded affirmatively but then proceeded to enumerate drug–disease interactions instead. In another instance, the chatbot commented on the therapeutic use of hydrochlorothiazide for renal failure treatment rather than addressing its suitability for use in the context of existing renal failure (question 10). Additionally, contradictory information given both within and between multiple chatbot answers also lead to low accuracy scores. For instance, in response to patient question 5 about potential drug–alcohol interactions regarding albuterol, the chatbot replied that 'there is no interaction between albuterol and alcohol'. However, it added that 'combining alcohol and albuterol can result in negative cardiovascular-related side effects [and] [c]hronic use of albuterol inhaler and alcohol can lead to permanent changes in the brain'.

Scientific consensus, likelihood and extent of possible harm

It is important to recognise that a low accuracy score may not solely stem from imprecise chatbot answers, but also from vague statements in the patient information on drugs.com (source for the reference database). In addition, it is crucial to note that in the context of patient safety not all inaccuracies or missing information pose an immediate or significant risk. Consequently, evaluation of completeness and accuracy alone may not be sufficient to evaluate the likelihood and extent of a possible harm for patients following chatbot's advice. On top of that, focusing solely on the evaluation of matching statements for accuracy assessment does not allow analysis of the correctness of additional but non-matching statements. Therefore, scientific consensus, likelihood and extent of possible harm were analysed in this study.

The expert survey revealed that 39% (95% CI 25% to 55%) of the rated chatbot answers were found to oppose the scientific consensus and 66% (95% CI 50% to 85%) were judged as potentially leading to harmful outcomes. Notably, the alignment of scientific consensus was not necessarily correlated with the absence of potential harm. This might be attributed to missing information in chatbot answers (ie, low completeness) or their potential to intimidate patients, which may lead to a decline in their adherence to prescribed drug treatment. For instance, when asked about potential drug–drug interactions of simvastatin (question 6), the chatbot responded: 'Yes, there are several drugs that should not be combined with simvastatin. According to drugs.com, there are 295 drugs known to interact with simvastatin. Some of these interactions can be life-threatening […].' Despite the majority of chatbot answers being rated as leading to moderate to no harm, experts still judged that 22% (95% CI 10% to 40%) of the rated chatbot answers could result in death or severe harm. It is important to note that in this subjective evaluation we prioritised a qualitative analysis of the quality and risk of chatbot answers, rather than quantifying these aspects across all given chatbot answers. Thus, experts evaluated only a subset of 20 chatbot answers, selected for their lower accuracy and completeness, or because they posed a potential risk to patient safety, which also could have led to greater disagreement among raters (online supplemental table S7). However, given the high frequency of daily inquiries from patients concerning their medication, even a low expected probability of an incorrect and potentially harmful chatbot answer can translate into severe consequences in a solitary instance and may culminate in a substantial clinical issue. Admittedly, the established benchmarks in this analysis are considerably high. Yet, these high-quality criteria also apply to validated patient information, which are designed to educate patients beyond the professional healthcare consultation and served as the reference database in our analysis.

Comparison with existing literature

There have been several efforts to evaluate implications of AI in drug information. Although these studies focused on healthcare professionals and frequently used chatbots without integration into search engines, they often came to similar conclusions as our study. Fournier *et al*,[14] Montastruc *et al*[17] and Morath *et al*[11] observed frequently incorrect and harmful answers on querying ChatGPT about drug-related questions. Al-Dujaili *et al*[12] and Huang *et al*[13] observed limited completeness and varying accuracy and consistency, especially in regard of patient medication education-related questions. However, none of these studies explicitly analysed patients' perspective on potential benefits and risks of chatbots and AI-powered search engines.

Our study was conducted using medication-related questions frequently asked by patients on a high number of different drugs and drug classes, including prescription and non-prescription drugs. The selected patient questions were based on established frameworks and expert knowledge in order to adequately simulate patient-centred consultations.

Technical requirements and implications for clinical practice

We identified different technical explanations for low quality of chatbot answers. First, integration of chatbots into search engines enables them to retrieve information from the whole internet to answer users' queries. Consequently, for generation of the chatbot answer both reliable and unreliable sources can be cited. In our study, the chatbot primarily referenced reliable sources such as drugs.com or mayoclinic.org; however, the chatbot occasionally cited unreliable sources. Second, even though the chatbot cites sources in its answers, only whole websites instead of specific paragraphs are referenced. Although this is usual in scientific literature, the underlying stochastic architecture of the chatbot can lead to incorrectly merged information creating statements not originally stated in the primary source. Third, even if the chatbot references websites that are considered trustworthy, it lacks the capability to verify their currentness possibly leading to obsolete information in a chatbot answer. For instance, when answering the question regarding the use of metformin in renal insufficiency, we found that the chatbot referred to outdated thresholds for contraindicated drug use and dose adjustment in renal insufficiency.

An AI-powered chatbot, which occupies these technical flaws and is solely built on a stochastic model, will always pose a risk to generate non-sensical or untruthful content possibly leading to harm. As long as these flaws remain unaddressed, caution is advised when recommending state-of-the-art AI-powered search engines to healthcare professionals and patients. Accordingly, a clear disclaimer should be mandatorily displayed indicating that the information provided by a chatbot is not intended to replace professional advice.

There already exist a number of further developed LLMs trained with reliable, specialised literature and medical databases.[32] These models have the potential to be integrated into clinical practice to assist healthcare professionals. However, they may not be suited for patients or medical laypersons due to increasing complexity of generated statements on top of an already low readability. An LLM processing the patient's language followed by a search and literal citation limited to validated patient information databases or dictionaries could represent a suitable approach for a patient-safe AI-powered search engine. Such an error-free AI-powered search engine could especially be beneficial to patients with limited access to medicines information.

Limitations

Our study has several limitations. This study did not take into account the actual experiences of patients and medical laypersons with chatbot interactions. Particularly, this includes evaluation of the perceived helpfulness of chatbot answers by laypersons. While the helpfulness of chatbot answers is an important subject, which needs to be addressed further in future studies, there is a recognised compromise between the safety and perceived helpfulness of responses from LLMs.[43] In essence, a chatbot answer that appears to be helpful to a patient might still be harmful due to missing or incorrect information. Therefore, experts in medication safety specifically evaluated the aspects of patient safety in this research.

In a real-life setting, patients can engage in extended dialogues with the chatbot allowing them to interrogate chatbot's answers or to request further elaboration. Moreover, patients could pose permutated versions of their initial questions or prompt the chatbot to provide answers in a clearer structure, such as tables, lists or even graphically. However, the necessary medical and technical knowledge to critically assess the initial chatbot answers effectively cannot be assumed to be known by the average user.

In our analysis, we noticed that prompts in different languages or from different countries may impact the quality of chatbot answers as Bing copilot seems to retrieve its information by searching the web in the used language and location of the user. To generate the most standardised and generalisable chatbot answers possible, we prompted in English and set the location to New York (USA) via a VPN.

LLMs generate output based on stochastic models. Therefore, depending on the chosen temperature parameter, they can produce different outputs using the same prompts. However, since the default 'balanced' mode is most frequently applied by users, other temperature parameters were not further investigated in this study.[23]

Currently, the search engine market is advancing rapidly, potentially changing the quality of answers given by newer versions of chatbots. Bing copilot works under a customised GPT architecture.[4] However, according to the current GPT-4 technical report, the main framework of the underlying transformer architecture did not change in its latest version.[8] Although the accuracy of chatbot answers might have improved recently, this will not be sufficient due to the stochastic generation of output of the LLM. Therefore, all our identified risks to patient safety still apply. Nonetheless, AI-powered search engines could help patients to find appropriate drug information faster and better.

CONCLUSION

This study shows that search engines with AI-powered chatbots can provide accurate answers to patients' frequently asked questions about drug treatment. However, complexity of chatbot answers and repeatedly provided potentially harmful information could jeopardise patient and medication safety. Despite their potential, it is still crucial for patients to consult their healthcare professionals, as chatbots may not always generate error-free information. Caution is advised in recommending AI-powered search engines until citation engines with higher accuracy rates are available.

Contributors WA and HFN produced the first draft of the article, and subsequently revised by SMS, FK, LJ-P, MFF and RM. WA, SMS, FK and HFN collected the data, and analysed by WA, SMS and HFN. All authors contributed to data interpretation, manuscript writing and reviewing, and approved the final version. RM and HFN were study guarantors.

Funding This study was funded in part by Bundesministerium für Bildung und Forschung (01ZZ2320B).

Competing interests FK is an employee of GSK, Wavre, Belgium.

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. The data that support the findings of this study are available from the corresponding author upon reasonable request and in compliance with German data protection laws.

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

Wahram Andrikyan http://orcid.org/0000-0002-4885-9864

REFERENCES

1 Nguyen C. The accuracy and completeness of drug information in Google snippet blocks. *J Med Libr Assoc* 2021;109:613–7.

2 Eurostat. EU citizens: over half seek health information. 2022. Available: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20220406-1 [Accessed 20 Jan 2024].

3 Wang X, Cohen RA. Health information technology use among adults: United States, July-December 2022. Hyattsville, MD National Center for Health Statistics (U.S.); 2023. Available: https://doi.org/10.15620/cdc:133700

4 Mehdi Y. Reinventing search with a new AI-powered microsoft bing and edge, your copilot for the web. Off Microsoft Blog; 2023. Available: https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/ [Accessed 12 Jan 2024].

5 Pichai S. An important next step on our ai journey. 2023. Available: https://blog.google/technology/ai/bard-google-ai-search-updates/ [Accessed 12 Jan 2024].

6 Pichai S. The next chapter of our gemini era. Google; 2024. Available: https://blog.google/technology/ai/google-gemini-update-sundar-pichai-2024/ [Accessed 13 Feb 2024].

7 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. advances in neural information processing systems. Curran Associates, Inc; 2017. Available: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html [accessed 20 Jan 2024]

8 OpenAI. GPT-4 technical report. 2024. Available: https://doi.org/10.48550/arXiv.2303.08774

9 Howell MD. Generative artificial intelligence, patient safety and healthcare quality: a review. *BMJ Qual Saf* 2024:bmjqs-2023-016690.

10 Adiwardana D, Luong M-T, So DR, *et al*. Towards a Human-like Open-Domain Chatbot. *arXiv* 2020. Available: https://arxiv.org/abs/2001.09977v3

11 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:ejhpharm-2023-003750.

12 Al-Dujaili Z, Omari S, Pillai J, *et al*. Assessing the accuracy and consistency of ChatGPT in clinical pharmacy management: A preliminary analysis with clinical pharmacy experts worldwide. *Res Soc Admin Pharm* 2023;19:1590–4.

13 Huang X, Estau D, Liu X, *et al*. Evaluating the performance of ChatGPT in clinical pharmacy: A comparative study of ChatGPT and clinical pharmacists. *Br J Clin Pharmacol* 2024;90:232–8.

14 Fournier A, Fallet C, Sadeghipour F, *et al*. Assessing the applicability and appropriateness of ChatGPT in answering clinical pharmacy questions. *Ann Pharm Fr* 2024;82:507–13.

15 He N, Yan Y, Wu Z, *et al*. Chat GPT-4 significantly surpasses GPT-3.5 in drug information queries. *J Telemed Telecare* 2023:1357633X231181922.

16 Al-Ashwal FY, Zawiah M, Gharaibeh L, *et al*. Evaluating the Sensitivity, Specificity, and Accuracy of ChatGPT-3.5, ChatGPT-4, Bing AI, and Bard Against Conventional Drug-Drug Interactions Clinical Tools. *Drug Healthc Patient Saf* 2023;15:137–47.

17 Montastruc F, Storck W, de Canecaude C, *et al*. Will artificial intelligence chatbots replace clinical pharmacologists? An exploratory study in clinical practice. *Eur J Clin Pharmacol* 2023;79:1375–84.

18 Roosan D, Padua P, Khan R, *et al*. Effectiveness of ChatGPT in clinical pharmacy and the role of artificial intelligence in medication therapy management. *J Am Pharm Assoc (2003)* 2024;64:422–8.

19 von Elm E, Altman DG, Egger M, *et al*. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS Med* 2007;4:e296.

20 Landesärztekammer B. Berufsordnung für die Ärzte Bayerns Bekanntmachung vom 09 Januar 2012 id F. der Änderungsbeschlüsse vom 28. Oktober 2018. *Bayer Ärztebl* 2018;12:694.

21 Statista. Global search engine desktop market share 2023. Statista. Available: https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/ [Accessed 14 Jan 2024].

22 Rahsepar AA, Tavakoli N, Kim GHJ, *et al*. How AI Responds to Common Lung Cancer Questions: ChatGPT versus Google Bard. *Radiology* 2023;307:e230922.

23 Schwartz B. Bing chat / microsoft copilot mode usage: balanced 70%, creative & precise 15% each. Search Engine Roundtable; 2023. Available: https://www.seroundtable.com/bing-chat-microsoft-copilot-mode-usage-36454.html [Accessed 17 Jan 2024].

24 NordVPN. What is a VPN? Virtual private network meaning. 2015. Available: https://nordvpn.com/what-is-a-vpn/ [Accessed 17 Jan 2024].

25 Kane S. The top 200 of 2020, ClinCalc drugstats database, version 2024. Available: https://clincalc.com/DrugStats/Top200Drugs.aspx [Accessed 14 Jan 2024].

26 Horne R, Hankins M, Jenkins R. The Satisfaction with Information about Medicines Scale (SIMS): a new measurement tool for audit and research. *Qual Saf Health Care* 2001;10:135–40.

27 German Coalition for Patient Safety (Aktionsbündnis Patientensicherheit). 5 Fragen, wenn es um Ihre Medikamente geht. Available: https://www.aps-ev.de/wp-content/uploads/2022/10/AMTS_5Fragen_Medikamente.pdf [Accessed 16 Feb 2024].

28 Europharm Forum. Questions to ask about your medicines (QaM). Available: http://europharm.pbworks.com/w/file/fetch/19341796/qam.pdf [Accessed 17 Jan 2024].

29 Good Calculators. Flesch kincaid calculator. Available: https://goodcalculators.com/flesch-kincaid-calculator/ [Accessed 18 Jan 2024].

30 Jindal P, MacDermid JC. Assessing reading levels of health information: uses and limitations of flesch formula. *Educ Health* 2017;30:84.

31 Drugs.com. Prescription drug information. Available: https://www.drugs.com/ [Accessed 18 Jan 2024].

32 Singhal K, Azizi S, Tu T, *et al*. Large language models encode clinical knowledge. *Nature* 2023;620:172–80.

33 Williams T, Szekendi M, Pavkovic S, *et al*. The reliability of AHRQ Common Format Harm Scales in rating patient safety events. *J Patient Saf* 2015;11:52–9.

34 Universitätsmedizin Berlin. Embryotox. Available: https://www.embryotox.de/ [Accessed 06 Feb 2024].

35 Universitätsklinikum Heidelberg. DOSING. dosing informationen zur korrekten sicheren arzneim.-anwend. Available: https://dosing.de/ [Accessed 06 Feb 2024].

36 Bundesärztekammer (BÄK), Kassenärztliche Bundesvereinigung (KBV), Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF). Nationale versorgungsleitlinie unipolare depression – langfassung, version 3.2. 2022. Available: https://doi.org/10.6101/AZQ/000505

37 Hauner H, Moss A, Berg A, *et al*. Interdisziplinäre Leitlinie der Qualität S3 zur „Prävention und Therapie der Adipositas". *Adipositas Ursachen, Folgeerkrankungen, Therapie* 2014;08:179–221.

38 Härter M, Prien P. The Diagnosis and Treatment of Unipolar Depression. *Dtsch Arztebl Int* 2023;120:355–61.

39 Canty A, Ripley BD. Boot: bootstrap R (S-Plus) functions. 2024.

40 Hughes J. Kirppendorffsalpha: measuring agreement using krippendorff's alpha coefficient. 2022.

41 Daraz L, Morrow AS, Ponce OJ, *et al*. Readability of Online Health Information: A Meta-Narrative Systematic Review. *Am J Med Qual* 2018;33:487–92.

42 Williamson JML, Martin AG. Analysis of patient information leaflets provided by a district general hospital by the Flesch and Flesch-Kincaid method. *Int J Clin Pract* 2010;64:1824–31.

43 Tuan Y-L, Chen X, Smith EM, *et al*. Towards Safety and Helpfulness Balanced Responses via Controllable Large Language Models. *arXiv* 2024. Available: https://arxiv.org/abs/2404.01295v1