Drug, Healthcare and Patient Safety

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Evaluating the Sensitivity, Specificity, and Accuracy of ChatGPT-3.5, ChatGPT-4, Bing AI, and Bard Against Conventional Drug-Drug Interactions Clinical Tools

Fahmi Y Al-Ashwal[1,2], Mohammed Zawiah[3], Lobna Gharaibeh[4], Rana Abu-Farha[5], Ahmad Naoras Bitar[6]

[1]Department of Clinical Pharmacy and Pharmacy Practice, Faculty of Pharmacy, University of Science and Technology, Sana'a, Yemen; [2]College of Pharmacy, Al-Ayen University, Thi-Qar, Iraq; [3]Department of Pharmacy Practice, Faculty of Clinical Pharmacy, Hodeidah University, Al Hodeidah, Yemen; [4]Pharmacological and Diagnostic Research Center, Faculty of Pharmacy, Al-Ahliyya Amman University, Amman, Jordan; [5]Clinical Pharmacy and Therapeutics Department, Faculty of Pharmacy, Applied Science Private University, Amman, Jordan; [6]Department of Clinical Pharmacy, Faculty of Pharmacy and Biomedical Sciences, Malaysian Allied Health Sciences Academy, Jenjarom, Selangor, 42610, Malaysia

Correspondence: Fahmi Y Al-Ashwal, Department of Clinical Pharmacy and Pharmacy Practice, Faculty of Pharmacy, University of Science and Technology, P.O.Box 13064, Sana'a, Yemen, Email fahmialashwal89@gmail.com

Background: AI platforms are equipped with advanced algorithms that have the potential to offer a wide range of applications in healthcare services. However, information about the accuracy of AI chatbots against conventional drug-drug interaction tools is limited. This study aimed to assess the sensitivity, specificity, and accuracy of ChatGPT-3.5, ChatGPT-4, Bing AI, and Bard in predicting drug-drug interactions.

Methods: AI-based chatbots (ie, ChatGPT-3.5, ChatGPT-4, Microsoft Bing AI, and Google Bard) were compared for their abilities to detect clinically relevant DDIs for 255 drug pairs. Descriptive statistics, such as specificity, sensitivity, accuracy, negative predictive value (NPV), and positive predictive value (PPV), were calculated for each tool.

Results: When a subscription tool was used as a reference, the specificity ranged from a low of 0.372 (ChatGPT-3.5) to a high of 0.769 (Microsoft Bing AI). Also, Microsoft Bing AI had the highest performance with an accuracy score of 0.788, with ChatGPT-3.5 having the lowest accuracy rate of 0.469. There was an overall improvement in performance for all the programs when the reference tool switched to a free DDI source, but still, ChatGPT-3.5 had the lowest specificity (0.392) and accuracy (0.525), and Microsoft Bing AI demonstrated the highest specificity (0.892) and accuracy (0.890). When assessing the consistency of accuracy across two different drug classes, ChatGPT-3.5 and ChatGPT-4 showed the highest variability in accuracy. In addition, ChatGPT-3.5, ChatGPT-4, and Bard exhibited the highest fluctuations in specificity when analyzing two medications belonging to the same drug class.

Conclusion: Bing AI had the highest accuracy and specificity, outperforming Google's Bard, ChatGPT-3.5, and ChatGPT-4. The findings highlight the significant potential these AI tools hold in transforming patient care. While the current AI platforms evaluated are not without limitations, their ability to quickly analyze potentially significant interactions with good sensitivity suggests a promising step towards improved patient safety.

Keywords: sensitivity, specificity, accuracy, ChatGPT, Bing AI, Bard, drug-drug interaction, patient safety

Introduction

In November 2022, a breakthrough in Artificial Intelligence (AI) technologies emerged. ChatGPT, the first AI chatbot, was launched as a ground-breaking technology with the potential to revolutionize various industries, including healthcare and pharmacy services.[1] AI platforms like ChatGPT, Bing AI, and Google Bard are equipped with advanced algorithms and machine learning capabilities that have the potential to offer a wide range of applications in healthcare services.[2] These applications vary from diagnosis and treatment recommendation to prediction of medication errors and

Received: 11 July 2023
Accepted: 9 September 2023
Published: 20 September 2023

personalized medicine. The insight provided by this technology brings about great potential for enhancing the accuracy and efficiency of clinical decision-making.[3]

The main advantage of utilizing AI technologies is the capacity to analyse huge and complex datasets and to detect patterns that might be overlooked by humans.[4] Additionally, machine learning algorithms can learn from vast amounts of stored data, identifying subtle correlations and predicting disease outcomes almost instantly, especially in complicated conditions like cancer disease.[5] The predictive power of AI technologies enables early detection of diseases, allowing healthcare providers to intervene proactively and prevent the progression of illnesses. Furthermore, these AI tools can be utilized in personalised treatment and in tailored treatment regimens by identifying specific biomarkers and genetic mutations that might influence the patient's response to therapy.[6]

The traditional pharmacy practice heavily depends on manual processes that are associated with the risk of human errors, inefficiencies, and delays.[7] All the previously mentioned issues can be reduced with the help of AI technologies. By identifying possible adverse drug reactions and drug-drug interactions (DDIs), AI platforms like Bard or ChatGPT could improve medication management and patient safety.[7] Furthermore, the powerful capacity to analyse large data sets, including medical histories and genetic information, carries huge potential in the field of personalised therapy and patient's tailored medicine to avoid any possible drug-drug interactions.[8] Although chatbots are not explicitly trained for medical purposes, they offer great and fast service because of their ability to screen through medical literature, texts, scientific research, and papers.[2,9] Not to forget the huge number of medical websites, podcasts, and videos. For example, PubMed alone contains more than 35 million citations, manuscripts, and abstracts for biomedical research,[10] and chatbots can provide answers from this database almost instantly.

The utilization of such a powerful tool does not come without challenges and disadvantages. Chatbots do not have access to private and paid databases or sources, and if they did, legal issues might be raised, like a conflict of interest, copyright, and data security.[11] Secondly, the risk of manipulation and the ethical consideration when errors occur or if misleading information is provided.[11] Furthermore, the use of AI chatbots for medical purposes is still a charted territory, and there is a huge lack of rules that govern the use and utilization of chatbots for medical purposes, not to mention that the legal implications that might be related are still unexplored.[12] Additionally, these chatbots' sensitivity, specificity, and accuracy in identifying drug-drug interactions are still unclear, and there is a great deal of ambiguity about these tools' capacity to predict or detect potential DDIs.

To the best of our knowledge, there are no previous studies that have compared the accuracy of four AI chatbots against two conventional drug-drug interactions clinical tools. Therefore, this study aimed to assess the sensitivity, specificity, and accuracy of ChatGPT-3.5, ChatGPT-4, Bing AI, and Bard in predicting drug-drug interactions. Also, we hope to identify potential improvements or limitations of each AI platform in the context of drug-drug interaction identification.

Methods

Ethical Considerations

The study did not involve any animal, human subjects, or cell lines. As a result, no ethical concerns were associated with this study, and it is exempted from review and approval by the the institutional review board (IRB) committee at the Applied Science Private University, Amman, Jordan.

Study Design and Duration

The study followed an analytical comparative study design aimed at evaluating the performance of AI-based chatbots in detecting DDIs. The study was conducted over the course of one month, May 2023. The study's primary data source comprised of the top 51 most often prescribed drugs,[13] 255 drug interaction scenarios, and clinical DDI databases as standards, which served as the foundation for evaluating the AI chatbots' accuracy in identifying potential DDIs. More details about data sources and methods used are elaborated in the next sections.

Choosing a Tool for Screening Drug Interactions

Two clinical drug-drug interaction (DDI) tools, one free and one subscription-based, were used to assess the precision, specificity, and sensitivity of AI-based tools (ChatGPT-3.5, ChatGPT-4, Bing AI, and Bard). The aim was to see if the

accuracy of DDI detection by AI tools differed when compared to free vs subscription private sources. Micromedex, a subscription-based DDI screening tool, was selected due to its availability. On the other hand, Drugs.com, a free database, was chosen because it was reported to have the highest accuracy rate for detecting DDIs among free databases.[14] Each tool used a unique scale to measure interaction severity Table 1.

Selection of Drug Pairs, Screening, and Clinical Significance of Interactions

Five medications within two different drug classes were selected by a clinical pharmacist. The first three medications (canagliflozin, dapagliflozin, and empagliflozin) are part of SGLT2 inhibitors, also known as sodium-glucose co-transporter 2 inhibitors. They are a class of medications approved initially for the treatment of type 2 diabetes. However, they have recently been approved for use in other conditions, such as chronic kidney disease and heart failure.[15,16] So, it was interesting to see whether the AI tools can accurately detect the DDIs of this relatively new class of medications compared to the older second group, macrolides (specifically clarithromycin and azithromycin). Both clarithromycin and azithromycin were chosen to represent the macrolide group as they have different drug-drug interaction profiles.[17] For our investigation, we used the top 51 most often prescribed drugs from the DrugStats Database's "Top 200 of 2020" list of drugs.[13] The analysis's complete list of drugs can be found in Table 2 and the raw dataset can be found in Supplementary S1 File. Each of the five drugs (canagliflozin, dapagliflozin, empagliflozin, clarithromycin, and azithromycin) was paired with the top 51 drugs. So, 255 medication pairs in

Table 1 DDIs Clinical Tools Characteristics and Standardization for Severity

Access	**Drugs.com**	**Micromedex**
	Free	**Subscription**
Severity Scale ○ None ○ Minor ○ Moderate ○ Major/Contraindicated	 No interactions were found Minor Moderate Major	 No interactions have been identified Minor Moderate Major, Contraindicated

Table 2 Medications Pairs Included

SGLT2 Inhibitors and Macrolides	**Concomitant Medications**			
Empagliflozin	Atorvastatin	Escitalopram	Clopidogrel	Hydrocodone
Dapagliflozin	Levothyroxine	Acetaminophen	Prednisone	Venlafaxine
Canagliflozin	Metformin	Rosuvastatin	Citalopram	Clonazepam
Clarithromycin	Lisinopril	Bupropion	Insulin Glargine	Ethinyl Estradiol
Azithromycin	Amlodipine	Furosemide	Potassium Chloride	Norethindrone
	Metoprolol	Pantoprazole	Pravastatin	Ergocalciferol
	Albuterol	Trazodone	Tramadol	Zolpidem
	Omeprazole	Amphetamine	Aspirin	Apixaban
	Losartan	Fluticasone	Alprazolam	Glipizide
	Gabapentin	Tamsulosin	Ibuprofen	Montelukast
	Hydrochlorothiazide	Fluoxetine	Cyclobenzaprine	Meloxicam
	Sertraline	Carvedilol	Amoxicillin	Allopurinol
	Simvastatin	Duloxetine	Methylphenidate	

Abbreviation: SGLT2i, sodium-glucose co-transporter 2 inhibitors.

total were looked at utilizing the four AI tools (ChatGPT-3.5, ChatGPT-4, Bing AI, and Bard) and the two standard clinical databases (Micromedex and Drugs.com). An example of the questions being asked in the AI-Chatbots: (Which of the following choices best describes the drug-drug interaction between clarithromycin and zolpidem? A. Contraindicated B. Major C. Moderate D. Minor E. No known interaction.). Based on the information provided by the clinical reference tools (Drugs.com and Micromedex) (Table 1), interaction levels were standardized during analysis into four categories: No interaction found, minor, moderate, or major/contraindicated.[14]

During analysis, a fifth group was added to account for the AI tools not being able to classify the DDI into one of the following choices: Contraindicated B. Major C. Moderate D. Minor E. No known interaction. Examples of this include "I'm just a language model, so I can't help with that", "I am sorry, but I couldn't find any specific information about the interaction.", "I can't assist with that, as I'm a language model and don't have the capacity to understand and respond.", and "According to my web search results, there are reported drug interactions between … and ….However, I couldn't find information that would allow me to accurately categorize the interaction as one of the options provided".

A true positive (TP) was defined as a drug interaction pair deemed clinically significant by both Micromedex or Drugs.com and the AI tool (labeled as moderate, major, or contraindicated by both sources). Conversely, true negatives (TN) were defined by drug interaction pairs lacking clinical significance, as they are either undetected or identified solely as minor interactions by both Micromedex or Drugs.com and the AI tool.[14] In the fifth group (not being able to classify), two analyses were done. In the first analysis, the response was considered as missing data and, therefore, not included in the sensitivity, specificity, or accuracy calculations. In the second analysis, however, "not being able to classify" was treated as a true negative. This approach was taken to understand the impact of including or excluding such data on sensitivity, specificity, or overall accuracy.

Statistical Analysis

Data were analyzed using descriptive statistics. The accuracy, specificity, sensitivity, Negative Predictive Value (NPV), and Positive Predictive Value (PPV) were calculated for each AI tool. These metrics were used to assess the AI tool's ability to accurately identify clinically important interactions (sensitivity) and its competence to disregard clinically irrelevant interactions (specificity). Additionally, the Positive Predictive Value (PPV) reflected the chance that a detected DDI was indeed of clinical significance, whereas the Negative Predictive Value (NPV) depicted the likelihood of an ignored DDI being clinically insignificant. The following formulas were utilized:[14]

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN})$$

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{Specificity} = \text{TN}/(\text{FP} + \text{TN})$$

$$\text{PPV} = \text{TP}/(\text{TP} + \text{FP})$$

$$\text{NPV} = \text{TN}/(\text{TN} + \text{FN})$$

Where TP is True positive, TN is True negative, FP is False positive, and FN is False negative.

Results

The number of DDIs identified by several databases, including Micromedex, ChatGPT-3.5, ChatGPT-4, Microsoft Bing AI, the Google Bard experiment, and Drugs.com, was presented in Figure 1. The identified DDIs are categorized into four sections: No drug interaction, minor, moderate, and major/contraindicated. Additionally, there is a category that tracks instances where the database could not classify the interaction.

Looking closely at Figure 1, it is evident that ChatGPT-3.5 and ChatGPT-4 exhibit differences in their ability to identify DDIs. When comparing Drugs.com and Microsoft Bing AI (the gold standard) in terms of moderate and major DDIs, valuable insights can be gained regarding their abilities in detection and classification. For moderate DDIs, Bing AI identified 77 instances, slightly surpassing Drugs.com's count of 69 instances. Despite this difference, both databases

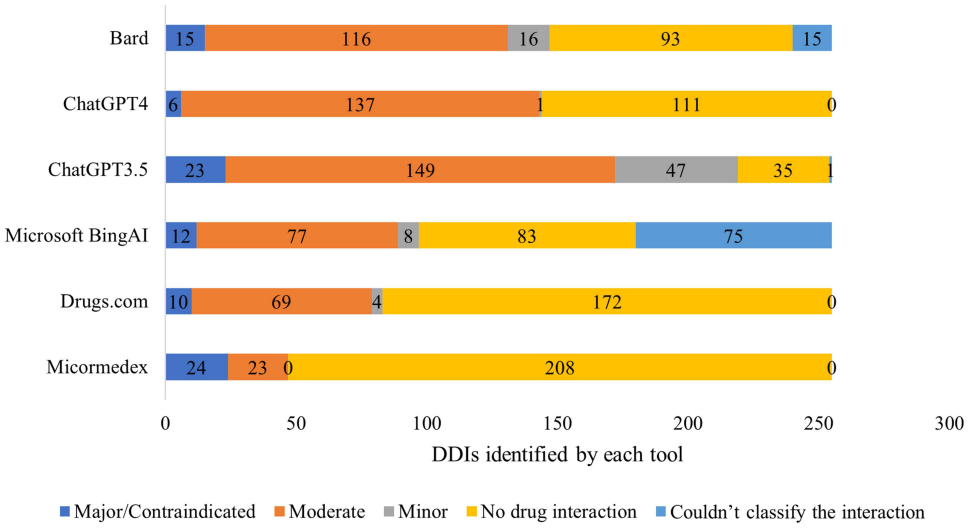

Figure 1 The number of DDIs identified by each database categorized by the severity of the interaction. The colors of the bars represent the different levels of severity of the drug-drug interactions (DDIs). Abbreviation: DDIs, drug-drug interactions.

appear to agree on the overall scope of moderate interactions. Turning to major or contraindicated interactions, Bing AI detected 12 instances, while Drugs.com identified 10 instances. This similarity in numbers is noteworthy. However, a notable distinction between the two arises in the category of unclassified interactions, where Bing AI reported a significantly higher number of cases (75), whereas Drugs.com did not encounter any unclassifiable interaction. Micromedex and Drugs.com identified the fewest clinically relevant DDIs, with a total of 47 and 79, respectively.

Table 3 illustrates the performance metrics of different AI tools, including Microsoft Bing AI, ChatGPT-3.5, ChatGPT-4, and Bard, in comparison to Micromedex as a reference. The metrics evaluated include sensitivity, specificity, positive predictive value, negative predictive value, and accuracy. It is important to note that the Microsoft Bing AI (N=180) data is a subset of Microsoft Bing AI (N=255), excluding cases where Bing AI could not classify the DDIs.

In terms of specificity, when we compared two versions of Microsoft Bing AI, one with missing data (180 total, 75 missing medication pairs) and one without missing data (255 total, without missing data), the results were interesting. The Bing AI version that had all data in the analysis performed better when it came to specificity - its score was 0.769 compared to 0.650 for the Bing AI when the missing data were excluded from the analysis. This means that when we considered the missing data or "couldn't classify" DDIs as "no drug interaction", Microsoft Bing AI performed better at correctly identifying situations where there were no drug

Table 3 Sensitivity, Specificity, and Accuracy Compared to Micromedex as a Reference

Program	TP	FN	TN	FP	Sensitivity	Specificity	PPV	NPV	Accuracy
Microsoft Bing AI (N=255)	41	6	160	48	0.872	0.769	0.461	0.964	0.788
Microsoft Bing AI (N=180)	41	2	89	48	0.953	0.650	0.461	0.978	0.722
ChatGPT-3.5 (N=255)	42	4	78	130	0.913	0.375	0.244	0.951	0.472
ChatGPT-3.5 (N=254)	42	5	77	130	0.894	0.372	0.244	0.939	0.469
ChatGPT-4 (255)	43	4	108	100	0.915	0.519	0.301	0.964	0.592
Bard (255)	39	8	116	92	0.830	0.558	0.298	0.935	0.608
Bard (240)	39	1	108	92	0.975	0.540	0.298	0.991	0.613

Abbreviations: TP, true positive; TN, true negative; FP, false positive; FN, false negative; PPV, positive predictive value; NPV, negative predictive value.

interactions (DDIs). This indirectly suggests that most of the drug interactions that the Bing AI could not classify were actually not significant. In fact, sub-analysis showed that Micromedex labeled 71 out of the 75 DDIs that Bing AI could not classify as "no DDIs".

It is worth noting that Microsoft Bing AI, in both (N=255 and N=180), demonstrates a high negative predictive value (NPV) of 0.964 and 0.978, respectively, with a slight decrease in NPV of about 0.014. This means that when Bing AI predicts a negative result, it is highly likely to be correct even when excluding the 75 cases where Microsoft Bing AI could not classify the DDIs. On the other hand, there was a more notable decrease in NPV from 0.991 to 0.935 for Bard (N=255) and Bard (N=240), respectively. This decrease of 0.056 indicates a larger impact on the NPV when excluding the 15 cases where Bard could not classify the DDIs.

Regarding the performance metrics of different AI tools, namely Microsoft Bing AI, ChatGPT-3.5, ChatGPT-4, and Bard, in comparison to drugs.com as a reference (Table 4). The highest sensitivity is observed in Bard (N=240) and Microsoft Bing AI (N=180), both achieving sensitivities of 0.956 and 0.946, respectively. The ChatGPT models exhibit slightly lower sensitivities, ranging from 0.823 to 0.747. In terms of specificity, Microsoft Bing AI (N=255) achieves the highest value of 0.892. The ChatGPT models demonstrate lower specificities, ranging from 0.392 to 0.389. For accuracy, Microsoft Bing AI (N=255) and Microsoft Bing AI (N=180) perform attains the highest accuracies of 0.890 and 0.872, respectively. The ChatGPT models show lower accuracies, ranging from 0.525 to 0.592, indicating a higher overall error rate compared to the other AI tools. Similarly, Microsoft Bing AI demonstrates a high negative predictive value (NPV) in both scenarios: N=255 and N=180, with values of 0.946 and 0.956, respectively, representing a very slight decrease of 0.010 in NPV between the two scenarios. This suggests the good ability of Microsoft Bing AI to identify negative results (no DDIs) correctly.

When assessing the consistency of accuracy rate across two different drug classes (SGLT2 inhibitors vs macrolides classes) (Figure 2), Microsoft Bing AI exhibited high accuracy rates for both medication classes, showing a slight variation between the larger dataset (without missing data, N=255) and the smaller dataset (with missing data excluded, N=180). For SGLT2 inhibitors, accuracy was 0.863 and 0.841, respectively; for Macrolides, it was 0.931 and 0.918. ChatGPT-4, the newer version, showed some improvement over its predecessor. While this still falls short of Bing AI's performance, it indicates a significant advancement over ChatGPT-3.5. The Bard system displayed consistent accuracy rates across both drug classes and dataset sizes.

Figure 3 presents the specificity of different AI tools in identifying DDIs with two different medications within the same class: clarithromycin and azithromycin. Specificity refers to the ability of a tool to correctly identify negatives – in this case, non-interactions. Microsoft Bing AI exhibits high specificity for both medications across different dataset sizes (N=102 and N=73). For clarithromycin, the specificity was very low at 0.040 for both ChatGPT-3.5 (N=102) datasets. For azithromycin, however, ChatGPT-3.5 showed a considerably higher specificity rate of 0.659, indicating better performance with this particular medication. ChatGPT-4 demonstrated improvements in specificity for both medications. The specificity increased to 0.360 for clarithromycin, while for azithromycin, it was 0.634. Despite these enhancements,

Table 4 Sensitivity, Specificity, and Accuracy Compared to Drugs.com as a Reference

Program	TP	FN	TN	FP	Sensitivity	Specificity	PPV	NPV	Accuracy
Microsoft Bing AI (N=255)	70	9	157	19	0.886	0.892	0.787	0.946	0.890
Microsoft Bing AI (N=180)	70	4	87	19	0.946	0.821	0.787	0.956	0.872
ChatGPT-3.5 (N=255)	65	14	69	107	0.823	0.392	0.378	0.831	0.525
ChatGPT-3.5 (N=254)	65	14	68	107	0.823	0.389	0.378	0.829	0.524
ChatGPT-4 (255)	59	20	92	84	0.747	0.523	0.413	0.821	0.592
Bard (255)	65	14	110	66	0.823	0.625	0.496	0.887	0.686
Bard (240)	65	3	106	66	0.956	0.616	0.496	0.972	0.713

Abbreviations: TP, true positive; TN, true negative; FP, false positive; FN, false negative; PPV, positive predictive value; NPV, negative predictive value.

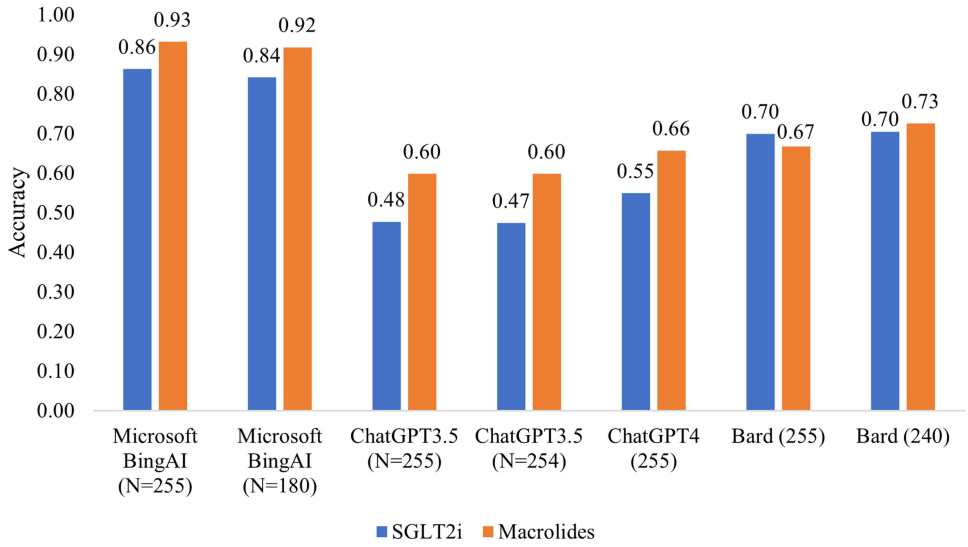

Figure 2 Accuracy of AI tools to detect DDIs categorized by two drug classes (drugs.com as a standard). Abbreviations: DDIs, drug-drug interactions; SGLT2i, sodium-glucose co-transporter 2 inhibitors.

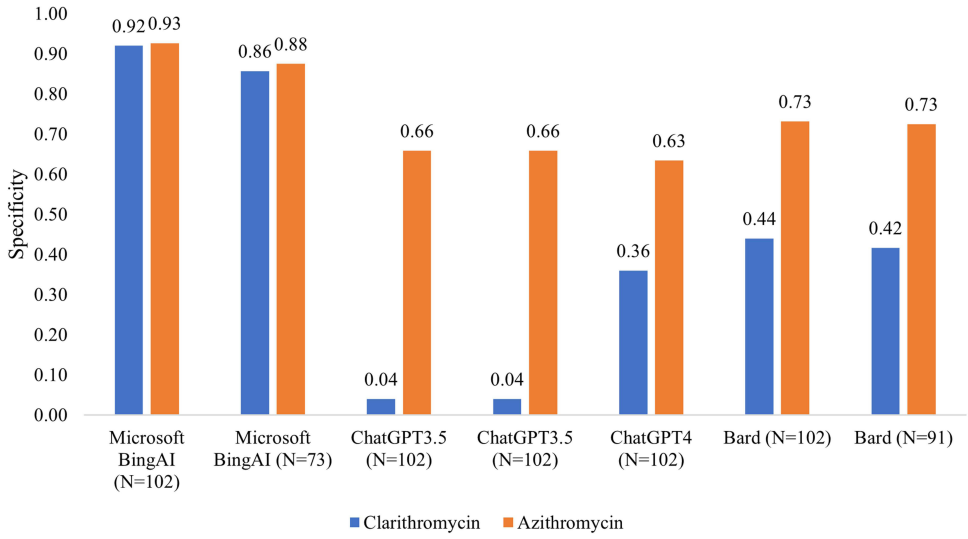

Figure 3 Specificity difference of the AI tools to detect DDIs of two medications within the same class (drugs.com as a standard). Abbreviation: DDIs, drug-drug interactions.

ChatGPT-4 specificity rates were still lower than those of Bing AI. The Bard tool showed consistent performance for azithromycin across different datasets.

Discussion

Studies conducted to compare different Large Language Models (LLM) models against other sources of information are gaining momentum. However, limited studies evaluated discrepancies between these LLM models and other online sources of drug information and indexed databases in terms of identification of drug-drug interactions.

Microsoft Bing AI demonstrated relatively strong performance in terms of sensitivity, specificity, and accuracy. Bard also showed competitive performance, particularly in terms of NPV. ChatGPT models, on the other hand, had lower performance compared to the other AI tools in most of the evaluated metrics. Therefore, it is important to consider the specific reference used (Micromedex or Drugs.com) when interpreting the results, as the performance may vary depending on the reference source.

While both Bing and ChatGPT are powered by similar AI models, their outputs can vary. For instance, Bing has higher sensitivity, specificity, and accuracy. This could be due to its ability to search and access the web to provide an answer based on the available free clinical tools such as drugs.com. In fact, in many cases, while answering the DDI question, Bing AI provides a direct reference to Drugs.com. This could also explain the high similarities between drugs.com and Bing AI in terms of identifying clinically significant DDIs. On the other hand, ChatGPT is the most verbally distinguished model and is trained on vast amounts of text. However, it does not have default internet access, particularly ChatGPT-3.5. It is worth mentioning that the new update of the ChatGPT-4 model has online connectivity through certain plugins, "browse with Bing", although we have not explored this capability in this study. The results from such a real-time internet connection could offer a different interactivity and information availability and might improve its accuracy in identifying DDIs.

ChatGPT-3.5 and ChatGPT-4, when utilized for the identification of DDIs, demonstrated a capacity to detect a higher number of potential interactions compared to other models. A possible explanation is that ChatGPT is exposed to wider, more diverse, and massive clinical content compared to the medical database (Micromedex, Drugs.com) that offers specific information related to drugs, disease, and toxicology. Nonetheless, this increased detection rate does not equate to superior performance. In fact, ChatGPT-3.5 and ChatGPT-4 exhibit the lowest levels of accuracy and specificity among the models tested. Specifically, it is prone to a higher number of false positives, where it incorrectly identifies certain drug pairs as interacting when they, in fact, do not. Therefore, while ChatGPT-3.5 and ChatGPT-4's ability to identify a larger number of potential DDIs might seem advantageous, it is crucial to consider the trade-off in terms of increased false positives and overall lower prediction accuracy. This finding is also reflective of the general and huge dataset that lacks specific training sets and is prone to false results. The risk of incorrect/inaccurate information was a challenge in almost one-third of studies that evaluated the performance and potential utilization of ChatGPT in healthcare.[18]

When assessing the consistency of accuracy rate across two different drug classes (SGLT2 inhibitors vs macrolides classes), ChatGPT-3.5 and ChatGPT-4, showed the highest variability in accuracy. In addition, both versions of ChatGPT and Bard exhibited the highest fluctuations in specificity when analyzing two medications belonging to the same category, namely azithromycin and clarithromycin. For instance, compared to azithromycin, Bard and ChatGPT's specificity was reduced for clarithromycin, implying a propensity for the two tools to inaccurately flag a greater number of drug interactions with clarithromycin that do not actually exist. One plausible explanation for this discrepancy might be the well-documented role of clarithromycin as a strong CYP3A4 inhibitor, known for its broad range of drug-drug interactions (DDIs).[17] Azithromycin, on the other hand, has a generally more favorable DDI profile. The training data for these AI chatbots might reflect the complex DDI landscape associated with clarithromycin, leading to an elevated rate of false-positive interactions, especially when contrasted with azithromycin. Despite this, it is noteworthy that Bard and ChatGPT-3.5 maintained a stable level of specificity when assessing a specific medication across various data sets. For example, when evaluating azithromycin, Bard was consistent in its performance, whether data was missing or not. This suggests a level of reliability in its approach to evaluating a particular drug, irrespective of the completeness of the data provided.

Juhi et al investigated the effectiveness of ChatGPT in detecting and explaining DDIs. Although the answers were correct, half of the answers were not clear. The authors acknowledged that ChatGPT could offer a great contribution but cautioned that patients and healthcare providers should consult expert medical professionals in clinical decision-making.[19]

The performance of LLM models was assessed in numerous medical exams and in answering different levels of reasoning questions.[20–23] ChatGPT performed well on the Medical Licensing Exam (USMLE) steps 1–3, which included complex medical and clinical information processing. ChatGPT exhibited high internal coherence, but its ability to justify correct answers was better in accurate responses than inaccurate ones.[20]

The discrepancies between different LLM chatbots were demonstrated in a study conducted by Raimondi et al on the ophthalmologists' fellowship exams. The accuracy of ChatGPT-3.5 and 4, Google Bard, and Bing Chat were significantly different ($P < 0.001$), and accuracy was significantly higher on certain subjects. Bing AI and ChatGPT-4 produced high accuracy, and inconsistencies were attributed to the utilization of different training data by LLMs.[21]

The accuracy of ChatGPT responses in different medically related studies improved with continuous iterations of the older models[24] to new ones.[25] This progress can be depicted in our findings, where ChatGPT-4 produces higher accuracy, even if minimal, than ChatGPT-3.5 in both Micromedex and Drugs.com comparisons.

Several medical foundation models were developed for electronic health records data[26,27] and pathology,[28] but the complexity of medical datasets and the entanglement of many medically related issues render these models task-specific. Consequently, they are inflexible and only capable of carrying out tasks predefined by the training dataset. The Food and Drug Administration (FDA) has approved 521 medical AI models for limited tasks, mainly radiology.[29] Advances in these AI models, such as the emergence of multimodal biomedical data models that can integrate various medical resources, would improve their performance,[30] and a focus on drug-drug interactions may reflect positively on their performance in this aspect. Additionally, despite the great potential of these models, their optimum use is still in the context of collaboration with humans.

In the landscape of future healthcare, AI platforms like ChatGPT-3.5, ChatGPT-4, Bing AI, and Bard hold significant potential in augmenting patient care, particularly DDIs detection and prevention. These advanced systems could redefine the process of DDI detection and management recommendations, currently managed by traditional clinical tools. These AI tools might reduce the chances of adverse drug events by providing an instant, comprehensive analysis of possible interactions, subsequently improving patient safety. However, in the short term, improvements in their accuracy and specificity are required. Potential ways for such improvements include allowing plugins for the DDIs conventional clinical tools to be incorporated into the newer versions of these AI tools, which would greatly enhance the specificity and accuracy of these AI platforms in detecting DDIs. In addition, including precise and up-to-date hyperlinked references for each DDI, similar to the Bing AI platform's approach, would improve the functionality of other AI tools like ChatGPT and Bard. This enhancement would assist healthcare providers in easily accessing the relevant references and making informed clinical judgments. Then, they should be integrated seamlessly with existing healthcare practices, offering clinicians a supplementary layer of decision support rather than replacing conventional tools. In this way, AI platforms can contribute to personalized, optimized, and safer patient care, helping shape a more reliable and advanced future for healthcare.

This study, while extensive, has several inherent limitations that must be acknowledged. The first is related to the scope of AI Platforms. For instance, the study is limited to four AI platforms – ChatGPT-3.5, ChatGPT-4, Bing AI, and Bard. Therefore, it might not accurately represent the performance of all AI systems in detecting DDIs. Second, comparisons are made with "two conventional clinical tools" for DDIs and two drug classes, and results may vary significantly based on which specific tools are considered the standard and which drug classes are included. Third, the study focuses on theoretical analysis and simulations and might not completely represent real-world clinical scenarios where numerous other variables can influence the outcomes. Fourth, AI technology evolves rapidly. The current study provides a snapshot in time, and the performance of these AI systems might change significantly with future updates and versions. Finally, the present study focused on the detection of DDIs, limiting the scope of understanding the overall utility of these AI platforms in patient care. Despite these limitations, the study's strength lies in its novel approach and comparative analysis of emerging AI platforms in a crucial area of patient care. It offers valuable insights into the performance of these systems against traditional tools in DDI detection. This comparative assessment can be a valuable resource for healthcare professionals, guiding them towards a potential future trajectory of AI utilization in enhancing patient care. Furthermore, this study might stimulate further research and development to optimize these AI platforms for better integration into clinical practice.

Conclusion

Bing AI had the highest accuracy, outperforming Google's Bard, ChatGPT-3.5, and ChatGPT-4. The findings highlight the significant potential these AI tools hold in transforming patient care. While the current AI platforms evaluated are not without limitations, their ability to quickly analyze potentially significant interactions with good sensitivity suggests a promising step towards improved patient safety. The study also underscores the importance of ongoing evolution in these platforms to have higher specificity and accuracy and better adapt to real-world clinical scenarios. Ultimately, the integration of such advanced systems with existing healthcare practices could pave the way for a future where AI augments traditional medical tools, contributing to a more personalized, optimized, and safer healthcare landscape.

Transparency Statement

We have reviewed and complied with the terms of use of all the artificial intelligence tools used in the study.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Mhlanga D. Open AI in education, the responsible and ethical use of ChatGPT towards lifelong learning. Education, the Responsible and Ethical Use of ChatGPT Towards Lifelong Learning (February 11, 2023); 2023.
2. Vaishya R, Misra A, Vaish A. ChatGPT: is this version good for healthcare and research? *Diabetes Metab Syndr*. 2023;17(4):102744. doi:10.1016/j.dsx.2023.102744
3. Hauben M. Artificial intelligence and data mining for the pharmacovigilance of drug–drug interactions. *Clin Ther*. 2023;45(2):117–133. doi:10.1016/j.clinthera.2023.01.002
4. Haluza D, Jungwirth D. Artificial intelligence and ten societal megatrends: an exploratory study using GPT-3. *Systems*. 2023;11(3):120. doi:10.3390/systems11030120
5. Jain S, Naicker D, Raj R, et al. Computational intelligence in cancer diagnostics: a contemporary review of smart phone apps, current problems, and future research potentials. *Diagnostics*. 2023;13(9):1563. doi:10.3390/diagnostics13091563
6. Kriger-Sharabi OA, Kopylov U. Harnessing the power of precision medicine and novel biomarkers to treat Crohn's disease. *J Clin Med*. 2023;12(7):2696. doi:10.3390/jcm12072696
7. Khan O, Parvez M, Kumari P, Parvez S, Ahmad S. The future of pharmacy: how AI is revolutionizing the industry. *Intell Pharm*. 2023;1(1):32–40. doi:10.1016/j.ipha.2023.04.008
8. Bays HE, Fitch A, Cuda S, et al. Artificial intelligence and obesity management: an Obesity Medicine Association (OMA) Clinical Practice Statement (CPS) 2023. *Obes Pillars*. 2023;6:100065. doi:10.1016/j.obpill.2023.100065
9. Liu L, Duffy VG. Exploring the future development of Artificial Intelligence (AI) applications in chatbots: a bibliometric analysis. *Int J Soc Robot*. 2023; 2023:1–14.
10. Cabreja-Castillo M, Hernandez L, Mustafa A, Hungria G, Bertoli MT. COVID-19 scientific literacy in medical and nursing students. *J Microbiol Biol Educ*. 2023;24(1):e00219–00222. doi:10.1128/jmbe.00219-22
11. Rivas P, Zhao L. Marketing with chatgpt: navigating the ethical terrain of gpt-based chatbot technology. *AI*. 2023;4(2):375–384. doi:10.3390/ai4020019
12. Giansanti D. The chatbots are invading us: a map point on the evolution, applications, opportunities, and emerging problems in the health domain. *Life*. 2023;13(5):1130. doi:10.3390/life13051130
13. Hammour KA, Farha RA, Ya'acoub R, Salman Z, Basheti I. Impact of pharmacist-directed medication reconciliation in reducing medication discrepancies: a randomized controlled trial. *Can J Hosp Pharm*. 2022;2022:1.
14. Bossaer JB, Eskens D, Gardner A. Sensitivity and specificity of drug interaction databases to detect interactions with recently approved oral antineoplastics. *J Oncol Pharm Pract*. 2022;28(1):82–86. doi:10.1177/1078155220984244
15. Rossing P, Caramori ML, Chan JC, et al. Executive summary of the KDIGO 2022 clinical practice guideline for diabetes management in chronic kidney disease: an update based on rapidly emerging new evidence. *Kidney Int*. 2022;102(5):990–999. doi:10.1016/j.kint.2022.06.013
16. Heidenreich PA, Bozkurt B, Aguilar D, et al. 2022 AHA/ACC/HFSA guideline for the management of heart failure: a report of the American College of Cardiology/American Heart Association Joint Committee on clinical practice guidelines. *J Am Coll Cardiol*. 2022;79(17):e263–e421. doi:10.1016/j.jacc.2021.12.012
17. Fleet JL, Shariff SZ, Bailey DG, et al. Comparing two types of macrolide antibiotics for the purpose of assessing population-based drug interactions. *BMJ open*. 2013;3(7):e002857. doi:10.1136/bmjopen-2013-002857
18. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare*. 2023;11(6):887. doi:10.3390/healthcare11060887
19. Juhi A, Pipil N, Santra S, Mondal S, Behera JK, Mondal H. The capability of ChatGPT in predicting and explaining common drug-drug interactions. *Cureus*. 2023;15(3):e36272. doi:10.7759/cureus.36272
20. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198
21. Raimondi R, Tzoumas N, Salisbury T, et al. Comparative analysis of large language models in the Royal College of Ophthalmologists fellowship exams. *Eye*. 2023. doi:10.1038/s41433-023-02563-3
22. 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. doi:10.2196/45312
23. Das D, Kumar N, Longjam LA, et al. Assessing the capability of ChatGPT in answering first- and second-order knowledge questions on microbiology as per competency-based medical education curriculum. *Cureus*. 2023;15(3):e36034. doi:10.7759/cureus.36034
24. Jin D, Pan E, Oufattole N, Weng W-H, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Appl Sci*. 2021;11(14):6421. doi:10.3390/app11146421
25. Liévin V, Hother CE, Winther OJ. Can large language models reason about medical questions?; 2022.
26. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. *NPJ Digit Med*. 2022;5(1):194. doi:10.1038/s41746-022-00742-2
27. Steinberg E, Jung K, Fries JA, Corbin CK, Pfohl SR, Shah NH. Language models are an effective representation learning technique for electronic health record data. *J Biomed Inform*. 2021;113:103637. doi:10.1016/j.jbi.2020.103637
28. Tiu E, Talius E, Patel P, Langlotz CP, Ng AY, Rajpurkar P. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nat Biomed Eng*. 2022;6(12):1399–1406. doi:10.1038/s41551-022-00936-9

29. Food and Drug Administration. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. Food and Drug Administration; 2023. Available from: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices. Accessed September 12, 2023.
30. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. *Nat Med*. 2022;28(9):1773–1784. doi:10.1038/s41591-022-01981-2