Exploratory Research in Clinical and Social Pharmacy 20 (2025) 100655

Contents lists available at ScienceDirect

Exploratory Research in Clinical and Social Pharmacy

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

Comparative evaluation of artificial intelligence platforms and drug interaction screening databases using real-world patient data

Bálint Márk Domián [a], Amir Reza Ashraf [a], András Tamás Fittler [a], Mátyás Káplár [b], Róbert György Vida [a,*]

[a] Department of Pharmaceutics, Faculty of Pharmacy, University of Pécs, Pécs, Hungary
[b] Institute of Psychology and Mental Health, University of Pannonia, Veszprém, Hungary

ARTICLE INFO

Keywords:
Artificial intelligence
Chatbot
Clinical pharmacy
Drug-drug interaction
Medication review
Patient safety
ChatGPT
Gemini
Copilot

ABSTRACT

Background: The use of multiple medications increases the risk of harmful drug-drug interactions (DDIs). Conventional DDI screening databases vary in coverage and often trigger low-relevance alerts, contributing to alert fatigue. Large language models (LLMs) have emerged as potential tools for DDI identification, however, their performance compared to established databases using real-world patient data remains under-explored.

Methods: In this exploratory study, we compared conventional database screening with LLM-based screening using anonymized medication lists from rheumatology patients. Lexicomp, Medscape and Drugs.com were used to compile a reference set of 204 clinically relevant interactions across 57 cases. Using identical prompts, we then queried ChatGPT, Google Gemini and Microsoft Copilot for interactions potentially requiring pharmacists' intervention. We calculated sensitivity, specificity, precision and F1 score.

Results: Compared to the reference set of 204 DDIs, ChatGPT identified 439, Gemini 1556, and Copilot 1813 potential interactions. While Gemini achieved the highest sensitivity (0.697), ChatGPT demonstrated higher specificity (0.868). All three platforms demonstrated low precision scores. Overall, ChatGPT achieved the highest performance by F1 score (0.2520), followed by Gemini (0.1933) and Copilot (0.1153). Our results suggest that no AI systems assessed achieve the required balance of precision and sensitivity for reliable clinical decision-making in DDI screening.

Conclusion: Although LLMs show promise as complementary tools in DDI screening, as they proved effective in identifying true interactions, they generate clinically inaccurate information due to hallucinations, which limits their reliability as standalone screening tools. Consequently, while LLMs could support clinical pharmacists in polypharmacy management, their outputs must always undergo professional validation to ensure patient safety.

1. Introduction

Polypharmacy, the regular use of multiple medications at the same time, is increasingly common, especially among older adults, creating heightened risk for drug–drug interactions (DDIs). The identification and management of DDIs represents a critical component of patient care and pharmacy services. However, conventional online drug interaction screening databases (e.g. Drugs.com Drug Interaction Checker) or interaction checkers integrated into pharmacy information management systems often vary in comprehensiveness, consistency, and clinical relevance.[1,2] Additionally, the excessive number of drug interaction warnings of low clinical significance may result in cognitive overload and alert fatigue. In practice, this high volume and variability of alerts can lead to missed critical interactions or unnecessary medication changes, compromising patient safety and therapeutic outcomes. Recent studies emphasize that even well-established databases may overlook relevant interactions or present conflicting categorizations, further challenging clinical pharmacists' ability to make informed decisions.[3,4]

Abbreviations: AI, Artificial Intelligence; ANOVA, Analysis of Variance; API, Active Pharmaceutical Ingredient; CYP, Cytochrome P450; DDI, Drug-Drug Interaction; FN, False Negative; FP, False Positive; HTA, Health Technology Assessment; LLM, Large Language Model; M, Mean; NLP, Natural Language Processing; p, *p*-value (significance); SmPC, Summary of Product Characteristics; STARD, Standard for Reporting Diagnostics Accuracy Studies; TN, True Negative; TP, True Positive; χ^2, Chi-squared test.

* Corresponding author at: Szigeti street 4, Pécs 7624, Hungary.
E-mail address: vida.robert@pte.hu (R.G. Vida).

https://doi.org/10.1016/j.rcsop.2025.100655
Received 30 July 2025; Accepted 7 September 2025
Available online 8 September 2025

These longstanding challenges and limitations of traditional screening tools highlight the need for more advanced and adaptable solutions. Artificial intelligence (AI) is rapidly transforming healthcare, with AI-powered tools emerging as promising complementary solutions for clinical decision support, diagnostics, and personalized medicine.[5,6] Within pharmacy practice specifically, these technologies show marked promise for medication use review and medication reconciliation, particularly when addressing complex polypharmacy regimens. These tools use large language models (LLMs) with natural language processing ability and leverage large-scale biomedical data to potentially provide more adaptive and clinically nuanced DDIs assessments.[7–9] Multiple recent studies have shown that mainstream generative AI models, such as ChatGPT, Google Gemini (formerly Bard), and Microsoft Copilot (formerly Bing AI), can provide guidance and recommendations when asked about acquiring medications via the internet, as well as recognizing and classifying drug interactions with differing levels of performance and correctness.[10–13] Although emerging evidence supports the potential of generative AI in pharmacy practice, early investigations revealed that the performance of AI models vary significantly depending on the platform and the framing of prompts, highlighting the need for structured benchmarking against conventional databases.[14,15]

Despite this variability, many healthcare professionals including clinical pharmacists are beginning to integrate these tools into their workflows.[16] Familiarity with digital technologies and commitment to evidence-based, rapid decision-making position make young healthcare professionals as early adopters of AI tools in clinical settings.[17] Studies have shown that pharmacy professionals, predominantly those in academic or hospital settings, are increasingly engaging with AI tools to assess interactions, improve patient counseling, and streamline clinical workflows.[18,19] The evolving expectations of modern healthcare, combined with the increasing complexity of polypharmacy, further amplify the appeal of AI-enhanced solutions for clinicians[20] and pharmacists as well. Nevertheless, despite growing adoption, studies on comparative performance evaluation of mainstream LLM-based chatbot platforms against established databases with real-world data remain scarce. This study addresses the gap by comparing the performance of conventional DDI databases versus AI chatbots using real-world polypharmacy patient data for structured, comparative benchmarking. We focus on clinically relevant drug interactions to evaluate the reliability and practical utility of these AI tools in supporting safe medication therapy management.

2. Methods

2.1. Patient data and ethical considerations

To determine eligibility for subsequent AI testing, preliminary DDI screening was performed on anonymized medication lists from 80 rheumatology patients included in our previously published cross-sectional observational study (sum = 661 medication; average = 11.6 medication/medication list; 89.5 % polypharmacy prevalence).[3] The institutional ethical review board approved the original data collection, and because all patient data were de-identified for this study, additional ethics approval was not required as no personal or identifiable information was accessed or analyzed.

2.2. Drug interaction database selection and DDI classification

To establish a reliable baseline for evaluating AI model performance, we utilized three widely recognized drug interaction databases: the UpToDate Lexicomp database (Wolters Kluwer Clinical Drug Information), the Medscape drug interaction checker (WebMD LLC.), and Drugs.com database (Drugsite Limited).[21–24] Throughout this study, the use of drug interaction screening databases is referred to as the standard or conventional method. Since relying on a single DDI source may yield inconsistent or incomplete results due to variations in database coverage and interpretation,[3,25–27] each medication list was checked across all three databases, and the interaction results were systematically documented according to each database's DDI severity category classifications.

2.3. DDI database category unification technique and medication list selection

Since these databases use different nomenclatures for DDI severity categorization, we consolidated the categories into a unified binary system to enable comparison. These included a "Clinically relevant" and a "Clinically NOT relevant" category (consisting of minor interactions requiring no action, and instances of no interaction) to enable comparison (Table 1.).

Database DDI results were paired with the relevant unified category, and we considered a DDI clinically relevant only if all three databases identified it with consensus. The remaining drug pairs were classified as "Clinically NOT relevant". With these consensus interactions, we constructed our reference database for AI comparison.

Following database screening, the medication lists were included in the AI analysis only when there was consensus across databases for at least one clinically relevant interaction pair, or when all databases agreed on flagging no interactions at all.

2.4. Artificial intelligence platforms

To ensure broad relevance and generalizability, the evaluated AI platforms were selected based on market share and size of their user base at the time of data collection in April 2025.[28] We evaluated OpenAI's ChatGPT (GPT-4 model), Google's Gemini 2.0 (Flash edition), and Microsoft Bing's Copilot, using the latest available versions at the time of investigation. All platforms were accessed using identical prompts to maintain objectivity.

2.5. Prompt development

We aimed to ensure clinical relevance during prompt development by mirroring the language and perspective a healthcare professional might use when assessing potential drug interactions. The prompt incorporated common categorizations used in reference databases with the following instruction:

"List those interactions that require intervention by a healthcare professional, such as a clinical pharmacist, when reviewing medication use, because they are categorized as 'Avoid combination,' 'Consider therapy modification,' 'Use alternative,' or 'Monitor therapy' drug interactions."

This phrasing was designed to focus on interactions that typically require professional oversight and intervention. Beside this instruction prompt, each query contained the medication list in form of active pharmaceutical ingredients (APIs) using their international nonproprietary names (INNs). We deliberately requested no explanations or justifications, requesting each AI to provide only a list of clinically relevant interaction pairs that require professional intervention. Since the AI

Table 1
Unified categorization of DDI severity across databases.

UpToDate Lexicomp	Medscape	Drugs.com	Unified categories
Avoid combination Consider therapy modification	Serious - Use alternative	Major	Clinically relevant
Monitor therapy	Significant - Monitor closely	Moderate	
No action needed	Minor	Minor	Clinically NOT relevant
No known interaction	No interaction	No interaction	

platforms generated responses automatically and independently without human interpretation during output generation, blinding procedures were not applicable. All AI platforms provided definitive interaction lists for each medication list, with no inconclusive or missing data encountered during data collection.

2.6. Data analysis

Model outputs were benchmarked against the reference baseline created by using conventional databases (Lexicomp, Medscape, Drugs.com) by the study investigators (licensed pharmacists). For each individual medication list, the total number of clinically relevant DDIs reported by each AI was recorded to determine the number of true positives (TP), and false positives (FP). False negative (FN) values were calculated from reference baseline values minus TP. True negative (TN) values were calculated from the theoretical maximum of possible pairwise interactions minus the sum of TP, FP and FN values for each given case. Using these values, we calculated the following performance metrics: sensitivity, specificity, precision, and F1 score for individual medication lists and then calculated the mean value of these metrics to assess the performance of the AI platforms. Detailed description of performance metrics for assessing AI based DDI classification is available in Supplement 1.

2.7. Statistical analysis

To compare the performance of the three AI platforms across multiple evaluation metrics, non-parametric statistical tests were employed due to violations of normality as assessed by the Kolmogorov-Smirnov test. Specifically, the Friedman test was used to detect overall differences among the models for sensitivity, specificity, precision, and F1 scores. When the Friedman test indicated significant differences, post-hoc pairwise comparisons were conducted using the Wilcoxon signed-rank test. Statistical analyses were performed using SPSS version 28 (IBM Corp.) software. All analyses were performed on a within-subjects basis using data from 57 matched cases, and statistical significance was determined at an alpha level of 0.05 (two-tailed). These non-parametric approaches ensured robust comparison of model performance despite non-normal distribution characteristics in the data.

2.8. Reporting guideline compliance

This study was designed and reported following the Standards for Reporting Diagnostic Accuracy Studies (STARD 2015) guidelines to ensure transparent and comprehensive reporting. Since the STARD-AI is currently under development and not yet finalized,[29] STARD 2015 represents the most appropriate available framework for evaluating AI-based drug-drug interaction screening.[30] For this purpose, the artificial intelligence platforms (ChatGPT, Gemini, and Microsoft Copilot) served as the evaluated index tests, while our drug-drug interaction classification dataset derived from conventional screening databases (Lexicomp, Medscape, and Drugs.com) served as the reference standard.

3. Results

3.1. Identification of all DDIs using the conventional method

The UpToDate Lexicomp database reported 554 interactions, classified as follows: 14 labeled “Avoid combination”, 102 as “Consider therapy modification”, 347 as “Monitor therapy”, and 91 as “No action needed”. The Medscape drug interaction checker yielded 740 DDIs, comprising of 77 “Serious-Use alternative”, 450 “Monitor closely”, and 213 “Minor” interactions. The Drugs.com interaction checker identified 835 DDIs, including 168 “Major”, 598 “Moderate”, and 69 “Minor” interactions. We identified a total of 2129 DDI signals across the three conventional drug interaction screening databases, for all potential DDI severity categories and including duplicates. The number of DDIs identified with different categories and as a total number are shown in Table 2. Collectively, the conventional databases flagged 1756 clinically relevant DDIs that require professional intervention.

3.2. Identification of clinically relevant DDIs and medication lists

There were substantial discrepancies among the three databases regarding the nomenclature and classification of DDIs. As described in the Methods section, we unified the interaction categories and identified DDIs that were consistently classified in the’Clinically relevant’ unified category across all three sources. After applying our unified categorization and matching process, we identified 204 DDIs classified as clinically relevant across 57 medication lists (reference standard). In case of the remaining 23 medication lists there was not a single drug-drug interaction with a consensus in the categorization of clinical relevance among the three selected conventional databases.

3.3. Identification of clinically relevant DDIs using AI platforms

Using the 57 individual medication lists, we then assessed the drug interaction screening capabilities of the three AI platforms against our reference standard. The AI platforms collectively flagged a total of 3808 DDI signals that require intervention. ChatGPT flagged the fewest clinically relevant DDIs, with 439 interactions requiring intervention. Gemini flagged 1556, while Microsoft Copilot reported the highest number, flagging 1813 interactions.

3.4. Performance evaluation of AI platforms

Besides the comparison of the number of clinically relevant DDI alerts in the different databases and platforms, we aimed to compare the performance of ChatGPT, Gemini and Copilot against the reference standard. Previous studies typically compared AI platforms to interaction databases based on drug-drug interaction pairs, not real-world medication lists (Supplement 2.). We calculated sensitivity, specificity, precision, and F1 score, along with the mean value of these metrics, to assess how effectively each AI platform identifies clinically relevant drug-drug interactions.

The Friedman test revealed significant differences among the models across all performance metrics (sensitivity, specificity, precision, F1 score), demonstrating that the models differ substantially in both individual and combined aspects of performance ($p < 0.001$). The results of performance evaluations are summarized in Table 3 and Fig. 1.

Assessment of performance evaluation (Table 3.) metrics and pairwise comparisons using the Wilcoxon signed-rank test revealed distinct strengths, limitations, and performance trade-off patterns across the three AI platforms, with potential implications for DDI screening clinical decision-making. For sensitivity, Gemini achieved the highest mean score ($M = 0.6966$), significantly outperforming both Copilot ($M = 0.5863$, $p = 0.021$) and ChatGPT ($M = 0.4683$, $p < 0.001$). This finding indicates that Gemini demonstrates superior capability in detecting actual drug interactions, correctly identifying approximately 20 % more clinically relevant DDIs than ChatGPT. ChatGPT demonstrated superior specificity ($M = 0.8683$), with significantly fewer false positives relative to true negatives than both Gemini ($M = 0.5995$, $p < 0.001$) and Copilot ($M = 0.4295$, $p < 0.001$). This high specificity indicates that ChatGPT correctly identifies approximately 87 % of non-interacting drug pairs as safe, compared with 60 % for Gemini and 43 % for Copilot. However, the dataset is highly imbalanced in the context of DDI screening, with an extreme number of true negative drug pairs and a relatively high number of false positive signals. Therefore, relying solely on specificity as a unique performance metric can be misleading, as it may overestimate the practical utility of the model. Although ChatGPT again outperformed the others ($M = 0.1937$) in terms of precision score, significantly exceeding Gemini ($M = 0.1297$, $p = 0.002$) and Copilot (M

Table 2
The number of DDIs and their categorization identified with the conventional method.

UpToDate Lexicomp		Medscape		Drugs.com		Unified categories	
Avoid combination	14	Serious - Use alternative	77	Major	168	Clinically relevant	1756
Consider therapy modification	102						
Monitor therapy	347	Significant - Monitor closely	450	Moderate	598		
No action needed	91	Minor	213	Minor	69	Clinically NOT relevant	373

Table 3
Mean diagnostic performance metrics and comparative analysis of all AI platforms.

Metric	ChatGPT Mean score (±SD)	Gemini Mean score (±SD)	Copilot Mean score (±SD)	Friedman χ^2 (df = 2), p
Sensitivity	0.4683 (±0.3907)	0.6966 (±0.4095)	0.5863 (±0.4655)	14.552, 0.001
Specificity	0.8683 (±0.1460)	0.5995 (±0.2676)	0.4295 (±0.3855)	59.760, <0.001
Precision	0.1937 (±0.1680)	0.1297 (±0.1247)	0.0738 (±0.0807)	38.233, <0.001
F1 score	0.2520 (±0.2006)	0.1933 (±0.1594)	0.1153 (±0.1166)	38.714, <0.001

= 0.0738, $p < 0.001$), all three platforms demonstrated notably low precision scores, indicating that the majority of flagged interactions were false positives. For example, the 19 % precision of ChatGPT means that only approximately 1 in 5 warnings represents a true interaction, whereas Copilot's 7 % precision translates to approximately 13 false alarms for every detected DDI. This generally high false positive rate of AI platforms could lead to unnecessary alerts and clinical interventions. The F1 score balances both detection capability and warning accuracy. ChatGPT achieved the highest overall performance by F1 score (M = 0.2520), followed by Gemini (M = 0.1933, $p = 0.015$) and Copilot (M = 0.1153, $p < 0.001$). Our study results indicate suboptimal overall performance across all platforms and suggest that none of the evaluated AI systems achieve the balanced precision-sensitivity performance necessary for reliable clinical decision support in drug interaction screening. Such low F1 scores reflect the inherent challenge of the highly imbalanced nature of drug interaction detection, where clinically relevant interactions represent a small fraction of all possible drug combinations.

4. Discussion

The potential impact of AI in healthcare is significant, especially in the area of clinical decision support, patient safety, DDI identification and prediction, and pharmacovigilance.31 The ability of AI-based solutions to analyze large, complex datasets, and detect patterns that humans might overlook is a key advantage for the assessment of DDIs.8,32,33

Our study builds upon this emerging evidence by systematically evaluating performance metrics across LLMs using real-world patient medication lists rather than isolated drug pairs and hypothetical patient profiles, which is a more clinically relevant approach that gives a better representation of real-world performance of these mainstream AI chatbots.

Our results show that while these chatbots can identify drug interactions in patients taking multiple medications, their performance varies widely with profound clinical implications. The higher sensitivity of Gemini makes it more suitable for safety-critical screening where missing interactions could have severe consequences, though its lower specificity may overwhelm clinicians with false alerts. Conversely, precision and F1 scores were uniformly poor across all AI platforms, indicating that current mainstream AI platforms are not yet ready, nor highly reliable for drug interaction screening in clinical practice. In particular, much higher precision is essential to minimize false positives, prevent alert fatigue and maintain clinical workflow efficiency.

Comparing studies on AI chatbot performance specific to DDI detection is challenging due to heterogeneous methods and outcome measurement.31 Seventeen such studies have been published over the last few years, highlighting a growing but still limited body of research in this area.4,8–12,14,15,18–20,34–39 These publications consistently demonstrate that the accuracy of LLMs can vary significantly. While some studies have shown that AI platforms can identify most interactions, they also generate a considerable number of false positives. For instance, Sicard et al. found that ChatGPT and Claude achieved approximately 99 % sensitivity for known adverse drug reaction-associated DDIs, yet their specificity remained low (between 0.64 and 0.68), leading to frequent misclassification of negative controls as interactions.39 Similarly, a 2024 study showed that ChatGPT-4.0 outperformed ChatGPT-3.5 in terms of overall accuracy but still exhibited low sensitivity. These findings, summarized in the Supplementary material (Table S1), underscore the need for further optimization of AI models before they can reliably support clinical pharmacy decision-making.10,11

It is important to note that the use of real medication lists, standardized drug pairs, or fixed drug pairs can affect the results. Radha Krishnan et al.'s 2024 study found that ChatGPT 3.5 exhibited a sensitivity of 24 % or less when given real patient profiles.10 In our study, ChatGPT demonstrated a sensitivity of 47 % when evaluating 57 real patient medication regimens, which is notably lower than the 91.5 % observed with professional curated lists.4 Other studies have also found higher sensitivity on simple benchmarks than on clinical profiles (see Supplementary File 2).

However, AI platform performance varied not only in complex scenarios with multiple medications, but also when assessing controlled drug-pairs. Al-Ashwal et al. (2023) found that ChatGPT 3.5 achieved an accuracy of only 47 %, compared to 79 % for Bing/Copilot. Meanwhile, Aksoyalp and Erdoğan (2024) reported sensitivities and specificities of 91 % and 97 % respectively, for ChatGPT 3.5 on 78 clopidogrel-specific DDIs. Overall, the literature echoes the cautious tone of our study. Currently, no LLM matches the performance of clinical DDI screenings, even when sensitivity is adequate, precision remains poor and hallucinations occur. These limitations are largely due to the fact that LLMs and AI chatbots were not specifically designed and trained for healthcare applications. They were trained to acquire general reasoning abilities based on publicly available data, which means they lack curated, context-specific knowledgebase required for reliable clinical decision-making.6,40,41 Furthermore, these models can introduce new errors through hallucinations. A striking example from Bischof et al. analysis occurred when ChatGPT falsely classified a magnesium supplement as an antacid, leading it to incorrectly conclude it affects the absorption of the immunosuppressant mycophenolate mofetil. Similar to some databases, LLMs can also struggle with APIs and drug classes, incorrectly attributing the specific characteristics of a single drug (e.g., a specific statin) to its entire therapeutic drug class.15 These examples again highlight the need for professional oversight over AI tools used in clinical context.39,42–44

DDIs are a well-recognized cause of adverse drug events and analyzing them has become increasingly complex. This complexity is due to the vast number of available drugs and supplements, the rise of

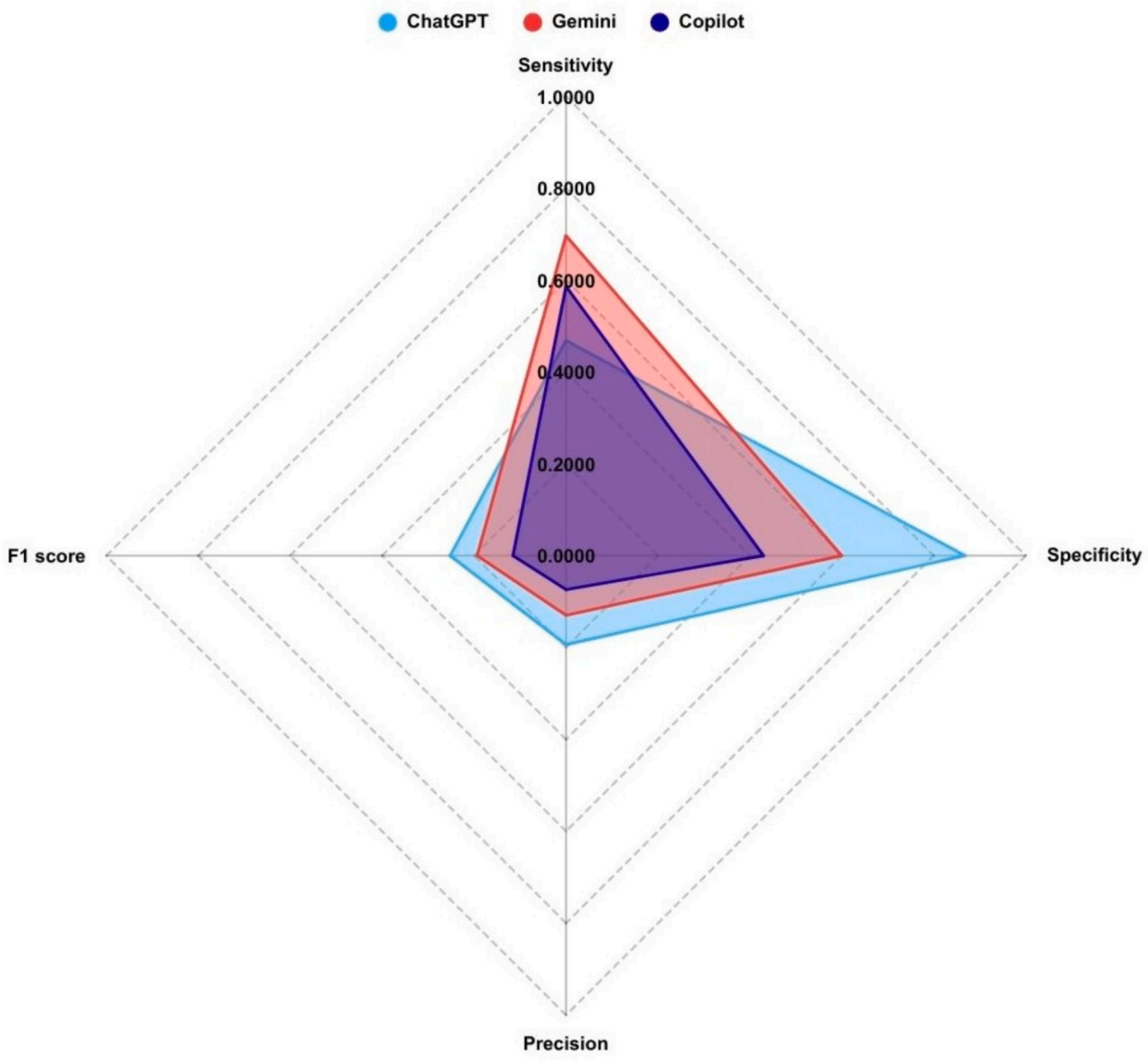

Fig. 1. Radar graph visualizing the mean performance metrics of AI platforms.

specialty medications (like targeted therapies and monoclonal antibodies, orphan drugs), the growing prevalence of polypharmacy and multimorbidity in an ageing population,[45] and the inherent limitations of current screening tools.[46] Previously we found that Drugs.com performed better for targeted therapy drug-drug interactions, while Upto-Date Lexicomp identified more drug-supplement interactions.[3,25–27] Other key limitations of conventional DDI screening include incomplete coverage of all commercially available health products and their APIs, ineffective handling of synonyms, and lack of visualizing features for interaction networks to identify core medicines that could be substituted to non-interacting alternatives, thereby preventing severe complications during polypharmacy.[3,47,48]

While LLM-based chatbots like ChatGPT show potential to enhance decision-making processes by overcoming some of the barriers existing in nowadays DDIs screening (e.g. nomenclature issues, product availability issues, etc.), and even offering user-friendly explanations or simplifying complex interactions, significant risks remain. Our findings support this caution. The lack of continuous, real-time updates and formal clinical validation means that over-reliance on AI chatbots without pharmacist consultation can lead to serious drug related problems, especially when evaluating complex cases or scenarios.[4,7,10,18]

4.1. Strengths and limitations

Our research has several strengths, including the use of real-world patient data and a large sample size to represent clinical practice. By selecting multiple AI platforms and focusing objectively on clinically relevant potential DDIs, our study contributes to previous literature in this field. However, several limitations must be acknowledged. We did not include data on the effects of potential DDIs on patients and did not perform a holistic assessment of the standardized medication lists. We did not perform supplement-drug interaction screening, as our methodology required potential DDIs to be present in multiple databases, a criterion that conventional drug interaction screening databases often fail to meet. The generalizability of our results is limited by specific cases and the choice of LLMs. As drug interaction literature and AI technologies evolve rapidly, our cross-sectional study provides only a snapshot of LLM performance at the time of investigation.

5. Conclusion

LLMs show significant promise as complementary tools for DDI screening, particularly in managing varying drug nomenclature and synonyms, areas where traditional standard database screening

platforms often struggle. Our findings demonstrate that LLMs are effective at identifying true drug interactions, however, several limitations prevent their adoption in everyday clinical application for screening potential. Most importantly, LLMs frequently generate clinically inaccurate information due to hallucinations, which could create patient safety risks. Additionally, AI chatbots may fail to identify clinically relevant potential DDIs, resulting in critical errors and inconsistencies in the outcomes. Moreover, the use of AI with patient information requires careful consideration of data ethics, patient privacy, and management of inaccurate and hallucinatory responses.[10,42–44] Evaluating the performance of LLMs as DDI tools remains challenging when studies use inconsistent methods and metrics, leading to methodological and interpretation biases. To improve comparability, we recommend using real-world medication-use data, interactions verified by multiple sources, standardized severity categories, and balanced metrics like the F1 score that account for both sensitivity and precision. Further research with standardized methodologies and comparable outcome metrics is needed to validate AI chatbots in controlled clinical settings to establish appropriate frameworks for using these technologies for polypharmacy management. Until these validations are achieved, mainstream AI platforms should be considered only as novel experimental tools that require supervision and fact-checking using established standard databases and cannot replace the clinical judgment of pharmacists and healthcare professionals.

CRediT authorship contribution statement

Bálint Márk Domián: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Amir Reza Ashraf:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **András Tamás Fittler:** Conceptualization, Funding acquisition, Methodology, Writing – original draft, Writing – review & editing. **Mátyás Káplár:** Methodology. **Róbert György Vida:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Ethical compliance

Not applicable/Not required for this study. All information collected from this study was from the public domain and the study did not involve any interaction with users.

Funding

The research was supported by the Hungarian Scientific Research Fund (grant NKFI-ID 143684).

Declaration of competing interest

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

Acknowledgments

This research was conducted to evaluate responses generated by various large language models. Accordingly, this article and its supplementary files contain responses and content generated by these models. The authors have used Trinka AI (Enago) for minor language editing purposes (grammar, spelling, wording). Trinka AI did not modify the manuscript's scientific content or interpret the study's data, analyses, or conclusions. All authors reviewed and approved the final manuscript. Authors reports no conflict of interest associated with this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rcsop.2025.100655.

Data availability

The datasets generated or analyzed during this study are available from the corresponding author upon reasonable request.

References

1. Phansalkar S, van der Sijs H, Tucker AD, et al. Drug-drug interactions that should be noninterruptive in order to reduce alert fatigue in electronic health records. *J Am Med Inform Assoc*. 2013;20:489–493. https://doi.org/10.1136/amiajnl-2012-001089.
2. Saverno KR, Hines LE, Warholak TL, et al. Ability of pharmacy clinical decision-support software to alert users about clinically important drug—drug interactions. *J Am Med Inform Assoc*. 2011;18:32–37. https://doi.org/10.1136/jamia.2010.007609.
3. Rajj R, Schaadt N, Bezsila K, et al. Vida, survey of potential drug interactions, use of non-medical health products, and immunization status among patients receiving targeted therapies. *Pharmaceuticals*. 2024;17:942. https://doi.org/10.3390/ph17070942.
4. Al-Ashwal FY, Zawiah M, Gharaibeh L, Abu-Farha R, Bitar AN. 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;15:137–147. https://doi.org/10.2147/DHPS.S425858.
5.. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ*. 2023;23:689. https://doi.org/10.1186/s12909-023-04698-z.
6. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI Chatbot for medicine. *N Engl J Med*. 2023;388:1233–1239. https://doi.org/10.1056/NEJMsr2214184.
7.. Zhang Y, Deng Z, Xu X, Feng Y, Junliang S. Application of artificial intelligence in drug–drug interactions prediction: A review. *J Chem Inf Model*. 2024;64:2158–2173. https://doi.org/10.1021/acs.jcim.3c00582.
8. Roosan D, Padua P, Khan R, Khan H, Verzosa C, Wu Y. Effectiveness of ChatGPT in clinical pharmacy and the role of artificial intelligence in medication therapy management. *J Am Pharm Assoc*. 2024;64:422–428.e8. https://doi.org/10.1016/j.japh.2023.11.023.
9. Kim WT, Shin J, Yoo I-S, et al. Medication extraction and drug interaction Chatbot: generative Pretrained transformer-powered Chatbot for drug-drug interaction. *Mayo Clinic Proc Digital Health*. 2024;2:611–619. https://doi.org/10.1016/j.mcpdig.2024.09.001.
10. Radha Krishnan RP, Hung EH, Ashford M, et al. Evaluating the capability of ChatGPT in predicting drug–drug interactions: real-world evidence using hospitalized patient data. *Br J Clin Pharmacol*. 2024;90:3361–3366. https://doi.org/10.1111/bcp.16275.
11. Thapa RB, Karki S, Shrestha S. Exploring potential drug-drug interactions in discharge prescriptions: ChatGPT's effectiveness in assessing those interactions. *Expl Res Clin Soc Pharm*. 2025;17, 100564. https://doi.org/10.1016/j.rcsop.2025.100564.
12. Most A, Chase A, Sikora A. *Assessing the Potential of ChatGPT-4 to Accurately Identify Drug-Drug Interactions and Provide Clinical Pharmacotherapy Recommendations*. 2024. https://doi.org/10.1101/2024.06.29.24309701.
13. Ashraf AR, Mackey TK, Fittler A. Search engines and generative artificial intelligence integration: public health risks and recommendations to safeguard consumers online. *JMIR Public Health Surveill*. 2024;10, e53086. https://doi.org/10.2196/53086.
14. Aksoyalp ZŞ, Erdoğan BR. Comparative evaluation of artificial intelligence and drug interaction tools: a perspective with the example of CLOPIDOGREL. *Ankara Universitesi Eczacilik Fakultesi Dergisi*. 2024;48:22. https://doi.org/10.33483/jfpau.1460173.
15.. Bischof T, Al Jalali V, Zeitlinger M, et al. Chat <scp>GPT</scp> vs. Clinical decision support systems in the analysis of drug–drug interactions. *Clin Pharmacol Ther*. 2025;117:1142–1147. https://doi.org/10.1002/cpt.3585.
16. Falconer N, Scott I, Barras M. Powered by <scp>AI</scp> : advancing towards artificial intelligence algorithms in Australian hospital pharmacy. *J Pharm Pract Res*. 2024;54:107–109. https://doi.org/10.1002/jppr.1922.
17.. Lambert SI, Madi M, Sopka S, et al. An integrative review on the acceptance of artificial intelligence among healthcare professionals in hospitals. *NPJ Digit Med*. 2023;6:111. https://doi.org/10.1038/s41746-023-00852-5.
18. 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. https://doi.org/10.7759/cureus.36272.
19. Hsu H-Y, Hsu K-C, Hou S-Y, Wu C-L, Hsieh Y-W, Cheng Y-D. Examining real-world medication consultations and drug-herb interactions: ChatGPT performance evaluation. *JMIR Med Educ*. 2023;9, e48433. https://doi.org/10.2196/48433.
20.. Kumari A, Kumari A, Singh A, et al. Large language models in hematology case solving: A comparative study of ChatGPT-3.5, google bard, and microsoft bing. *Cureus*. 2023. https://doi.org/10.7759/cureus.43861.

21.. Alkhalid Z, Birand N. Determination and comparison of potential drug–drug interactions using three different databases in Northern Cyprus community pharmacies. *Niger J Clin Pract*. 2022;25:2005–2009. https://doi.org/10.4103/njcp.njcp_448_22.

22.. UpToDate Lexicomp. https://www.uptodate.com/contents/table-of-contents/drug-information; 2025.

23.. Medscape''s Drug Interaction Checker. https://reference.medscape.com/drug-interactionchecker; 2025.

24.. Drugs.com. https://www.drugs.com/drug_interactions.html; 2025.

25.. Végh A, Lankó E, Fittler A, et al. Identification and evaluation of drug–supplement interactions in Hungarian hospital patients. *Int J Clin Pharm*. 2014;36:451–459. https://doi.org/10.1007/s11096-014-9923-z.

26. Ábrahám BL. Investigation and identification of drug supplement interactions in a population with unipolar depression. *Eur J Hosp Pharm*. 2017;24:A175–A177.

27. A NBVRLABL Somogyi-Végh. Gyógyszerkölcsönhatások kiszűrésére szolgáló adatbázisok értékelése: ellentmondások és egyezőségek [comprehensive evaluation of drug interaction screening programs: discrepancies and concordances]. *Orv Hetil*. 2015;5.

28.. Firstpagesage.com. https://firstpagesage.com/seo-blog/generative-ai-statistics; 2025.

29.. Sounderajah V, Ashrafian H, Golub RM, et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open*. 2021;11, e047709. https://doi.org/10.1136/bmjopen-2020-047709.

30. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *BMJ*. 2015, h5527. https://doi.org/10.1136/bmj.h5527.

31.. Ong JCL, Chen MH, Ng N, et al. A scoping review on generative AI and large language models in mitigating medication related harm. *NPJ Digit Med*. 2025;8:182. https://doi.org/10.1038/s41746-025-01565-7.

32. 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;13:1563. https://doi.org/10.3390/diagnostics13091563.

33. Vaishya R, Misra A, Vaish A. ChatGPT: is this version good for healthcare and research?, Diabetes & Metabolic Syndrome. *Clini Res Rev*. 2023;17, 102744. https://doi.org/10.1016/j.dsx.2023.102744.

34. Albogami Y, Alfakhri A, Alaqil A, et al. Safety and quality of AI chatbots for drug-related inquiries: a real-world comparison with licensed pharmacists. *Digit Health*. 2024;10. https://doi.org/10.1177/20552076241253523.

35. Bull D, Okaygoun D. Evaluating the performance of ChatGPT in the prescribing safety assessment: implications for artificial intelligence-assisted prescribing. *Cureus*. 2024. https://doi.org/10.7759/cureus.73003.

36. Salama AH. The promise and challenges of ChatGPT in community pharmacy: a comparative analysis of response accuracy. *Pharmacia*. 2024;71:1–5. https://doi.org/10.3897/pharmacia.71.e116927.

37. van Nuland M, Erdogan A, Açar C, et al. Performance of ChatGPT on factual knowledge questions regarding clinical pharmacy. *J Clin Pharmacol*. 2024;64:1095–1100. https://doi.org/10.1002/jcph.2443.

38. Chase A, Most A, Sikora A, et al. Evaluation of large language models' ability to identify clinically relevant drug-drug interactions and generate high-quality clinical pharmacotherapy recommendations. *Am J Health Syst Pharm*. 2025. https://doi.org/10.1093/ajhp/zxaf168.

39.. Sicard J, Montastruc F, Achalme C, et al. Can large language models detect drug–drug interactions leading to adverse drug reactions? *Ther Adv Drug Saf*. 2025;16. https://doi.org/10.1177/20420986251339358.

40.. Fiordelisi M, Masucci S, Bianco A, et al. ChatGPT, alleato del farmacista clinico nella verifica delle herbal-drug interactions: potenzialità e limiti. *Recenti Prog Med*. 2024;115:558–559. https://doi.org/10.1701/4365.43601.

41. Zhang X, Tsang CCS, Ford DD, Wang J. Student pharmacists' perceptions of artificial intelligence and machine learning in pharmacy practice and pharmacy education. *Am J Pharm Educ*. 2024;88, 101309. https://doi.org/10.1016/j.ajpe.2024.101309.

42. Ray PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Syst*. 2023;3:121–154. https://doi.org/10.1016/j.iotcps.2023.04.003.

43. Huang X, Estau D, Liu X, Yu Y, Qin J, Li Z. Evaluating the performance of ChatGPT in clinical pharmacy: a comparative study of ChatGPT and clinical pharmacists. *Br J Clin Pharmacol*. 2024;90:232–238. https://doi.org/10.1111/bcp.15896.

44. Ranchon F, Chanoine S, Lambert-Lacroix S, Bosson J-L, Moreau-Gaudry A, Bedouch P. Development of artificial intelligence powered apps and tools for clinical pharmacy services: a systematic review. *Int J Med Inform*. 2023;172, 104983. https://doi.org/10.1016/j.ijmedinf.2022.104983.

45.. Wastesson JW, Morin L, Tan ECK, Johnell K. An update on the clinical consequences of polypharmacy in older adults: a narrative review. *Expert Opin Drug Saf*. 2018;17:1185–1196. https://doi.org/10.1080/14740338.2018.1546841.

46. Gutiérrez-Igual S, Lucas-Domínguez R, Sendra-Lillo J, Martí-Rodrigo A, Crespo IR, Montesinos MC. Impact of pharmacist-led interventions in identifying and resolving drug related problems and potentially inappropriate prescriptions among rural patients: a pilot study. *Expl Res Clin Soc Pharm*. 2024;16, 100536. https://doi.org/10.1016/j.rcsop.2024.100536.

47.. Suriyapakorn B, Chairat P, Boonyoprakarn S, et al. Comparison of potential drug-drug interactions with metabolic syndrome medications detected by two databases. *PLoS One*. 2019;14, e0225239. https://doi.org/10.1371/journal.pone.0225239.

48.. Monteith S, Glenn T. A comparison of potential psychiatric drug interactions from six drug interaction database programs. *Psychiatry Res*. 2019;275:366–372. https://doi.org/10.1016/j.psychres.2019.03.041.