Available online at *www.ujpronline.com*
Universal Journal of Pharmaceutical Research
An International Peer Reviewed Journal
ISSN: 2831-5235 (Print); 2456-8058 (Electronic)

REVIEW ARTICLE

PERFORMANCE OF AI-GENERATED DRUG–DRUG INTERACTION ALERTS VERSUS PHARMACIST ASSESSMENT: A SYSTEMATIC REVIEW

Seerat Shahzad[1*], Ayesha Afzaal[1], Rida Afzaal[1], Asma Ashraf[1], Hafiz Muhammad Bilal[1], Syed Hamid Hussain Zaidi[1], Abdul Rehman Abid[1], Zarfshan Shahzad[2]

[1]Department of Pharmacy Practice, Faculty of Pharmaceutical Sciences, Lahore University of Biological and Applied Sciences, Lahore, Pakistan. [2]Department of English, Faculty of Languages, National University of Modern Languages, Islamabad, Pakistan.

Article Info:

Article History:
Received: 9 April 2026
Reviewed: 11 May 2026
Accepted: 13 June 2026
Published: 15 July 2026

Cite this article:
Shahzad S, Afzaal A, Afzaal R, Ashraf A, Bilal HM, Zaid SHH, Abid AR, Shahzad Z. Performance of AI-Generated drug–drug interaction alerts versus pharmacist assessment: A systematic review. Universal Journal of Pharmaceutical Research 2026; 11(3): 73-82.
http://doi.org/10.22270/ujpr.v11i3.1575

***Address for Correspondence:**
Dr. Seerat Shahzad, Department of Pharmacy Practice, Faculty of Pharmaceutical Sciences, Lahore University of Biological and Applied Sciences, Lahore, Pakistan.
Tel: +92-3457288899;
E-mail: *seeratshahzad3@gmail.com*

Abstract

Background: Artificial Intelligence (AI) is increasingly being used in the medicine administration and drug–drug interaction (DDI) screening domain in healthcare. However, its reliability in comparison with the pharmacist review is unknown. This systematic review aimed to compare the effectiveness of DDI alerts generated by AI to DDI alerts generated by pharmacists.
Methodology: This review was done in accordance with PRISMA 2020 guidelines and registered with PROSPERO (CRD420251153581). The publications since 2022 were obtained from PubMed, Embase, Scopus, Web of Science, Cochrane Library, IEEE Xplore, and arXiv. The Mixed Methods Appraisal Tool (MMAT) 2018 was used to evaluate the quality of the studies.
Results: Fourteen studies of AI systems like ChatGPT, Gemini, Bing AI, and Claude were selected. AI models performed well in general drug information retrieval and were found to have significant shortcomings when it comes to clinical DDI screening. Low accuracy, inconsistent responses, severity and onset not assessed and clinically important interaction not included were common problems. Pharmacist review and validated databases like Lexicomp and Micromedex were more reliable to assess DDI.
Conclusion: AI systems can provide support for drug information and may be suitable but are currently not sufficient for autonomous drug information screening. However, pharmacist oversight will remain essential to patient safety, and additional studies are required to develop customizable AI systems for pharmacies that incorporate trusted clinical decision support tools.
Keywords: Artificial intelligence, clinical decision support, clinical pharmacy, drug-drug interactions, pharmacist, systematic review.

INTRODUCTION

The safe and effective use of medications represents a fundamental cornerstone of modern healthcare systems worldwide, yet it remains an increasingly formidable challenge in an era of unprecedented therapeutic complexity[1]. This complexity is determined by factors including polypharmacy, aging population, and the introduction of important biological therapies, all of which increase the risk of drug interactions and adverse events[2]. Landmark studies from the Institute of medicine established the medical errors that are leading to cause of death, this finding that remains tragically relevant today[3]. Worldwide, medication errors and adverse drug reactions are represent as substantial public health burden. Lamdmark studies indicates the account for approximately 5-8% of all hospital admissions[4], with more recent analyses evaluate the associated annual costs in United States alone exceed as $40 billion[5]. This complexity of modern pharmacotherapy that continues to increase with the introduction of new molecular substance, biological therapies, and specialized medications with the narrow therapeutic indices, further complicating the medication safety landscape[6]. This is the evidenced by regulatory data showing as steady rise for the approval of complex specialty drugs, which require sophisticated management strategies to ensure the safe and effective use[7]. Between medication safety concerns and preventable harm stemming from drug-drug interactions (DDIs) represent a particularly significant clinical and economic burden that contributing the substantially to patient morbidity, mortality, and avoid hospitalization worldwide[8].

The evolution of drug interaction screening has followed an interesting pace through modern medical

history, review broader technology and scientific advancement[9]. In the early to mid-20th century, DDI knowledge occupy primarily in the expertise of seasoned clinicians and pharmacists who relies on personal experience, limited case reports, and emerge pharmacological principles for identify the potential interactions[10]. This era was specifying by discovery of interactions and typically identify only after adverse events is occurred in patients[11]. The discovery of thioridazine and quinidine's QT-prolonged effects serve as a classic example of this model, where the cardiotoxic risk was recognized after fatal outcomes were reportes in the literature[12]. The turn of the millennial shows exponential growth in computing power and the exposure of the internet, facilitates the development of more refined databases such as Micromedex and Lexicomp that could be updated by more frequently and accessed more readily[13]. This period also reflects the incorporation of screening tools into electronic health records (EHRs), enabling real time interventions while simultaneously introducing challenges such as alert fatigue excessive[14].

Landmark studies have shown that clinicians may override 90% of drug safety alerts significantly reducing their effectiveness[15]. This historical evolution of DDI management reflects an ongoing effort to balance comprehensive screening with clinical usuability, a challenge that continues to persist as new technologies emerge[16]. For decades, the primary defense against DDIs has relied on the advance expertise of clinical pharmacists supported by integrated clinical decision supported system (CDSS) together forming a multi-layer safety network[17]. These healthcare professionals provide an vital cognitive layer, going beyond automated alerts to interpret the risks within the full context of a patient's unique clinical profile including comorbidities, organ functions, genetic factors, concurrent therapies, and therapeutic goals as defines by their standards of practice ("Standards of Practice for Clinical Pharmacists," 2014) [18].

The pharmacist's role in clinical practice also includes DDI, which requires a number of specific skills or knowledge[19]. They have detailed pharmacological knowledge of interaction mechanisms, and they understand the difference between pharmacokinetic and pharmacodynamic interactions, which affect the absorption, distribution, metabolism and excretion of the drugs and which affect the action of the drugs at receptor sites, respectively[20]. Secondly, they use their clinical judgment to evaluate if proposed interactions are relevant, depending on the patient's characteristics (age, renal and hepatic function, genetic polymorphism and disease states)[21]. Third, they integrate the practical knowledge of what medicines are available in the formularies, as well as the costs of alternative therapies recommended by their physicians[22]. This multi-dimensional expertise represents the termination of extensive education, training, and clinical experience that cannot be readily reduced to algorithmic processes that supports most clinical decision support systems[23]. This expert judgment is consistently guide by gold-standard, pharmacist-validated database such as Lexicomp® and Micromedex®, which serve as the reliable, evidence-based sources for DDI information and are recognized the primary source for clinical decision support[24].

These comprehensive resources are carefully developed and continuous updated by teams of clinical pharmacists, physicians, and drug information specialists who systematically review primary litera-ture, case reports, and pharmacological data to generate the evidence-based recommendations[24]. This develop-ment of these resources involves standardized evidence graded-systems, ongoing literature surveillance, and formalized review procedures to ensure the information remains both comprehensive and accurate[25]. The integration of human expertise with validated technology constitute the current 'pharmacist standard', severing as the essential benchmark against which any new innovation must be rigorously evaluated[26]. This is an example of a cooperative model that brings together the strengths of people and technology in a complementary way for safety[27]. It combines the computational power and broad coverage of databases with the clinical judgment and contextual understanding of the experienced pharmacists[28]. This partnership has proven highly effective in avert significant medication-related harm, with their systematic reviews exhibit that pharmacist intervention significantly reduces the adverse drug event[29]. However, it is remain limited by resource restraint as the pharmacist availability cannot always match to the volume of medication requires review, creating a gap between the ideal and practical application of that standard[30].

In the context of pharmacy practice, this technology theoretically offer a several advantages over traditional information sources: 24/7 availability, rapid processing of complex questions, ability to synthesize information from the multiple sources; and the capacity is explain to the concepts in various levels of complexity suitable to different users (patient vs. professionals)[30]. These characteristics suggest to the potential for serving as both clinical decision support tools for their healthcare professionals and educational resources for patients[31]. Collectively, these studies suggest that LLMs may hold valuable applications in domain requiring rapid information retrieval and composite, particularly for educational purposes or as initial screening tools when used with human mistake, a cautious approach strongly advocated for by conveners in the field[32].

The collective evidence paints a regarding picture regarding the preparation of LLMs for DDI screening. These models also shows a critical "contextual blindness", failing to integrate the patient-specific variables required for personalized risk assessment, which sharply limits to their clinical utility[33]. The literature, however, remains fractured and methodologically diverse. Consequently, a specific and urgent gap exists: there is no systematic a quantitative synthesis that resolves these contraindications by rigorously comparing AI performance to the steadfast benchmark of pharmacist expertise. This present review seeks to fill this gap by offering an evidence

based evolution to inform the future of AI in medication safety.

MATERIALS AND METHODS

Study design

The Preferred Reporting Items for Systematic reviews and meta analysis (PRISMA) 2020 guidelines were followed in this systematic review. The study protocol was registered in the international prospective register of systematic reviews (PROSPERO) (registration number: CRD420251153581) before the study commenced there by reducing reporting bias and safeguarding the methodological rigor of the review.

Search strategy

To investigate coverage a systematic search was performed, across the following electronic databases: PubMed/MEDLINE, Embase, Scopus, Web of Science, Cochrane Library, IEEE Xplore, and arXiv (both clinical and technological databases). A combination of controlled vocabulary terms and free-text terms were used in the search strategy, which were associated with artificial intelligence, drug-drug interactions and pharmacist standards. Boolean operators (AND, OR) and proximity commands were employed to get the best of both world's – comprehensive coverage and relevant results. Only articles from January 2022 or newer were searched, to ensure that this covers the latest developments in the field of advanced generative AI.

Inclusion criteria

Only those studies that fulfilled the following criteria were included:

- Assessed artificial intelligence systems developed to screen drug-drug interactions.
- Performed comparisons of AI system outputs with reference standards of pharmacist review or pharmacist-validated databases.
- Quantitative measures of reported performance, e.g. sensitivity, specificity, or other diagnostic accuracy measures.
- Applied comparative study designs, such as diagnostic accuracy research, cross-sectional research, cohort research (prospective or retrospective), or experimental comparisons.

Exclusion criteria

The studies were not included according to the following criteria:

- Concentrated on non-drug interaction activities only e.g. drug-disease interactions or pharmacogenomics.
- Did not have a direct pharmacist-based comparator, where other AI systems were used to compare.
- The study design used was non-comparative in nature such as narrative reviews, scoping reviews, editorials, commentaries, and case reports.
- Published before 2022 due to the fast development of AI capabilities.
- Not published in the English language, because of limits in translation resources.

Data extraction

Standardized form was used in a pilot test to extract the data to ensure uniformity and thoroughness of the data. The authors, publication year, country of origin were extracted from the characteristics of the study, the complete details of artificial intelligence specifications (version of the model, access details), methodological parameters (sample characteristics, study design), reference standards specifications, and the complete outcome data (all reported performance metrics and qualitative findings) were extracted. The extraction process was carried out by two independent reviewers with any discrepancies being resolved by consensus.

Quality assessment

The quality of the methodological framework of the studies contained was determined with the use of the Mixed Methods Appraisal Tool (MMAT) version 2018. The tool was selected due to the applicability that had been developed to the different study methods anticipated in this review that include quantitative descriptive, non-randomized, and mixed methods designs. The MMAT enabled systematic evaluation of five critical methodological qualities of any type of studies, such as elements of sampling approach, suitability of measurements and analysis of data. This was used by two independent research reviewers for each study, and in the event of differing opinions, consensus was reached. The results of this quality analysis were directly applied to interpret the results and the determination of the overall confidence of the synthesized evidence.

RESULTS

A flow diagram of the process of how the studies were selected for this systematic review is shown in Figure 1 (PRISMA). A total of 2036 records were identified by using the database searching method, namely Science Direct, Scopus, Web of Science, PubMed, Google Scholar, ProQuest, and the Directory of Open Access Journals (DOAJ) with 320, 585, 452, 348, 180, 141, and 100 records respectively. Of the 1,296 studies remaining after 713 duplicates and 27 renounced studies were removed, 174 were excluded due to screening failures. Of the 1,296 studies that were left after removing 713 duplicate studies and 27 renounced studies, 174 were excluded because of screening failures. After screening for titles and abstracts, 1180 records were excluded.

116 studies were identified for retrieval and 14 studies were not retrieved (9 conference papers and 5 case papers). An additional 102 studies were evaluated for eligibility. In the full-text assessment, 88 studies were excluded on various reasons, including non-AI generated studies (n=26), absence of pharmacist comparators (n=13), wrong population or setting (n=18), insufficient empirical data (n=15), editorial/protocol/review articles (n=7), irrelevant outcomes (n=7), duplicate publications (n=1), and unavailable full texts (n=1). Finally, the systematic review included 14 studies.

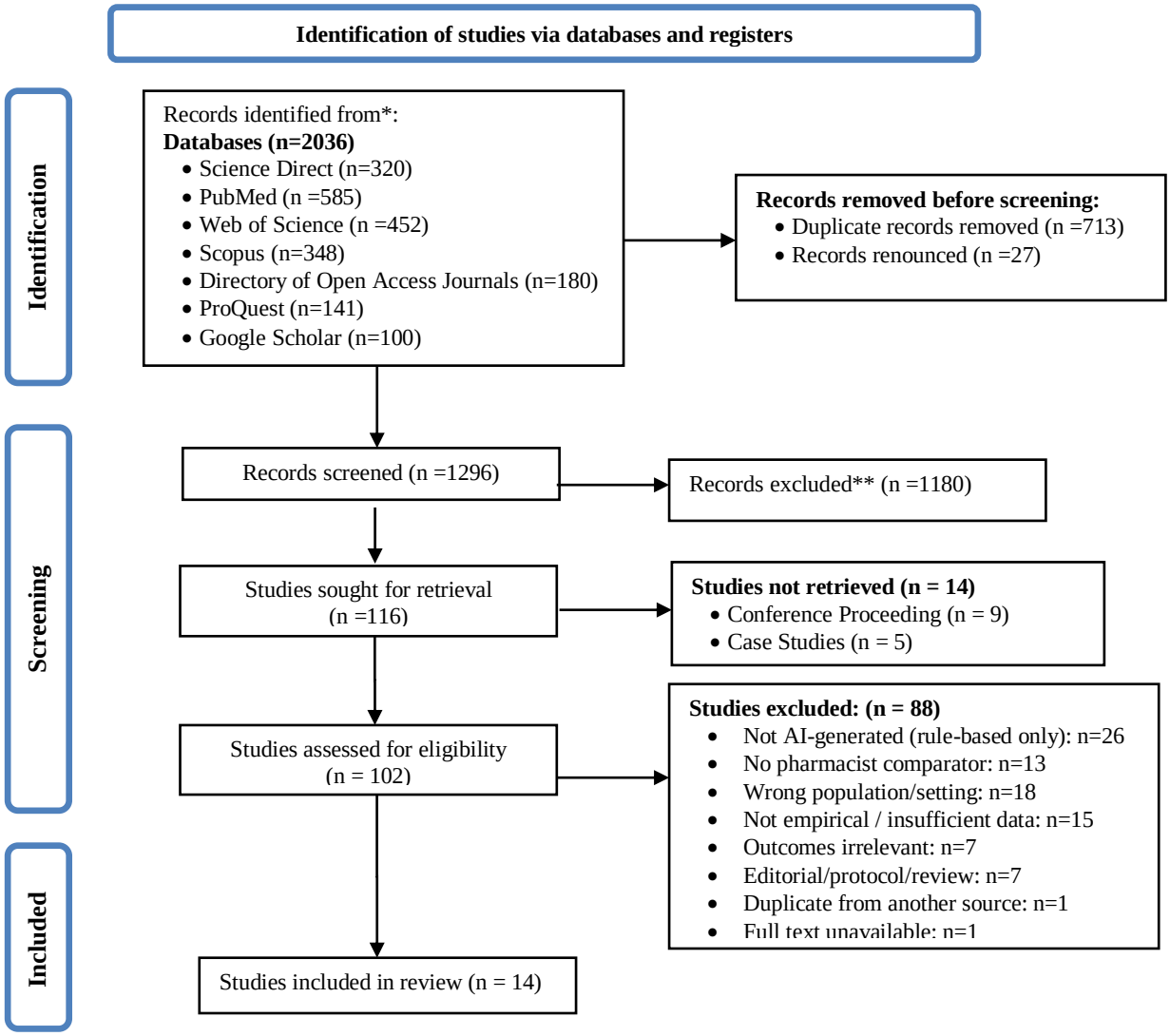

Figure1: PRISMA flow chart of the systematic review.

Characteristics of the study

The systematic review relies on 14 studies that reflect a worldwide research project and 4 continents, with developed and developing systems having a very strong contribution. The geographical spread shows that there is a focused research interest in the United States (3 studies), but the country shows a high level of international involvement by countries such as Saudi Arabia, Nepal, Iraq, Germany, Austria, Jordan, China, the Netherlands, and multi-national partnerships. This popularity highlights the universal fact of the potential of AI to change medication safety practices in diverse healthcare settings and resource environments.

Temporally, the studies included are representative of the fast-changing nature of this field, with 2023-2025 publications. The increased focus in the studies, mostly on the 2024-2025 period (9 out of 14 studies), reflects the growing research interest that coincided with the emergence of powerful Large Language Models (LLMs) in the public domain. This temporal scope underscores the nature of the present systematic review, which incorporates the most recent evidence within the context of rapidly evolving technologies.

The reviewed studies offer a comprehensive overview of the current AI ecosystem as applied to healthcare solutions.Large language models dominate this technology landscape with open AI ChatGPT variations (particularly GPT-3.4,GPT-4 andGPT-4o) emerging as the most widely tested platforms 10 out 14 studies utilize them. The study represents the competitive market of AI by conducting parallel evaluations of Bard/Gemini, Bing AI, and Claude models of Google, Microsoft, and Anthropic.

One of the most useful contributions is made by the studies which performed comparative analysis of several AI systems at the same time. A particularly interesting study tested eight different AI platforms including regional specialists (ERNIE Bot, Doubao) and international leaders (GPT-4o, Gemini-1.5-Pro, Claude-3.5-Sonnet), which allows relevant conclusions about performance differences between different architectural strategies and training methods. Such a comparative method makes a great contribution to our knowledge of the comparative advantages and restrictions of the existing AI environment.

The development of model capabilities can be tracked through the chronological development of research, with subsequent research adopting more developed versions (GPT-4, GPT-4o) having significant empirical performance improvements over previous ones.

Table 1: Summary of included studies.

S. N.	Author, year, Country	AI model/ version	DDI types evaluated	Comparator (Pharmacist Expertise)	Study design and Samples /Patients	Reference Standard	Diagnostic Metrics (AI vs. Pharmacist)	Key recommendations and limitations
1.	Albogami *et al.* (2024)[34] Saudi Arabia	GPT-4, GPT-3.5, GPT-3	Drug interactions and Dosage/	Licensed pharmacists (unstated years)	Cross-sectional. 70 inquiries.	Pharmacist Consensus	Accuracy: 64.3%; Detail: GPT-4 provided proactive risk info in 70% vs. 25.7% by pharmacists	GPT-4 is comparable to pharmacists for safety. Limited to Saudi Arabia.
2.	Thapa *et al.* (2025)[35] Nepal	Chat GPT-4.0	General DDIs	Pharmacist role implied	Cross-sectional. 301 patients.	Micromedex DDI Checker	Occurrence: 100% agreement with Micromedex; Severity: 37.3% concordance; Onset: 65.2%; $p < 0.001$.	ChatGPT-4.0 detects DDI occurrence perfectly. No live pharmacist comparison.
3.	Kim *et al.* (2025)[36] USA	Model: GPT-4, GPT-4o, GPT-3.5	Rx-OTC (43 pairs), Rx-Herbal	Pharmacist role implied	Comparative study. 73 DDI pairs.	Lexicomp (Gold Standard Database)	Accuracy: <50%; Cohen's κ: Poor agreement; GPT-4/4o improved for high-risk DDIs	Off-the-shelf LLMs underperform Lexicomp. No direct pharmacist-AI comparison.
4.	Sulaiman *et al.* (2023)[37] Iraq	Google Bard	Anti Microbials and other Rx	Pharmacist role implied	Cross-sectional. 414 prescriptions.	Lexicomp online	DDI Identification: 68 (Bard) vs. 90 (Lexicomp); κ (risk rating): 0.01 (nil-slight); κ (severity): 0.02 (slight); κ	Lexicomp remains authoritative; Bard lacks precision for clinical use. No live pharmacist comparison.
5.	Radha Krishnan *et al.* (2024)[38] US/UK/	Chat GPT-3.5	General DDIs (hospitalized patients)	Licensed pharmacists (unstated years)	Real-world retrospective. 120 patients	Pharmacist Evaluation	Sensitivity: Low (varies by prompt); Specificity: High; AUC-ROC: Not reported; Cohen's κ: 0.077–0.143	ChatGPT-3.5 lacks reliability for DDI detection. Small sample; single AI version tested
6.	Stroop *et al.* (2025)[39] Germany	GPT-4	Pain therapy DDIs	Physicians & pharmacists (unstated years)	Cross-sectional survey. Not specified	German Drug Directory	Accuracy: 96%; Completeness: 63%; SUS Score: 83.38	GPT-4 is accurate but incomplete for pain therapy DDIs. No sensitivity.
7.	Shin *et al.* (2024)[40] USA	Chat GPT-4	Drug interactions	Community pharmacists (unstated years)	Case-based evaluation. Not specified	Pharmacist Evaluation	Qualitative agreement: Satisfactory for drug info; Comparable for rash/eyelid cases	LLMs show promise but require validation for diverse tasks. No quantitative metrics.
8.	Bischof *et al.* (2025)[41] Vienna	ChatGPT (version un-specified)	Contraindicated severe	CDSS (MediQ, Lexicomp, Micromedex)	Retrospective Cohort. 30 patients	Multi disciplinary team	CDSSs: 280 pDDIs (3 contraindicated, 13 severe, 264 moderate) ChatGPT: 80 pDDIs	ChatGPT not recommended for pDDI screening. Small sample size.
9.	Al-Ashwal *et al.* (2023)[42] Jordan	ChatGPT-3.5 ChatGPT-4 Microsoft Bing AI	255 clinically relevant drug pairs	Conventional DDI tools	Comparative diagnostic accuracy study	Standard DDI reference tools	Bing AI (Accuracy: 0.788-0.890), ChatGPT-3.5/4 showed highest accuracy fluctuations.	Bing AI demonstrated best overall performance, Unspecified reference standard tools,
10.	Khatri *et al.* (2025)[43] United States	ChatGPT-3.5 & ChatGPT-4	Broad Drug Information Questions	Lexicomp	Serial Cross-Sectional. 30 questions	Lexicomp (Wolters Kluwer)	Accuracy: • ChatGPT-3.5: 30% (9/30) • ChatGPT-4: 40% (12/30)	Healthcare professionals should use ChatGPT with caution. No direct pharmacist adjudication.
11.	Li *et al.* (2025)[44] China	AI systems: Kimi; GPT-4o ERNIE Bot;	Included in Prescription Review	6 experienced clinical pharmacists	Mixed Methods 48 clinically. validated questions	Pharmacist consensus on predefined criteria	Overall Performance (Mean Composite Score): • Best: DeepSeek-R1 (9.3-9.4/10)	Generative AI is a promising assistant tool but requires human oversight due to high-risk errors.
12.	Salama *et al.* (2024)[45] UAE	ChatGPT	Drug-drug interactions	Standard answers	Comparative evaluation. 80 questions	Unspecified "standard answers"	DDI Screening: 30% (6/20 correct), Adverse Effects: 65%, Drug Dosage: 35%,	AI models like ChatGPT require significant enhancement. Unspecified ChatGPT version.
13.	Roosan *et al.* (2023)[46] United States	ChatGPT-4.0	Included in MTM cases	Two clinical pharmacists who validated case difficulty	Case-based evaluation. 39 patient cases	Unspecified answers by 2 clinical pharmacists	Accuracy: 100% (39/39 cases); Key outcomes: Correctly detected drug interactions and generated therapy recommendations.	ChatGPT-4.0 may assist pharmacists in MTM. Extreme outlier result.
14.	Nuland *et al.* (2024)[47] Netherlands	ChatGPT	Broad clinical pharmacy knowledge	Pharmacist's performance.	Comparative cross-sectional. 264 questions	Answer key from the Dutch clinical pharmacist	Accuracy: • ChatGPT: 79% • Pharmacists (2022 avg): 66%	Overall, ChatGPT performs better on factual knowledge questions. Tests knowledge, not clinical.

Nonetheless, one of the major methodological issues that can be identified in a number of studies consists of the inability to indicate the version of AI that was tested, which is a serious gap in the area of reproducibility because the performance of the model iterations varies significantly. The studies included are quite diverse in terms of methodological approaches, as they use different methods to answer the main research question of AI competency in DDI screening. Cross-sectional analysis (8 studies) is the most common research design and in this case AI performance is compared to reference standards at one time. Comparative diagnostic accuracy studies, retrospective cohort studies and case-based evaluations complement this approach.

The unit of analysis can differ quite widely in dependence on the study, as it is associated with various possible implementation scenarios:

Drug-pair studies (as many as 255 drug pairs) allow the knowledge breadth of DDI to be evaluated systematically. Problems by simulated clinical scenarios (70-80 inquiries) are used to determine practical problem solving skills. Studies that use actual or simulated patient situations (30-414 patients) are used to assess performance in clinical settings of relevance. Studies that encompass DDI screening in the wider context of medication therapy management (MTM).

Although such a methodological heterogeneity makes synthesis difficult, it offers complementary information about AI performance in various possible applications, ranging between rapid reference checking and full-scale medication review. The main strength of this evidence base is that the pharmacist-driven reference standards were used consistently, which confirms the fundamental comparator framework of this systematic review. The research makes use of three major groups of reference standards: Six reports were based on the use of standard clinical decision support systems (Lexicomp, Micromedex, Drugs.com) as reference standards.These databases are a reflection of institutionalized pharmacist knowledge, and include evidence-based advice that has been formulated through a systematic literature surveillance and formal review process by clinical pharmacists and physicians.

Five of these studies utilized direct evaluation by licensed pharmacists, with some being individual review of a single pharmacist and others being formal consensus evaluation by multiple clinical pharmacists. This method best reflects the clinical judgment which is the contemporary standard of care. In practice, clinical decision making is a collaborative process and two studies used multidisciplinary team evaluation, a combination of pharmacist, physician and other specialists' assessments of reference standards.The fact that such rigorous reference standards are applied in the majority of studies contributes greatly to the validity of the results and allows to meaningfully compare AI performance with the existing best practices.

The overall description of the studies included sheds light on a research area that is characterized by both potential and serious methodology issues. A number of common limitations affect the overall evidence base, such as the size of samples used to restrict statistical power and generalizability, inconsistent reporting of key parameters, such as AI model versions and immediate engineering approaches, and a bias towards detectivity rather than clinical practice and workflow support. This high level of methodological heterogeneity, though indicative of an emerging and fast developing field, poses a critical inhibitor to making convergent conclusions. This fragmentation and the existing evidence gaps are what makes the rigorous, standardized synthesis of this systematic review necessary and the following meta-analysis of more specific and generalizable estimates of AI performance relative to the standard, pharmacist-led care.

Mixed methods appraisal tool

Among the selected studies, the quality of methodology was critically appraised using Mixed Methods Appraisal Tool (MMAT) 2018. Given the diverse nature of the evidence base, the selection of the MMAT was appropriate, as it measures the methodological rigour of various forms of study (qualitative, quantitative (RCT, non-RCT, descriptive) and mixed methods), all of which were represented.The review was done by two reviewers who did it independently. Both studies were compared to two screening question and then five study-specific criteria were used based on the design. The answers were taken as Yes, No or Can not tell. Any inconsistencies among reviewers were to be addressed by discussion in order to come up with a consensus.

The methodological quality of the included studies was assessed using the Mixed Methods Appraisal Tool 2018. All 14 studies met the two initial MMAT screening criteria, as they had clear research questions and collected data that were appropriate for addressing those questions. Most of the included studies were quantitative descriptive in design, while a smaller number included quantitative non-randomized, qualitative, or mixed-methods components. Overall, the methodological quality of the evidence was moderate. Common strengths across the included studies included relevant sampling strategies, appropriate measurement of outcomes, and the use of pharmacist review, pharmacist consensus, pharmacist-validated databases, or multidisciplinary expert assessment as reference standards. These elements strengthened the relevance of the evidence for comparing AI-generated outputs with pharmacist-led or pharmacist-validated standards.

Several methodological limitations were also identified. These included incomplete reporting of sample representativeness, unclear assessment of non-response bias, limited control of confounding in some studies, and insufficient reporting of important AI-related methodological details such as model version, prompt structure, evaluation protocol, and reproducibility procedures. These issues limited comparability between studies and reduced the certainty of the overall conclusions.

Taken together, the MMAT appraisal supports a cautious interpretation of the findings. Although the included studies provide useful evidence regarding the

performance of AI systems in DDI screening and related clinical pharmacy tasks, methodological heterogeneity and reporting gaps indicate the need for more standardized future research.

DISCUSSION

This systematic review focused on the performance of the AI-generated drug-drug interaction (DDI) alerts with pharmacists. The 14 studies synthesis demonstrates a field characterized by a high rate of innovation, as well as the high level of limitations and persistent ones. The major theme that appears is that of unfulfilled potential: although AI proves to be exceptionally good at information retrieval, its existing implementation to the high-stakes DDI-screening process is undermined by the crucial lack of reliability, clinical sensitivity, and safety. We can conclude that AI is, as of today, incapable of substituting the skills of a pharmacist, yet, through a cautious use of this technology, it can become a useful tool.

One of the main problems of the interpretation of the literature is the seeming contradiction between the studies. This review solves this by defining the situations in which AI is most effective and ineffective, which is essential to determine its role in pharmacy.

Investigations by Van Nuland[47] and Albogami[34], has a positive perspective and demonstrates that AI may be safer and more proactive in providing accurate information than pharmacists would be on questions of factual knowledge. This is the main strength of AI as a strong information retrieval and education tool. The fact that it can quickly create information based on its extensive training corpus makes it very useful in answering specific pharmaco-logical questions or providing patient-friendly explana-tions, which may give pharmacists additional time to perform more complex tasks.

Yet, this strength has no reliable transfer to the dynamic and high stakes task of detecting clinical DDI. A larger amount of data studies by Bischof *et al.*[41], Salama[45], and Kim *et al.*[36], shows deep flaws. When Bischof discovered that ChatGPT failed to identify the important QTc-prolonging interactions and Salama found that ChatGPT only detected DDI screening with 30 percent accuracy, this is a pattern of failure in clinical reasoning in practice. This is further developed in the study by Thapa *et al.*[35], which illustrates an essential lack of connection: AI systems are often able to identify an interaction, but struggle to accurately measure its severity (37.3% accuracy) or to predict when it will occur (65.2% accuracy) – the exact information needed to make clinical decisions.

This dichotomy gives rise to two groups of AIs: an AIs that works well as a knowledge store and another one that fails as a clinical decision-maker. The favourable results are mostly relevant to the former, i.e. pattern recognition and recall of information based on its training data. The latter negative results reveal the weaknesses of the latter, which needs to be judged in a particular setting, stratify risk, and apply it to specific patient cases. This contradiction is therefore solved by acknowledging the fact that performance is very task specific. It is possible that AI is prepared to be used as a fast-access source or educational tool, but it is not prepared to make independent clinical decisions in the field of DDI management. In addition to simple accuracy measures, this review determines certain, interconnected failure modes that restrict the clinical utility and safety of existing AI systems to a serious degree.

The most serious flaw of any clinical tool is the absence of a life-threatening event. The inability of the models to reliably detect QTc-prolonging DDIs[41] is a fatal blind spot. This is augmented by the fact that the use of LLMs has been documented to engage in hallucinating or confabulating information. A model may not just fail to capture an actual interaction, but may even create a non-existent one, or a management strategy that may sound plausible but which is totally incorrect. This puts the clinicians in a situation of a double jeopardy that they will not trust the occurrence or absence of an AI-generated alert.

AI models tend to generate long, confident-sounding statements that give the impression of knowledge. Nonetheless, as a study by Li *et al.*[44], it may cover up important omissions and a basic lack of contextual sensitivity. As opposed to a pharmacist who considers the renal functionality, hepatic condition, age, genetic makeup and comorbidities of a patient, AI gives generalized, one-fit-all responses. As an example, the interaction between a renally excreted drug can be moderate in the case of a healthy adult and severe in the case of an elderly patient with chronic kidney disease-a difference that AI does not reliably distinguish. This is not personalized making its outputs clinically insignificant.

Total 90 percent inconsistency rate cited by Bischof *et al.*[41], is the opposite of the medical practice. A pharmacist should have the capacity to believe that the same clinical question is going to have the same evidence-based response at all times. The unreliability of LLMs is devastating since the prompt can produce random results due to its probabilistic characteristics. This is compounded by the black box problem; in many cases, it is impossible to find out why an AI provided a particular response, and it is hard to test its reasoning, as well as learn its lessons, which further weakens clinical trust.

This review should be understood considering the significant methodological weaknesses and heterogeneity of the original studies. Although such variability is a sign of an emerging field, it makes it difficult to draw homogenous conclusions. Due to its fast iterative nature (e.g. GPT-3.5 to GPT-4o), a study published in 2023 can be assessing a far less able system in the year 2025. Numerous studies did not indicate the specific version that was tested and this seriously affected reproducibility. The performance of AI is extremely sensitive to the phrasing of a question. The results of a study conducted with complex, clinically based prompts will not be the same as those obtained with simple queries. This was a variable that was hardly controlled or reported in details. This indicates a potential methodological bias such as

unblinded assessor or less demanding case scenarios that can blow up performance measures. This heterogeneity not only prevented the realization of formal meta-analysis, but it also highlights the necessity to implement standardized evaluation frame-works (like the CONSORT-AI extension) in future studies. It also implies that the efficacy of any particular AI tool should be tested on the local clinical setting before it is allowed to be used.

Considering the shortcomings of AI, this review is a strong support of the necessity of the clinical pharmacist. The pharmacist-standard is not a database-search, it is a complex, multi-dimensional cognitive task. It involves examining the quality of evidence of a possible DDI. The pharmacist tailors the risk, depending on the health profile of the patient, including age and kidney functionality and other conditions. The pharmacist is then faced with a clinical judgment, balancing the possible harm of an interaction with the vital good that the medication does to the patient. Development of action management plans.

Li used direct pharmacist evaluation as the reference standard and all found the performance of AI to be wanting[44]. This highlights the fact that the gold standard is still the use of professional human judgment to validate and pharmacist edited databases such as Lexicomp and Micromedex. The interplay between the clinical reasoning of the pharmacist and these sources of authority is a strong, valid, and patient-centred care that the existing AI will never be able to replace[47]. The pharmacist role is therefore becoming more of a clinical integrator and decision-maker rather than an information verifier who is also able to take advantage of technology and exercise an irreplaceable human judgment.

Future research implications

In our findings, we provide the following evidence-based recommendations to manoeuvre through the given landscape and lead on the future development.

For Clinical Drug Drug Interactions (DDIs), we strongly discourage using large language models (LLMs) like ChatGPT or Google Gemini, where they are publicly and readily available.The risk of not identifying interactions or giving false advice is too high to continue using them. The dangers of the dangerous omissions, hallucinations, and inconsistencies are too high to be approved of patient care. Pharmacist-led review should be the standard practice and should not be set aside because of the availability of validated clinical decision support systems (CDSS) in EHRs. In case of the AI application, the role of the latter must be entirely supportive. Potential applications of low-risk could be drafting of patient education material, and all results could be strictly checked and placed in context by a qualified pharmacist.

For future Research, development and Policy, create and test pharmacy-specific AI. Shift to specialized AI applications that work with trusted data on drugs and particular pharmacy work, and not general-purpose chatbots. One of the future directions is to combine LLMs with old-fashioned, rule-oriented CDSS. The rule-based system would serve as a safety net by making sure that critical DDIs are not overlooked, whereas the LLM would be able to help with complex reasoning, literature summarization, and communication with the patients. The field needs to shift toward a standardized set of reporting policies including what AI model was applied, how it was activated, and in which cases it was not able to do so, so that comparisons can be made with other studies. This will enable making of meaningful comparisons and cumulative knowledge building. The research should transition beyond accuracy alone and investigate the possibilities of safely and smoothly introducing AI tools into clinical workflows to reduce cognitive load and improve the efficiency. Finally, there needs to be research conducted to quantify the effects of the AI-supported pharmacy on the bottom line on patient outcomes and adverse drug events as well as hospital readmission rates.

Limitations of the study

There are a number of limitations of this systematic review that should be taken into account when interpreting the results. First, there was significant methodological variation in the studies included, both in terms of the AI models assessed, as well as the study design, the reference standards, the outcome measures, and reporting methods. Such methodologic differences made it difficult to make direct comparisons between studies and prevented the use of quantitative meta-analysis. Second, the majority of these studies used simulated clinical scenarios, drug-pair assessments or structured questionnaires instead of clinical implementation, which may reduce the generalizability of the results. Third, some studies failed to provide explicit information on the versions of AI models employed, the strategies applied, and/or the evaluation protocols, limiting the replicability and comparability of the findings. Fourth, only English-language publications from 2022 forward were included, which could have led to language and publication bias. Lastly, because artificial intelligence is a field that is constantly evolving, the capabilities of existing models could change from the ones outlined in the studies included in this review and the results may not be seen as representative of the most current advances in AI.

CONCLUSIONS

Based on the synthesis of the evidence in general, the current state of functionality of the artificial intelligence as an information retrieval and learning tool is not adequate and safe to conduct autonomous clinical decision-making in screening of drug-drug interaction due to the severe gap in its reliability, clinical sensitivity, and contextual reasoning. The findings, including some adjustments like the methodological heterogeneity, absence of reporting on AI versions, and high proportion of simulated tests compared to real-life tests, are clear indicators of the need to have the clinical pharmacist as the gold standard. Integration The future of integration hinges on the development of hybrid, pharmacy-specific AI tools, which run under the watch of pharmacists and the focus of research, has shifted to standardized

assessment, workflow integration, and ultimately, the impact on hard patient outcomes.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude towards the department of Pharmacy Practice, Faculty of Pharmaceutical Sciences, Lahore University of Biological and Applied Sciences, Lahore, Pakistan for all the support and co-operation provided in the course of this research.

AUTHOR'S CONTRIBUTIONS

Shahzad S: conceptualization, methodology, literature search, writing. **Afzaal A:** study screening, data extraction. **Afzaal R:** literature search, data extraction. **Ashraf A:** quality assessment, analysis. **Bilal HM:** writing, manuscript review. **Zaidi SHH:** methodology, manuscript review. **Abid AR:** supervision, manuscript review. **Shahzad Z:** editing and proofreading. Final manuscript was checked and approved by all authors.

DATA AVAILABILITY

The associated author can provide the empirical data used to support the study's conclusions upon request.

CONFLICT OF INTEREST

There are no conflicts of interest associated with this work.

REFERENCES

1. Ali HM, Doghish MA, Mageed AS, *et al*. Innovations in biosensor technologies for healthcare diagnostics and therapeutic drug monitoring: Applications, recent progress, and future research challenges. Sensors (Basel, Switzerland) 2024; 24(16): 5143. *https://doi.org/10.3390/s24165143*
2. Aggarwal P, Woolford SJ, Patel HP. Multi-morbidity and polypharmacy in older people: Challenges and opportunities for clinical practice. Geriatrics 2020; 5(4): 85. *https://doi.org/10.3390/geriatrics5040085*
3. Shojania KG, DIxon-Woods M. Estimating deaths due to medical error: The ongoing controversy and why it matters. BMJ Qual Saf 2017; 26(5):423-428. *https://doi.org/10.1136/bmjqs-2016-006144*
4. Mekonnen AB, Alhawassi TM, McLachlan AJ, *et al*. Adverse drug events and medication errors in African Hospitals: A systematic review. Drugs - Real World Outcomes 2017; 5(1):1-24. *https://doi.org/10.1007/s40801-017-0125-6*
5. Grosse SD, Nelson RE, Nyarko KA, *et al.* The economic burden of incident venous thromboembolism in the United States: A review of estimated attributable healthcare costs. Thromb Res 2016; 137:3-10. *https://doi.org/10.1016/j.thromres.2015.11.033*
6. Prajapati RN, Bhusan B, Singh K, *et al*. Recent advances in pharmaceutical design: Unleashing the potential of novel therapeutics. Cur Pharm Biotech 2024; 25:2024. *https://doi.org/10.2174/0113892010275850240102105033*
7. Gupta DK, Tiwari A, Yadav Y, *et al*. Ensuring safety and efficacy in combination products: Regulatory challenges and best practices. Front Med Tech 2024; 6:1377443. *https://doi.org/10.3389/fmedt.2024.1377443*
8. Zheng WY, Richardson LC, *et al*. Drug-drug interactions and their harmful effects in hospitalised patients: A systematic review and meta-analysis. Eur J Clin Pharmacol 2017; 74(1):15-27. *https://doi.org/10.1007/s00228-017-2357-5*
9. Gittelman M. The revolution re-visited: Clinical and genetics research paradigms and the productivity paradox in drug discovery. Res Policy 2016; 45(8): 1570-1585. *https://doi.org/10.1016/j.respol.2016.01.007*
10. Abougalambou SSI, Alenezi TN. Knowledge and information sources of potential drug-drug interactions of healthcare professionals among Buraydah Hospitals. J Pharm Policy Pract 2023; 16(1); 131. *https://doi.org/10.1186/s40545-023-00642-0*
11. Cai R, Liu M, Hu Y, *et al.* Identification of adverse drug-drug interactions through causal association rule discovery from spontaneous adverse event reports. Artif Intell Med 2017; 76: 7. *https://doi.org/10.1016/j.artmed.2017.01.004*
12. Nachimuthu S, Assar MD, and Schussler JM. Drug-induced QT interval prolongation: Mechanisms and clinical management. Ther Adv Drug Saf 2012; 3(5): 241. *https://doi.org/10.1177/2042098612454283*
13. Yamin M. Information technologies of 21st century and their impact on the society. Int J Inf Tech 2019; 11(4): 759. *https://doi.org/10.1007/s41870-019-00355-1*
14. McGreevey JD, Mallozzi CP, Perkins RM, *et al.* Reducing alert burden in electronic health records: State of the art recommendations from four health systems. Appl Clin Inform 2020; 11(1): 1. *https://doi.org/10.1055/s-0039-3402715*
15. Rush JL, Ibrahim J, Saul K, *et al*. Improving patient safety by combating alert fatigue. J Grad Med Educ 2016; 8(4): 620. *https://doi.org/10.4300/JGME-D-16-00186.1*
16. Ren HC, He C, Wan H. Assessing DDI risks in drug discovery and development: Approaches, challenges, and trends. Expert Opin Drug Metab Toxicol 22(1): 65–82. *https://doi.org/10.1080/17425255.2025.2607019*
17. Sutton RT, Pincock D, Baumgart DC, *et al.* An overview of clinical decision support systems: Benefits, risks, and strategies for success. NPJ Digit Med 2020; 3(1):17. *https://doi.org/10.1038/s41746-020-0221-y*
18. Fahim YA, Hasani IW, Kabba S, *et al*. Artificial intelligence in healthcare and medicine: Clinical applications, therapeutic advances, and future perspectives. Eur J Med Res 2025; 30(1): 848. *https://doi.org/10.1186/s40001-025-03196-w*
19. Russ-Jara AL, Elkhadragy N, Arthur KJ, *et al*. Cognitive task analysis of clinicians' drug-drug interaction management during patient care and implications for alert design. BMJ Open 2023; 13(12): e075512. *https://doi.org/10.1136/bmjopen-2023-075512*
20. Rajput A. Investigation of mechanisms and clinical implications of drug interactions: A review. Trends Pharm Nanotechnol 6: 41-54.
21. Lauschke VM, Ingelman-Sundberg M. The importance of patient-specific factors for hepatic drug response and toxicity. Int J Mol Sci 2016; 17(10): 1714. *https://doi.org/10.3390/ijms17101714*
22. Johnson ST, Kier KL, Douglas JS, *et al.* Formulary management challenges and opportunities: 2020 and beyond - An opinion paper of the drug information practice and research network of the American College of Clinical Pharmacy. JACCP J Am Coll Clin Pharm 2021; 4(1): 81-91. *https://doi.org/10.1002/jac5.1332*
23. Van Baalen S, Boon M, and Verhoef P. From clinical decision support to clinical reasoning support systems. J Eval Clin Pract 2021; 27(3): 520-528. *https://doi.org/10.1111/jep.13541*
24. Mountford CM, Lee T, De Lemos J, *et al.* Quality and usability of common drug information databases. Can J Hosp Pharm 2010; 63(2): 130. *https://doi.org/10.4212/cjhp.v63i2.898*
25. Alexander KM, Waer ML, Field C, *et al.* Advancing systems-based patient workup skills through redesign of a critical care pharmacy elective. Am J Pharm Educ 2026; 90(3): 101950. *https://doi.org/10.1016/j.ajpe.2026.101950*
26. Alqahtani SS, Menachery SJ, Alshahrani A, *et al.* Artificial intelligence in clinical pharmacy-A systematic

review of current scenario and future perspectives. Digit Heal 2025; 11: 20552076251388144. *https://doi.org/10.1177/20552076251388145*

27. Wu PPY, Fookes C, Pitchforth J, *et al*. A framework for model integration and holistic modelling of socio-technical systems. Decis Support Syst 2015; 71: 14-27. *https://doi.org/10.1016/j.dss.2015.01.006*
28. Mertens JF, Kempen TGH, Koster ES, *et al.* Cognitive processes in pharmacists' clinical decision-making. Res Soc Adm Pharm 2024; 20(2): 105-114. *https://doi.org/10.1016/j.sapharm.2023.10.007*
29. Kelly WN, Ho MJ, Smith T, *et al*. Association of pharmacist interventions with adverse drug events and potential adverse drug events. Pharmacoepidemiol Drug Saf 2024; 33(7): e5853. *https://doi.org/10.1002/pds.5853*
30. Tariq RA, Vashisht R, Sinha A, *et al*. Medication Dispensing Errors and Prevention. StatPearls 2024. *https://www.ncbi.nlm.nih.gov/books/NBK519065/.*
31. Hill A, Morrissey D, Marsh W. What characteristics of clinical decision support system implementations lead to adoption for regular use? A scoping review. BMJ Heal Care Informatics 2024; 31(1): e101046. *https://doi.org/10.1136/bmjhci-2024-101046*
32. Gong EJ, Bang CS, Shin YS. Applications of large language models in medical research: From systematic reviews to clinical studies. Bioeng 2026; 13(3): 365. *https://doi.org/10.3390/bioengineering13030365*
33. Ke Y, Yang R, Lie SA *et al*. Mitigating cognitive biases in clinical decision-making through multi-agent conversations using large language models: Simulation study. J Med Internet Res 2024; 26(1): e59439. *https://doi.org/10.2196/59439*
34. Albogami Y, Alfakhri A, Alhossan A, *et al.* Safety and quality of AI chatbots for drug-related inquiries: A real-world comparison with licensed pharmacists. Digit Heal 2024; 10. *https://doi.org/10.1177/20552076241253523*
35. Thapa R, Karki S, Shrestha S, *et al.* Exploring potential drug-drug interactions in discharge prescriptions: ChatGPT's effectiveness in assessing those interactions. Elsevier 2026. *https://doi.org/10.1016/j.rcsop.2025.100564*
36. Kim J, KinCaid JWR, Rao AS, *et al*. Risk stratification of potential drug interactions involving common over-the-counter medications and herbal supplements by a large language model. J Am Pharm Assoc 2025; 65(1): 102304. *https://doi.org/10.1016/j.japh.2024.102304*
37. Sulaiman DM, Shaba SS, Almufty HB, *et al.* Screening the drug-drug interactions between antimicrobials and other prescribed medications using google bard and Lexicomp® OnlineTM Database. Cureus 2023; 15(9). *https://doi.org/10.7759/cureus.44961*
38. 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(12): 3361-3366. *https://doi.org/10.1111/bcp.16275*
39. Stroop A, Stroop T, Zawy Alsofy S, *et al.* Assessing GPT-4's accuracy in answering clinical pharmacological questions on pain therapy. Br J Clin Pharmacol 2025; 91(8):2294-2303. *https://doi.org/10.1002/bcp.70036*
40. Shin E, Hartman M, Ramanathan M. Performance of the ChatGPT large language model for decision support in community pharmacy. Br J Clin Pharmacol 2024; 90(12): 3320-3333. *https://doi.org/10.1111/bcp.16215*
41. Bischof T, Al Jalali V, Zeitlinger M*et al*. Chat GPT vs. clinical decision support systems in the analysis of drug-drug interactions. Clin Pharmacol Ther 2025; 117(4): 1142-1147. *https://doi.org/10.1002/cpt.3585*
42. 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: 137-147. *https://doi.org/10.2147/DHPS.S425858*
43. Khatri S, Sengul A, Moon J, *et al.* Accuracy and reproducibility of ChatGPT responses to real-world drug information questions. JACCP J Am Coll Clin Pharm 2025; 8(6): 432-438. *https://doi.org/10.1002/jac5.70038*
44. L. Li, Du P, Hiang X, *et al*. Comparative analysis of generative artificial intelligence systems in solving clinical pharmacy problems: Mixed methods study. JMIR Med Informatics 2025;13: 1-14. *https://doi.org/10.2196/76128*
45. 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*
46. 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 2024; 64(2): 422-428.e8. *https://doi.org/10.1016/j.japh.2023.11.023*
47. Nuland MV, Erdogun V, Aca C, *et al*. Performance of ChatGPT on factual knowledge questions regarding clinical pharmacy. J Clin Pharmacol 2024; 64(9): 1095-1100. *https://doi.org/10.1002/jcph.2443*