JHPHS
Journal of Hospital Pharmacy and Health Services

Original Paper

Open Access

Capacity of ChatGPT, Deepseek, and Gemini in predicting major potential drug interactions in adults within the Intensive Care Unit

Tácio Mendonça LIMA[1]

[1]Departamento de Farmácia e Administração Farmacêutica, Universidade Federal Fluminense, Niterói, RJ, Brasil.

Corresponding author: Lima TM, taciolima@id.uff.br

Submitted: 10-02-2025 Resubmitted: 12-03-2025 Accepted: 12-03-2025

Double blind peer review

Abstract

Objective: evaluate the ability of the ChatGPT v.3.5, DeepSeek v-3, and Gemini 2.0 flash to accurately predict major potential drug interactions (DIs) in critically ill patients. **Methods:** A list of 20 DIs was compiled from previously published literature. The Micromedex and Drugs.com databases were used as references. A specific prompt was designed to interact with the tools. The generated responses were stored for subsequent analysis by a pharmacist. Specificity, sensitivity, negative predictive value (NPV), positive predictive value (PPV), accuracy, and agreement were calculated for each tool based on the responses regarding DDI severity, which were categorized into five levels: contraindicated, major, moderate, minor, and no interaction. Additionally, the responses related to the mechanism of action and recommended management for each DDI were categorized as "adequate and accurate," "adequate but inaccurate", and "inadequate." **Results:** When the Micromedex was used as a reference, ChatGPT performed better, achieving an accuracy rate of 75%, while DeepSeek and Gemini scored 70% and 65%, respectively. Overall, there was an improvement in the performance of all tools when Drugs.com was used as the reference, with accuracy rates of 80% for DeepSeek and 75% for both ChatGPT and Gemini. However, the agreement on the severity of DDIs between the tools and references was 0.354 (weak) for Drugs.com and 0.410 (moderate) for Micromedex. In general, two "inadequate" responses and 10 "adequate but inaccurate" responses regarding the mechanism of action and recommended management were observed when compared with Micromedex (14 DDIs analyzed), while eight "inadequate" responses and 21 "adequate but inaccurate" responses were found when compared with Drugs.com (17 DDIs analyzed). **Conclusion:** The tools analyzed show promise to assist healthcare professionals in predicting DDI in adults hospitalized in the intensive care unit (ICU). However, their use should be approached with caution, as they may generate incorrect/inaccurate information. Additional advancements are required to ensure their reliable application in clinical practice.

Keywords: Artificial Intelligence, ChatGPT, Deepseek, Gemini, Drug Interactions, Intensive Care Units.

Capacidade do ChatGPT, Deepseek e Gemini em prever as principais interações medicamentosas potenciais em adultos internados em Unidade de Terapia Intensiva

Resumo

Objetivo: analisar a capacidade das ferramentas ChatGPT v.3.5, DeepSeek v-3 e Gemini 2.0 flash em prever as principais interações medicamentosas (IM) potenciais encontradas em pacientes críticos. **Métodos:** uma lista de 20 IMs foi elaborada a partir da literatura previamente publicada. Utilizou-se as bases de dados Micromedex e Drugs.com como referência. Para interagir com as ferramentas, elaborou-se um comando de entrada específico. As respostas foram registradas para análise posterior por um farmacêutico. Foram calculados os parâmetros de especificidade, sensibilidade, valor preditivo negativo (VPN), valor preditivo positivo (VPP), acurácia e concordância para cada ferramenta, com base nas respostas referentes à gravidade das interações medicamentosas (IM), as quais foram categorizadas em cinco níveis: contraindicada, maior, moderada, menor e sem interação. Respostas geradas pelas ferramentas relacionadas ao mecanismo de ação e conduta a ser tomada frente a uma IM foram categorizadas em "adequadas e precisas", "adequadas e imprecisas" e "inadequadas". **Resultados:** Quando comparadas com o Micromedex, o ChatGPT obteve um melhor desempenho, com uma taxa de acurácia de 75%, enquanto o DeepSeek e Gemini obtiveram taxas de 70% e 65%, respectivamente. Houve uma melhoria geral no desempenho de todas as ferramentas quando o padrão de referência foi o Drugs.com, com taxas de acurácia de 80% para o DeepSeek e 75% para o ChatGPT e Gemini. Por outro lado, a concordância das informações sobre gravidade das IM entre as ferramentas e as referências foi de 0,354 (fraca) para Drugs.com e 0,410 (moderada) para Micromedex. De forma geral, foram observadas duas respostas "inadequadas" e 10 "adequadas e imprecisas" sobre mecanismo de ação e conduta ao comparar com o Micromedex (14 IMs analisadas) e oito "inadequadas" e 21 "adequadas e imprecisas" em comparação com Drugs.com (17 IMs analisadas). **Conclusão:** as ferramentas analisadas possuem

potencial em auxiliar o profissional de saúde na previsão de IM em adultos internados em Unidade de Terapia Intensiva (UTI), porém seu uso deve ser com cautela devido as informações equivocadas e imprecisas que podem ser geradas. Mais avanços são necessários para que possa ser utilizada de forma confiável.

Palavras-chave: Inteligência Artificial, ChatGPT, Deepseek, Gemini, Interações Medicamentosas, Unidades de Terapia Intensiva.

Introduction

Drug interactions (DIs) can be defined as the concomitant administration of two or more medications that may lead to a clinically relevant outcome related to efficacy, safety, or both.[1] They can result in adverse drug events and negative health consequences, especially in hospitalized patients due to polypharmacy and comorbidities.[2]

Critically ill patients are more likely to develop DIs due to frequent physiological changes, such as impaired absorption and reduced renal and hepatic function.[3] A systematic review estimated that 58% of these patients experienced at least one DI in the Intensive Care Unit (ICU), ranging from 1 to 5 interactions per patient,[4] although variations in study locations, patient characteristics, DI definitions, and methodological aspects may influence these findings.[3] Furthermore, a study identified that 65% of patients exposed to DIs developed preventable adverse events, with more than half classified as severe,[5] highlighting the significance of the issue.

In this context, the integration of technology into the detection and prediction of DIs has advanced significantly in clinical practice with the use of Artificial Intelligence (AI). Software applications,[6,7] Clinical Decision Support System (CDSS) alerts,[8] and machine learning,[9,10] are increasingly employed for DI identification and management. Among these technologies, the use of natural language models (NLMs), which utilize algorithms to comprehend and generate human-like conversations, has been growing and becoming more popular in healthcare-related contexts.[11]

Examples of AI-based NLMs include ChatGPT, Gemini, and DeepSeek. Initially launched in November 2022 by OpenAI, ChatGPT learns the nuances of natural language from large datasets, enabling the generation of coherent, human-like text.[12] Gemini (the successor to Google Bard) was announced by Google in December 2023 and is designed to understand and generate content in multiple formats (text, images, audio, and video), as well as to interpret and produce code in various programming languages.[13] Recently, in January 2025, the Chinese company DeepSeek released the latest version of its technology, introducing significant improvements in context comprehension, text generation, and efficiency.[14]

Previous studies have evaluated the performance of AI tools in accurately identifying DIs, mainly involving ChatGPT.[15-18] Juhi et al.[15] analyzed 40 drug pairs extracted from scientific literature, concluding that ChatGPT is a partially effective tool for predicting and explaining DIs. Similarly, Al-Ashwal et al.[16] evaluated drug pairs from a dataset of the 51 most prescribed medications, highlighting the potential of ChatGPT, Google Bard, and Bing AI tools to significantly enhance patient care. On the other hand, Aksoyalp et al.[17] reported that ChatGPT should not be used as a reference in clinical practice, while Krishnan et al.[18] suggested that ChatGPT has low performance in identifying DIs due to its low sensitivity.

To our knowledge, no study has evaluated the use of AI tools for detecting DIs related to critically ill patients or tested DeepSeek and Gemini for this purpose. Therefore, this study aims to analyze the ability of ChatGPT, DeepSeek, and Gemini to predict the main potential DIs found in critically ill patients.

Methods

Study Design and Period

This is a cross-sectional analytical study, with the internet as the primary data source, conducted in February 2025.

Selection of Databases

Two well-established DI databases, one free and one subscription-based, were used to assess the accuracy and completeness of the information generated by the AI tools ChatGPT v.3.5 (trained with data up to June 2024), DeepSeek v-3 (trained with data up to July 2024), and Gemini 2.0 Flash (trained with data up to September 2024). Micromedex, a subscription-based DI detection tool, was selected for its accessibility and reliability. Drugs.com, a free database, was chosen because it demonstrated higher accuracy in detecting DIs among other free resources.[19] ChatGPT v.3.5, DeepSeek v-3, and Gemini 2.0 Flash were selected for their free access and user-friendly interfaces.

List of Drug Interactions

A list of the most relevant DIs detected in adult ICU patients was compiled based on a previously published systematic review by Fitzmaurice et al.[4] The study's top 20 most frequent DIs were selected, as shown in Table 1. Subsequently, two drugs (a pair) were selected to verify each DI in the respective programs.

Table 1. List of potential drug interactions selected for the study, classified by severity according to the Micromedex and Drugs.com databases.

Main Drug Interactions	Severity	
	Micromedex	Drugs.com
Calcium and ceftriaxone	Contraindicated	No interactions
Fluconazole and omeprazole	Major	Moderate
Amphotericin B and prednisolone	No interactions	Moderate
Tacrolimus and prednisolone	No interactions	No interactions
Ondansetron and amiodarone	Major	Major
Fentanyl and midazolam	Major	Moderate
Amphotericin B and furosemide	No interactions	Moderate
Aspirin and enoxaparin	Major	Major
Ranitidine and morphine	No interactions	Moderate
Diltiazem and methylprednisolone	Moderate	Moderate
Aspirin and clopidogrel	Major	Moderate
Fentanyl and tramadol	Major	Major
Midazolam and tramadol	Major	Moderate
Insulin and aspirin	Moderate	Moderate
Phenytoin and omeprazole	Major	Moderate
Metoprolol and insulin	Moderate	Moderate
Norepinephrine and propofol	No interactions	No interactions
Sulfamethoxazole + trimethoprim and voriconazole	No interactions	Minor
Pantoprazole and mycophenolate mofetil	Major	Moderate
Amiodarone and simvastatin	Major	Major

*Brazilian Society of Hospital Pharmacy and Health Services

JHPHS Journal of Hospital Pharmacy and Health Services

Development of Input Command (Prompt)

The prompt was developed in Brazilian Portuguese using the PACIF acronym (Role, Action, Context, Intent, and Format) to ensure that the AI tools understand the scenario and deliver results aligned with the user's objectives. First, the desired role of the AI is defined. Next, the expected action is directed by providing the necessary context to ensure a proper understanding of the question. Then, the intent of the AI response is stated to meet specific needs. Finally, the format of the desired response is defined.

Thus, the following prompt was established: "*Act as a healthcare professional and identify whether there is an interaction between drugs A and B in the context of an Intensive Care Unit. Describe the severity of the interaction (contraindicated, major, moderate, minor, or no interaction), its documentation quality (excellent, good, fair, or limited), and the recommended course of action. Present the results in a list format.*"

Data Collection and Analysis

A user with a free account on ChatGPT, DeepSeek, and Gemini interacted with the AI language models. For the Micromedex database, the mobile application available on the Android device was accessed using a username and password provided to professionals registered with the Federal Pharmacy Council of Brazil. In contrast, data from the Drugs.com tool were collected using an incognito browser tab on a computer to ensure unbiased and secure access. Data collection was conducted on February 5, 2025. The responses from the databases and AI tools were reviewed by a pharmacist and recorded in a preformatted Microsoft Excel spreadsheet.

Based on the information provided by the tools and adapted from the study by Cedraz & Santos Junior,[20] the severity levels of DIs were standardized into five categories:

- Contraindicated: The medications should not be used concomitantly.
- Major: May present a life-threatening risk and/or require intervention to minimize or prevent severe adverse effects.
- Moderate: May exacerbate the patient's condition and/or require a therapy modification.
- Minor: Clinical effects are limited. Manifestations may include an increase in the frequency or severity of adverse effects but generally do not require a major therapy adjustment.
- No significant interaction: No significant interaction was found.

Additionally, accuracy was determined using sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), calculated as follows[16]:

- Sensitivity = TP / (TP + FN)
- Specificity = TN / (FP + TN)
- PPV = TP / (TP + FP)
- NPV = TN / (TN + FN)
- Accuracy = (TP + TN) / (TP + FN + TN + FP)

Where: TP (True Positive) refers to a DI classified as relevant by both Micromedex or Drugs.com and the tested AI tool (classified as moderate, major, or contraindicated by both sources). TN (True Negative) refers to a DI considered irrelevant, as it was either undetected or identified only as minor by both Micromedex or Drugs.com and the tested AI tool.[16]

The AI tools' ability to accurately detect contraindicated, major, or moderate DIs according to the databases was defined as sensitivity, while their ability to disregard minor or non-significant interactions was defined as specificity. PPV represents the likelihood that an interaction detected by the AI tools is clinically significant. Additionally, NPV indicates the probability that interactions not detected by the AI tools are insignificant.[16]

Fleiss' kappa coefficients (κ) were calculated to measure the agreement between Micromedex, Drugs.com, and the AI tools regarding DI severity.[17] Moreover, the AI tools' responses on the mechanism of action and recommended management of DIs were categorized as "adequate" (when the information was consistent) or "inadequate" (when the information was inconsistent) compared to the tested databases. "Adequate" responses were further classified as "accurate" (clear and aligned with the tested databases) or "inaccurate" (unclear and/or including non-relevant additional information).

Descriptive statistics, including simple frequencies and percentages of responses generated by the databases and AI tools, were analyzed using Microsoft Excel. Accuracy levels were classified as follows: low: 0–30%; moderate: >30% and <90%; high: >90%. Fleiss' κ was calculated using SPSS v.25, with agreement levels categorized as very weak: 0–0.2; weak: 0.21–0.40; moderate: 0.41–0.60; good: 0.61–0.80; very good: >0.8021. A p-value < 0.05 for κ indicated that the agreement between the databases and AI tools was unlikely to have occurred by chance.

Results

The number of DIs analyzed and their respective severity levels are presented in Figure 1. The Micromedex database identified three DIs as moderate in severity, ten as major, and one as contraindicated, while six tested DIs were not assigned a severity level (Amphotericin B and Prednisolone, Tacrolimus and Prednisolone, Amphotericin B and Furosemide, Ranitidine and Morphine, Norepinephrine and Propofol, and Sulfamethoxazole + Trimethoprim and Voriconazole). On the other hand, the Drugs.com database detected one DI classified as minor, 12 as moderate, and four as major, while the severity of three tested DIs was not classified (Calcium and Ceftriaxone, Tacrolimus and Prednisolone, and Norepinephrine and Propofol). Regarding AI tools, all identified a higher number of DIs classified as major: 13 by ChatGPT and 11 by both DeepSeek and Gemini. Additionally, moderate-severity DIs were detected (ChatGPT = 3, DeepSeek = 6, Gemini = 7), minor-severity DIs (ChatGPT = 3, DeepSeek = 2, Gemini = 1), and contraindicated DIs (n = 1 for all). Notably, none of the AI tools identified DIs as having no severity.

The results for sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, and agreement of all analyzed tools concerning DI severity are described in Table 2. The highest percentage of correct responses (accuracy) compared to Micromedex was obtained by ChatGPT (15 out of 20, 75%), while the lowest was by Gemini (13 out of 20, 65%). Compared to Drugs.com, the highest accuracy was observed in DeepSeek (0.800), followed by ChatGPT and Gemini, both with 0.750.

Figure 1. Potential drug interactions categorized by severity.

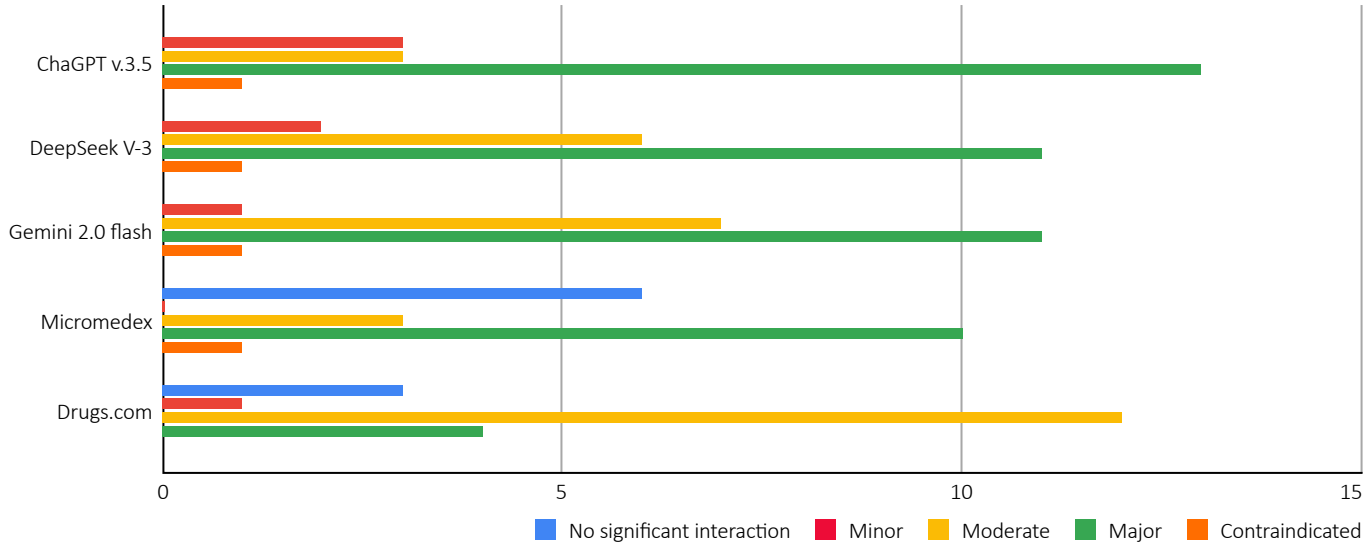

Table 2. Sensitivity, specificity, accuracy, and agreement values between the programs regarding the severity of the 20 potential drug interactions analyzed.

Artificial Intelligence Tools	VP	FN	VN	FP	Measures		VPP	VPN	% Accuracy	CKF (p-value)
					Sensitivity	Specificity				
Micromedex as a reference										
ChatGPT v.3.5	13	1	2	4	0,929	0,333	0,765	0,667	75	
DeepSeek v-3	13	1	1	5	0,929	0,167	0,722	0,500	70	0,410 (0,000)
Gemini 2.0 flash	13	1	0	6	0,929	0,000	0,684	0,000	65	
Drugs.com as a reference										
ChatGPT v.3.5	14	2	1	3	0,875	0,250	0,824	0,333	75	
DeepSeek v-3	16	1	0	3	0,941	0,000	0,842	0,000	80	0,354 (0,000)
Gemini 2.0 flash	15	1	0	4	0,938	0,000	0,789	0,000	75	

Legend: CKF (Fleiss' Kappa coefficient), FN (false negative), FP (false positive), VN (true negative), VP (true positive), VPP (positive predictive value), VPN (negative predictive value).

Sensitivity was high across all AI tools compared to Micromedex (0.925 for all). Regarding Drugs.com, the values were 0.875 for ChatGPT, 0.941 for DeepSeek, and 0.938 for Gemini. Conversely, when compared to Micromedex, specificity values were low for ChatGPT (0.333), DeepSeek (0.167), and Gemini (0.000). Regarding Drugs.com, the values were 0.250 for ChatGPT and 0.000 for both DeepSeek and Gemini.

The PPV of ChatGPT, DeepSeek, and Gemini, when compared to Micromedex, were 0.765, 0.722, and 0.684, respectively. These values increased when compared to Drugs.com, reaching 0.842 for DeepSeek, 0.824 for ChatGPT, and 0.789 for Gemini, highlighting their ability to identify more severe interactions. Conversely, only ChatGPT and DeepSeek presented NPV values when compared to Micromedex (0.667 and 0.500, respectively), and only ChatGPT showed NPV when compared to Drugs.com (0.333). These results indicate a limited ability to identify interactions of no or minor severity.

Agreement between AI tools and Micromedex was classified as moderate ($\kappa = 0.410$; $p < 0.000$), while the agreement with Drugs.com was considered weak ($\kappa = 0.354$; $p < 0.000$).

Data on all evaluated DIs and their respective generated responses are presented in Appendix 1. A total of 14 and 17 DIs were analyzed regarding the responses generated by AI tools on the mechanism of action, compared to Micromedex and Drugs.com, respectively. Only one response from ChatGPT was considered inadequate when compared to Micromedex. On the other hand, in comparison with Drugs.com, 10 responses were deemed adequate but imprecise, while three were considered inadequate—four imprecise responses were associated with ChatGPT, and two inadequate responses with Gemini. Regarding responses generated on clinical management, compared to Micromedex, seven responses were considered adequate but imprecise, and one was deemed inadequate, with four of these imprecise responses associated with ChatGPT. When comparing the same responses with Drugs.com, 11 were classified as adequate but imprecise, and four were considered inadequate, with six inadequate responses linked to Gemini. The number and classification of responses generated by AI tools compared to Micromedex and Drugs.com are shown in Table 3. Table 4 provides examples of inadequate responses on the mechanism of action and recommended management generated by AI tools compared to Micromedex and Drugs.com.

JHPHS Journal of Hospital Pharmacy and Health Services

Table 3. Responses generated by Artificial Intelligence tools compared to Micromedex and Drugs.com.

Artificial Intelligence Tools	Response Classification	Mechanism of Action (%)	Conduct (%)
Micromedex as a reference (n = 14)			
ChatGPT v.3.5	Inadequate	1 (7,2)	1 (7,2)
	Adequate and precise	13 (92,8)	10 (71,4)
	Adequate and imprecise	-	3 (21,4)
DeepSeek v-3	Inadequate	-	-
	Adequate and precise	14 (100)	12 (85,7)
	Adequate and imprecise	-	2 (14,3)
Gemini 2.0 flash	Inadequate	-	-
	Adequate and precise	14 (100)	12 (85,7)
	Adequate and imprecise	-	2 (14,3)
Drugs.com as a reference (n = 17)			
ChatGPT v.3.5	Inadequate	1 (5,9)	1 (5,9)
	Adequate and precise	12 (70,6)	13 (76,5)
	Adequate and imprecise	4 (23,5)	3 (17,6)
DeepSeek v-3	Inadequate	1 (5,9)	1 (5,9)
	Adequate and precise	13 (76,5)	12 (70,6)
	Adequate and imprecise	3 (17,6)	4 (23,5)
Gemini 2.0 flash	Inadequate	2 (11,8)	2 (11,8)
	Adequate and precise	12 (70,6)	11 (64,7)
	Adequate and imprecise	3 (17,6)	4 (23,5)

Table 4. Examples of inadequate responses regarding the mechanism of action and recommended conduct generated by AI tools compared to Micromedex and Drugs.com.

Potential drug interactions (PDI)	Responses generated by Artificial Intelligence tools				
	ChatGPT v3.5	DeepSeek v-3	Gemini 2.0 flash	Micromedex	Drugs.com
MX-TMP and voriconazole	Mechanism: It may result in an increased risk of hepatic toxicity (SMX-TMP inhibits the enzymes CYP2C19, and CYP3A4, where voriconazole is metabolized). Conduct: Monitor the plasma levels of voriconazole (adjust if necessary) and monitor liver function.	Mechanism: It potentiates the adverse effects of voriconazole, such as hepatotoxicity (SMX-TMP inhibits the enzymes CYP2C19, CYP3A4, and CYP2C9). Conduct: Monitor the plasma concentrations of voriconazole, if available, to adjust the dose and avoid toxicity.	Mechanism: It may increase the plasma levels of voriconazole due to the inhibition of the enzymes CYP2C9 and CYP2C8 by SMX-TMP, which could lead to hepatotoxicity. Conduct: Contraindicated.	No interaction.	Mechanism: Limited data suggests it may rarely prolong the QT interval. Conduct: Seek immediate medical attention if symptoms such as dizziness, fainting, palpitations, irregular heartbeat, shortness of breath, or syncope occur, as they may indicate torsades de pointes.

Legend: IM (drug interactions), SMX-TMP (Sulfamethoxazole + trimethoprim).

Discussion

This is the first study to analyze the ability of three AI tools (ChatGPT, DeepSeek, and Gemini) to predict the 20 most common DIs found in adult patients admitted to the ICU. The evaluated tools demonstrated satisfactory levels of accuracy and sensitivity compared to Micromedex and Drugs.com, with ChatGPT demonstrating relatively superior performance. Despite the AI tools showing precision in detecting relevant DIs, concerns regarding the content of their generated responses suggest that these tools are not yet fully prepared for unrestricted use in clinical practice.

It is undeniable that generative AI has significantly impacted people's daily lives and various fields of knowledge, including healthcare. In this field, these technologies play an increasingly relevant role in optimizing diagnoses[22], personalizing therapies, and supporting clinical[23] decision-making[24], promoting greater efficiency and precision in patient care. Additionally, the tools analyzed in this study are free to use, facilitating their adoption by healthcare professionals. However, it is important to highlight that studies indicate limited acceptance of AI integration into clinical practice among healthcare professionals.[25,26] Therefore, efforts aimed at AI literacy and training, interdisciplinary collaboration with technology specialists, and improving user experience in the healthcare context are essential to ensure efficient and sustainable patient-centered care.[25,27]

JHPHS
Journal of Hospital Pharmacy and Health Services

The AI tools demonstrated the ability to detect a higher number of potential DIs compared to the reference databases. These databases provide specific information about drugs and diseases, whereas AI tools use broader and more diverse sources of information. However, a higher detection rate does not necessarily translate into superior performance, as these tools exhibited higher false positive rates, incorrectly identifying interactions where none actually exist.[16] A scoping review on the use of ChatGPT in pharmaceutical practice highlighted that limitations, particularly the provision of imprecise information, represent a significant challenge to its incorporation into professional routines.[28]

Sensitivity rates ranging from 0.875 to 0.941 and specificity rates from 0.0 to 0.333 indicate that approximately 87% to 94% of severe DIs and 0% to 33% of non-severe DIs were correctly identified by the AI tools. These findings suggest that AI has the potential to help prevent adverse events associated with DIs and contribute to reducing hospital stays.[29,30] The findings of this study align with a previous study in which ChatGPT v.3.5 demonstrated a sensitivity of 89.4% and specificity of 37.2% for DI detection using the same reference databases.[16] Another study also reported similar results, where ChatGPT v.3.5 showed a sensitivity of 90.5% and specificity of 50% when using the Lexicomp Drug Interactions database as a reference.[17] Conversely, Krishnan et al. found sensitivity and specificity values of 24.3% and 92.9%, respectively, when using real-world DIs and comparing results to the prior experience of clinical pharmacists, which may explain the discrepancies among findings.

This study identified low to moderate agreement between the severity classifications provided by the AI tools compared to Drugs.com and Micromedex, respectively. Aksoyalp et al.[17] also observed low agreement in their study. Additionally, discrepancies were noted in some information regarding the mechanism of action and recommended management, both between the AI tools and the reference databases as well as among the databases themselves. Previous studies have shown that agreement among different DI detection databases is variable, raising concerns about their reliability in clinical practice.[31-33] Despite Micromedex[34,35] and Drugs.com[36,37] being widely used in clinical settings, it is crucial to consider the specific reference adopted when interpreting results, as performance may vary depending on the chosen source. Complementing information with multiple sources[38,39] and considering real-world scenarios is essential to ensure the most appropriate decision-making.

The use of AI tools in clinical practice for obtaining drug-related information is becoming increasingly prominent. This study, through a structured and robust methodology, highlighted the potential and challenges associated with using these tools. Furthermore, it is the first study to evaluate the use of DeepSeek and Gemini in this context, providing an innovative perspective on their utility. However, this study has some limitations. The analyzed interactions were hypothetical, involving two drugs independently, which does not reflect the reality of adult ICU patients, where multiple medications are administered simultaneously. Additionally, the study did not account for the clinical significance of DIs, which may limit its ability to predict real-world outcomes. Lastly, AI tools are updated regularly, meaning that the results obtained in the present study may change in future evaluations.

Conclusion

The AI tools analyzed in this study demonstrated high sensitivity and a moderate level of accuracy, suggesting their potential to assist healthcare professionals in predicting DIs in adult patients admitted to ICUs. However, the agreement between the severity classifications generated by the AI tools and the reference databases was categorized as weak to moderate. Additionally, inadequate and imprecise responses were observed regarding both the mechanisms of action and the recommended management of DIs, highlighting the need for cautious use of these tools.

Further advancements and research involving these and other AI tools are necessary to achieve adequate levels of safety and quality, enabling their integration into healthcare practice, particularly in the context of DI prediction.

Funding Sources

The author declares that this research received no funding.

Conflict of Interest Statement

The author declares no conflicts of interest regarding this article.

Statement on the Use of Generative AI in Writing

During the preparation of this manuscript, the author used ChatGPT and DeepSeek to enhance the clarity and linguistic quality of the text. After applying these technologies, the author carefully reviewed and edited the content as needed, assuming full responsibility for the final published article.

JHPHS
Journal of Hospital Pharmacy and Health Services

References

1. Dimakos J, Douros A. Methodological Considerations on the Use of Cohort Designs in Drug-Drug Interaction Studies in Pharmacoepidemiology. Curr Epidemiol Rep 2024;11(3): 175–183. doi: 10.1007/s40471-024-00347-1.
2. Aksoy N, Ozturk N. A meta-analysis assessing the prevalence of drug-drug interactions among hospitalized patients. Pharmacoepidemiol Drug Saf. 2023;32(12):1319-1330. doi: 10.1002/pds.5691.
3. Bakker T, Dongelmans DA, Nabovati E, *et al*. Heterogeneity in the Identification of Potential Drug-Drug Interactions in the Intensive Care Unit: A Systematic Review, Critical Appraisal, and Reporting Recommendations. J Clin Pharmacol. 2022;62(6):706-720. doi: 10.1002/jcph.2020.
4. Fitzmaurice MG, Wong A, Akerberg H, *et al*. Evaluation of Potential Drug-Drug Interactions in Adults in the Intensive Care Unit: A Systematic Review and Meta-Analysis. Drug Saf. 2019;42(9):1035-1044. doi: 10.1007/s40264-019-00829-y.
5. Klopotowska JE, Leopold JH, Bakker T, *et al*. Adverse drug events caused by three high-risk drug-drug interactions in patients admitted to intensive care units: A multicentre retrospective observational study. Br J Clin Pharmacol. 2024;90(1):164-175. doi: 10.1111/bcp.15882.
6. Roblek T, Vaupotic T, Mrhar A, *et al*. Drug-drug interaction software in clinical practice: a systematic review. Eur J Clin Pharmacol. 2015;71(2):131-42. doi: 10.1007/s00228-014-1786-7.
7. Apidi NA, Murugiah MK, Muthuveloo R, *et al*. Mobile Medical Applications for Dosage Recommendation, Drug Adverse Reaction, and Drug Interaction: Review and Comparison. Ther Innov Regul Sci. 2017;51(4):480-485. doi: 10.1177/2168479017696266.
8. Felisberto M, Lima GDS, Celuppi IC, *et al*. Override rate of drug-drug interaction alerts in clinical decision support systems: A brief systematic review and meta-analysis. Health Informatics J. 2024;30(2):14604582241263242. doi: 10.1177/14604582241263242.
9. Zhang T, Leng J, Liu Y. Deep learning for drug-drug interaction extraction from the literature: a review. Brief Bioinform. 2020;21(5):1609-1627. doi: 10.1093/bib/bbz087.
10. Dou M, Tang J, Tiwari P, *et al*. Drug–Drug Interaction Relation Extraction Based on Deep Learning: A Review. ACM Computing Surveys. 2024;56(6):1-33. doi: 10.1145/3645089.
11. Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. NPJ Digit Med. 2021 Jun 3;4(1):93. doi: 10.1038/s41746-021-00464-x.
12. OpenAI. ChatGPT-3.5. Available online: https://chatgpt.com/. Available in: 5 feb 2025.
13. Google. Gemini 2.0 flash. Available online: https://gemini.google.com/app?hl=pt-BR. Available in: 5 feb 2025.
14. DeepSeek. DeepSeek V-3. Available online: https://www.deepseek.com/. Available in: 5 feb 2025.
15. Juhi A, Pipil N, Santra S, *et al*. The Capability of ChatGPT in Predicting and Explaining Common Drug-Drug Interactions. Cureus. 2023;15(3): e36272. doi: 10.7759/cureus.36272.
16. Al-Ashwal FY, Zawiah M, Gharaibeh L, *et al*. Evaluating the Sensitivity, Specificity, and Accuracy of ChatGPT-3.5, ChatGPT-4, Bing AI, and Bard Against Conventional Drug-Drug Interactions Clinical Tools. Drug Healthc Patient Saf. 2023;15:137-147. doi: 10.2147/DHPS.S425858.
17. Aksoyalp ZS, Erdoğan BR. Comparative evaluation of artificial intelligence and drug interaction tools: a perspective with the example of clopidogrel. J Fac Pharm Ankara. 2024;48(3):1011-1020. doi: 10.33483/jfpau.1460173.
18. Krishnan RPR, 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. doi: 10.1111/bcp.16275.
19. Bossaer JB, Eskens D, Gardner A. Sensitivity and specificity of drug interaction databases to detect interactions with recently approved oral antineoplastics. J Oncol Pharm Pract. 2022;28(1):82–86. doi:10.1177/1078155220984244.
20. Cedraz KN, Santos Junior MC. Identificação e caracterização de interações medicamentosas em prescrições médicas da unidade de terapia intensiva de um hospital público da cidade de Feira de Santana, BA. Rev Soc Bras Clin Med. 2014;12(2):1-7.
21. Altman D.G. Practical Statistics for Medical Research. London: Chapman and Hall; 1991.
22. Kumar Y, Koul A, Singla R, *et al*. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. J Ambient Intell Humaniz Comput. 2023;14(7):8459-8486. doi: 10.1007/s12652-021-03612-z.
23. Rezayi S, R Niakan Kalhori S, Saeedi S. Effectiveness of Artificial Intelligence for Personalized Medicine in Neoplasms: A Systematic Review. Biomed Res Int. 2022;2022:7842566. doi: 10.1155/2022/7842566.
24. Elhaddad M, Hamam S. AI-Driven Clinical Decision Support Systems: An Ongoing Pursuit of Potential. Cureus. 2024 Apr 6;16(4):e57728. doi: 10.7759/cureus.57728.
25. Hoffman J, Hattingh L, Shinners L, *et al*. Allied Health Professionals' Perceptions of Artificial Intelligence in the Clinical Setting: Cross-Sectional Survey. JMIR Form Res. 2024 Dec 30;8:e57204. doi: 10.2196/57204.
26. Sadiq F, Sadiq F, Gul R, *et al*. Knowledge, Attitude, and Practice (KAP) Regarding the Use of Artificial Intelligence in Hospital Settings in Mardan, Khyber Pakhtunkhwa, Pakistan. Cureus. 2024 Dec 9;16(12):e75355. doi: 10.7759/cureus.75355.
27. Jena A, Chaudhary N, Manohar B, *et al*. Knowledge and Perception of Artificial Intelligence Amidst the Health Professionals-A Web-Based Survey. J Pharm Bioallied Sci. 2024 Dec;16(Suppl 5):S4365-S4367. doi: 10.4103/jpbs.jpbs_647_24.
28. Lima TM, Bonafé M, Baby AR, *et al*. ChatGPT in Pharmacy Practice: Disruptive or Destructive Innovation? A Scoping Review. Sci. Pharm. 2024;92(4):58. doi:10.3390/scipharm92040058.

29. Lima EDC, Camarinha BD, Ferreira Bezerra NC, *et al*. Severe Potential Drug-Drug Interactions and the Increased Length of Stay of Children in Intensive Care Unit. Front Pharmacol. 2020;11:555407. doi: 10.3389/fphar.2020.555407.

30. Schmitt JP, Kirfel A, Schmitz MT, *et al*. The Impact of Drug Interactions in Patients with Community-Acquired Pneumonia on Hospital Length of Stay. Geriatrics (Basel). 2022;7(1):11. doi: 10.3390/geriatrics7010011.

31. Shariff A, Belagodu Sridhar S, Abdullah Basha NF, *et al*. Assessing Consistency of Drug-Drug Interaction-Related Information Across Various Drug Information Resources. Cureus. 2021;13(3):e13766. doi: 10.7759/cureus.13766.

32. Kontsioti E, Maskell S, Bensalem A, *et al*. Similarity and consistency assessment of three major online drug-drug interaction resources. Br J Clin Pharmacol. 2022;88(9):4067-4079. doi: 10.1111/bcp.15341.

33. Carollo M, Crisafulli S, Selleri M, *et al*. Agreement of Different Drug-Drug Interaction Checkers for Proton Pump Inhibitors. JAMA Netw Open. 2024;7(7):e2419851. doi: 10.1001/jamanetworkopen.2024.19851.

34. Yamagata AT, Barcelos Júnior RMC, Galato D, *et al*. Perfil dos estudos de interações medicamentosas potenciais em hospitais brasileiros: revisão integrativa. Rev Bras Farm Hosp Serv Saude. 2018;9(4):e094.003. doi: 10.30968/rbfhss.2018.094.003.

35. Rothgeb A, Beckett RD, Daoud N. Off-label use information in electronic drug information resources. J Med Libr Assoc. 2022;110(4):471-477. doi: 10.5195/jmla.2022.1419.

36. Marcath LA, Xi J, Hoylman EK, Kidwell KM, *et al*. Comparison of Nine Tools for Screening Drug-Drug Interactions of Oral Oncolytics. J Oncol Pract. 2018 Jun;14(6):e368-e374. doi: 10.1200/JOP.18.00086.

37. Pinkoh R, Rodsiri R, Wainipitapong S. Retrospective cohort observation on psychotropic drug-drug interaction and identification utility from 3 databases: Drugs.com®, Lexicomp®, and Epocrates®. PLoS One. 2023;18(6):e0287575. doi: 10.1371/journal.pone.0287575.

38. 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(11):e0225239. doi: 10.1371/journal.pone.0225239.

39. Hecker M, Frahm N, Bachmann P, et al. Screening for severe drug-drug interactions in patients with multiple sclerosis: A comparison of three drug interaction databases. Front Pharmacol. 2022;13:946351. doi: 10.3389/fphar.2022.946351.