Generative AI Platforms in Predicting Drug Interactions: Case of Digoxin and Warfarin

Cahya Permana Apriansah[1], Erna Kristin[2], Soni Siswanto[3], Purwantiningsih Purwantiningsih[3], and Agung Endro Nugroho[3*]

[1] Master of Clinical Pharmacy Program, Faculty of Pharmacy, Universitas Gadjah Mada, Yogyakarta, Indonesia
[2] Department of Pharmacology & Therapy, Faculty of Medicine, Public Health and Nursing, Universitas Gadjah Mada, Yogyakarta, Indonesia
[3] Department of Pharmacology and Clinical Pharmacy, Faculty of Pharmacy, Universitas Gadjah Mada, Yogyakarta, Indonesia
nugroho_ae@ugm.ac.id

Abstract. Drug interactions are a major concern in clinical therapy due to their potential to affect both efficacy and safety. Medications with a narrow therapeutic index, such as digoxin and warfarin, are especially susceptible. With rapid advancements in technology, generative AI platforms like ChatGPT-4o mini, Copilot, and Gemini have emerged as potential tools for identifying drug interactions. However, their validity compared to established resources like UpToDate Lexidrug remains uncertain. This study aimed to evaluate the performance of the three generative AI platforms in detecting drug interactions, using UpToDate Lexidrug as a reference. An observational, cross-sectional design with comparative analysis was applied, and data were analyzed using descriptive statistics. Results showed that ChatGPT-4o mini achieved sensitivity (0.704), specificity (0.714), positive predictive value or PPV (0.812), negative predictive value or NPV (0.579), and accuracy (0.708). Copilot recorded sensitivity (0.859), specificity (0.312), PPV (0.686), NPV (0.558), and accuracy (0.660). Gemini achieved sensitivity (0.889), specificity (0.429), PPV (0.732), NPV (0.688), and accuracy (0.722). Among the platforms, Gemini obtained the highest sensitivity, NPV, and accuracy, while ChatGPT-4o mini demonstrated the best specificity and PPV. These results show that AI has the potential to support drug interaction screening. However, some outputs were inaccurate or incomplete, so generative AI should not replace healthcare professionals in making clinical decisions.

Keywords: Drug Interactions, Validity, Artificial Intelligence, Digoxin, Warfarin.

1 Introduction

Drug interaction is a change in the effect or properties of a drug caused by the concurrent administration of one or more drugs, food, or beverages. Drug interactions are known to be a significant contributing factor to adverse drug reactions and can affect

V. D. S. Handayani et al. (eds.), *Proceedings of the 6th International Conference on Bioinformatics, Biotechnology, and Biomedical Engineering (BIOMIC 2025)*, Advances in Health Sciences Research 94,
https://doi.org/10.2991/978-94-6239-762-0_29

drug response [1]. Drug-drug interactions can be defined as a changed effect of a drug (its effect can be increased or decreased) due to the presence of another drug when administered concurrently [2], [3], [4]. Drug interactions are common occurrences, but they can be hazardous to patient safety [3], [5].

Drugs with a narrow therapeutic index are highly susceptible to drug interactions, particularly those related to pharmacokinetics [6]. Many clinically relevant drug interactions involve drugs with a narrow therapeutic index [7]. Examples of drugs with a narrow therapeutic index include digoxin and warfarin [8].

Digoxin interacts with several drugs, including macrolide antibiotics, which can disrupt normal gut microbiota, leading to higher absorption concentrations; metoclopramide, which can decrease digoxin absorption; and indomethacin and spironolactone, which can reduce digoxin clearance [9].

Warfarin has numerous drug interactions that can increase the risk of adverse effects or decrease its anticoagulant effect. Concomitant use of antiplatelets, fibrinolytic, antiarrhythmics, antimicrobials, nonsteroidal anti-inflammatory drugs, and other anticoagulants requires careful consideration when administered with warfarin [10], [11]. Several antimicrobials directly interfere with warfarin metabolism through the inhibition of CYP enzymes, including metronidazole, trimethoprim-sulfamethoxazole, and ciprofloxacin [12], [13], [14].

Although drug interactions cannot always be avoided, they are often predictable [5], [15]. It is essential to identify and address issues related to drug interactions accurately. One method to ensure effective management of these interactions is through the use of widely accessible online drug information databases, which enhance usability and accessibility for pharmacists [16]. UpToDate Lexidrug, Micromedex, Medscape, Drugs.com, and DrugBank are examples of conventional drug databases that can be accessed online via websites or applications. These databases differ in availability; some, such as Drugs.com and DrugBank, are freely available, while others, like UpToDate Lexidrug and Micromedex, necessitate paid subscriptions.

In the contemporary era, the landscape of technology and information is undergoing rapid advancement, exemplified by the emergence of Artificial Intelligence (AI), which can potentially influence various aspects of life, including healthcare and pharmacy. Freely accessible generative AI platforms, such as ChatGPT-4o mini, Copilot, and Gemini, are equipped with sophisticated algorithms and possess machine-learning capabilities that hold promise in offering solutions to diverse challenges within healthcare services [17]. For instance, within the domain of drug interactions, the use of AI platforms in identifying potential adverse drug reactions and drug interactions can enhance medication management and patient safety [17], [18]. Although generative AI platforms such as ChatGPT-4o mini, Copilot, and Gemini are not explicitly trained for medical purposes, they possess significant capabilities in the rapid screening of medical literature, scientific research, and scholarly papers [17], [19].

While AI platforms offer numerous advantages, they also have limitations and challenges in their application. Notably, these platforms lack access to proprietary databases or subscription-based resources [17], [20]. The use of AI platforms for medical purposes is an area that requires further in-depth exploration, and a comprehensive regulatory framework governing the utilization of AI in healthcare is

still largely underdeveloped [17], [21]. Moreover, the accuracy, sensitivity, and specificity of each generative AI platform in identifying drug interactions remain unclear, and ambiguity persists regarding their ability to detect or predict potential drug interactions reliably [17].

Thus, this study aims to evaluate the differences in accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of each generative AI platform in detecting or predicting digoxin and warfarin interactions using cross-sectional study approaches. This research is expected to pave the way for the future integration of generative AI platforms usage into routine clinical practice.

2 Methodology

2.1 Study Design and Setting

This study employed an observational, cross-sectional design utilizing a comparative analysis research method to evaluate the performance of AI in detecting drug-drug interactions. This study was conducted online by accessing conventional drug interaction databases, namely UpToDate Lexidrug, and generative AI platforms including ChatGPT-4o mini, Copilot, and Gemini

2.2 Research Subject

The subjects in this study were generative AI platforms, including ChatGPT-4o mini, Copilot, and Gemini, and the conventional drug database, UpToDate Lexidrug, which served as the comparator/reference. The drugs used in this study were digoxin and warfarin, along with other medications that have potential interactions with digoxin and warfarin based on data from the FDA (Coumadin and Lanoxin Labeling-Package Insert), a systematic review and meta-analysis article [22], and the Indonesian National Formulary. Tables 1 and 2 present a detailed list of drugs used in this study.

2.3 Study Instruments

The research instrument comprised prompts entered into the question field of each generative AI platform. The prompts or questions posed were identical to those used in previous research [17], namely:

- *Which of the following choices best describes the drug-drug interaction between digoxin and "……."?*
  A. No known interaction
  B. Minor
  C. Moderate
  D. Major
- *Which of the following choices best describes the drug-drug interaction between digoxin and "……."?*
  A. No known interaction
  B. Minor
  C. Moderate

D. Major

Table 1. List of Drugs Used (Digoxin Interactions with Other Drugs)

Cardiovascular Drugs			
Amiodarone	Atorvastatin	Captopril	Carvedilol
Diltiazem	Verapamil	Nicardipine	Nifedipine
Telmisartan	Spironolactone	Ticagrelor	Cholestyramine
Ivabradine			
Anti-infectives			
Clarithromycin	Itraconazole	Lopinavir + Ritonavir	Gentamicin
Erythromycin	Tetracycline	Quinine	Trimethoprim
Azithromycin	Ketoconazole	Rifampicin	
Antineoplastics and Immunomodulators			
Cyclosporin	Lapatinib		
Gastrointestinal Drugs			
Omeprazole	Esomeprazole	Lansoprazole	Antacids
Metoclopramide	Sucralfate	Kaolin + Pectin	Sulfasalazine
Psychopharmaca			
Sertraline	Alprazolam		
Analgesics, Antipyretics, Non-Steroidal Anti-Inflammatories, Antigouts			
Diclofenac Sodium	Ibuprofen		
Hormones, Other Endocrine Drugs and Contraceptives			
Metformin	Acarbose		
Respiratory Drugs			
Salbutamol			
Antiepileptics – Anticonvulsants			
Phenytoin			
Antiallergy and Drugs for Anaphylaxis			
Epinephrine	Norepinephrine	Dopamine	
Peripheral Muscle Relaxants and Cholinesterase Inhibitors			
Succinylcholine			
Vitamins and Minerals			
Calcium gluconate	Calcium carbonate	Calcium lactate	

2.4 Data Collection

The severity levels of drug interactions in the prompts were aligned with the severity categories found in UpToDate Lexidrug, which consist of no known interaction, minor, moderate, and major, to determine the accuracy, sensitivity, and specificity based on the obtained true positive, true negative, false positive, and false negative values.

Interactions classified as major or moderate in UpToDate Lexidrug were considered true positives (TP) when concurrently identified as major or moderate in generative AI

platforms, and false negatives (FN) when classified as minor or no known interaction in those AI platforms. Conversely, minor or no known interaction in UpToDate Lexidrug were designated as true negatives (TN) if they were also categorized as minor or no known interaction in generative AI platforms, and false positives (FP) if they were categorized as major or moderate in generative AI platforms [17], [23], [24]. Fig. 1. provides a summary of the data collection procedure.

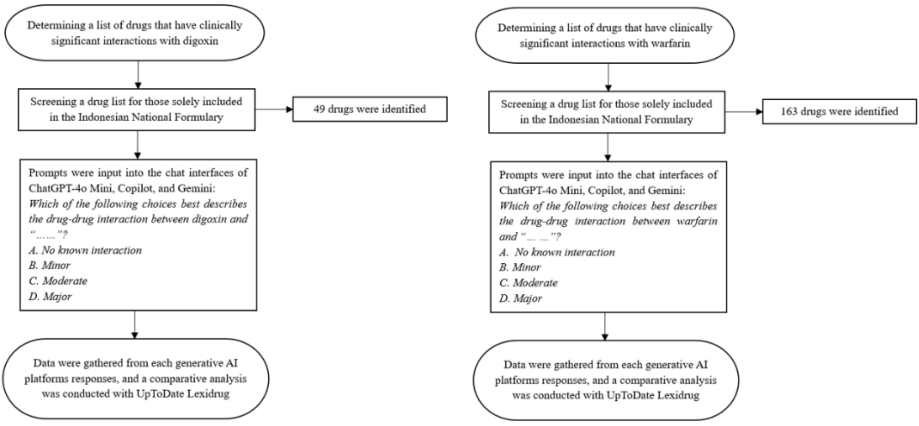

Fig. 1. Study flowchart

2.5 Data Analysis

Descriptive statistics were employed to analyze the data collected in this study. Accuracy, specificity, sensitivity, positive predictive value, and negative predictive value were determined for each generative AI platform, namely ChatGPT-4o mini, Copilot, and Gemini. These metrics were used to evaluate the capacity of each generative AI platform to identify clinically relevant interactions (sensitivity) correctly and to exclude clinically irrelevant interactions (specificity) correctly. Positive predictive value (PPV) reflects the probability that a database-detected interaction is genuinely significant. Conversely, negative predictive value (NPV) indicates the likelihood that interactions deemed absent by the database are indeed insignificant [17], [23], [25]. The formulas used for these measurements were as follows [17], [23], [24], [26] :

Accuracy = (TP + TN) / (TP + FP + TN + FN)

Sensitivity = TP / (TP + FN)

Specificity = TN / (FP + TN)

PPV = TP / (TP + FP)

NPV = TN / (TN + FN)

where TP = true positive; TN = true negative; FP = false positive; FN = false negative; PPV = positive predictive value; and NPV = negative predictive value. Data are presented as numerical values. An accuracy, sensitivity, specificity, PPV, and NPV value of the generative AI platforms approaching 1 indicates that the results obtained closely align with the drug interaction analysis conducted by UpToDate LexiDrug.

2.6 Ethical Considerations

Data were obtained and analyzed from generative AI platforms and drug databases for this study. As no human or animal subjects were involved, and language model responses were generated without the use of patient information, ethical committee approval was not required.

Table 2. List of Drugs Used (Warfarin Interactions with Other Drugs)

Cardiovascular Drugs			
Amiodarone	Propranolol	Verapamil	Amlodipine
Atorvastatin	Cilostazol	Diltiazem	Aspirin
Clopidogrel	Ticagrelor	Pravastatin	Rosuvastatin
Simvastatin	Fenofibrate	Gemfibrozil	Cholestyramine
Isosorbide Dinitrate	Atenolol	Bisoprolol	Metoprolol
Carvedilol	Adenosine	Digoxin	Lidocaine
Doxazosin	Hydrochlorothiazide	Imidapril	Irbesartan
Candesartan	Captopril	Clonidine	Lisinopril
Methyldopa	Nifedipine	Nicardipine	Nimodipine
Perindopril arginine	Ramipril	Telmisartan	Valsartan
Furosemide	Streptokinase		
Antineoplastics and Immunomodulators			
Capecitabine	Bicalutamide	Cyclosporine	Imatinib mesylate
Nilotinib	Tamoxifen	Cyclophosphamide	Fluorouracil
Gefitinib	Ifosfamide		
Anti-infectives			
Sulfamethoxazole + Trimethoprim	Lopinavir + Ritonavir	Cefoperazone + Sulbactam	Benzathine benzylpenicillin
Phenoxymethyl penicillin	Amoxicillin + Clavulanic acid	Ampicillin + Sulbactam	Procaine benzylpenicillin
Fluconazole	Metronidazole	Voriconazole	Rifampicin
Acyclovir	Ciprofloxacin	Terbinafine	Clarithromycin
Erythromycin	Isoniazid	Itraconazole	Ketoconazole
Efavirenz	Azithromycin	Spiramycin	Cefadroxil
Cephalexin	Cefazolin	Cefepime	Cefixime
Cefoperazone	Cefotaxime	Cefpirome	Ceftazidime
Ceftriaxone	Cefuroxime	Sulfadiazine	Levofloxacin

Table 2. *Continued)*

Anti-infectives			
Moxifloxacin	Ofloxacin	Clindamycin	Amoxicillin
Ampicillin	Doxycycline	Oxytetracycline	Tetracycline
Amikacin	Gentamicin	Streptomycin	Chloramphenicol
Meropenem	Pyrimethamine	Vancomycin	Quinine
Psychopharmaca			
Fluvoxamine	Alprazolam	Fluoxetine	Escitalopram
Sertraline	Amitriptyline	Maprotiline	
Topical Skin Drugs			
Miconazole			
Antiepileptics – Anticonvulsants			
Carbamazepine	Phenobarbital	Phenytoin	Valproate
Gastrointestinal Drugs			
Omeprazole	Ranitidine	Esomeprazole	Lansoprazole
Antacids	Cisapride	Sucralfate	Ondansetron
Loperamide	Sulfasalazine		
Analgesics, Antipyretics, Non-Steroidal Anti-Inflammatories, Antigouts			
Allopurinol	Celecoxib	Sodium Diclofenac	Ibuprofen
Ketoprofen	Ketorolac	Mefenamic acid	Metamizole
Paracetamol	Tramadol	Gabapentin	Pregabalin
Fentanyl	Hydromorphone	Codeine	Morphine
Oxycodone	Pethidine	Remifentanil	Sufentanil
Antimigraine and Antivertigo			
Ergotamine + Caffeine			
Hormones, Other Endocrine Drugs and Contraceptives			
Desogestrel + Ethinylestradiol	Levonorgestrel + Ethinylestradiol	Desogestrel	Lynestrenol
Prednisone	Pioglitazone	Glimepiride	Carbimazole
Levothyroxine	Propylthiouracil	Thiamazole	
Drugs Affecting Blood			
Dabigatran etexilate	Heparin	Enoxaparin	
Vitamins and Minerals			
Calcitriol	Ferrous sulfate		

3 Result and Discussion

The identified digoxin and warfarin interactions from the three AI platforms are presented in Fig. 2. The identified drug interactions are categorized into four sections: no known interaction, minor, moderate, and major.

Fig. 2 shows that each AI platform displays different capabilities in identifying drug interactions. ChatGPT-4o Mini produced the highest number of "no known interaction" responses compared to Copilot and Gemini, yielding 15 cases. In comparison, Copilot and Gemini reported 8 and 6 cases, respectively, out of 212 cases. Despite these discrepancies, all three AI platforms seemed to agree that the "no known interaction" category had the fewest occurrences compared to the other three severity categories. The detailed percentage breakdown of the combined digoxin and warfarin interaction categories with other drugs for each AI platform is as follows: ChatGPT-4o Mini 7.07 %, Copilot 3.77 %, and Gemini 2.83 %, as shown in Table 3.

Subsequently, regarding the minor category, the results indicate that ChatGPT-4o Mini produced the highest number of minor responses compared to Gemini and Copilot. Moreover, the number of minor responses identified by ChatGPT-4o Mini showed a considerable difference compared to Gemini and Copilot. Specifically, ChatGPT-4o Mini identified 80 (37.74 %) cases, Gemini identified 42 (19.81 %) cases, and Copilot identified 35 (16.51 %) cases out of a total of 212 (100 %) cases. The minor category ranked second among the four interaction categories for the ChatGPT-4o Mini platform, while it ranked third for both Copilot and Gemini.

Fig. 2. Each database's drug interaction findings are quantified and organized by severity, with bar color-coding denoting the severity level

Regarding the moderate category, the results revealed that Gemini produced the highest number of moderate responses compared to Copilot and ChatGPT-4o Mini. Moreover, the moderate category exhibited the highest count among the four categories (no known interaction, minor, and major), with Gemini identifying 118 (55.66%) cases, Copilot 104 (49.06%) cases, and ChatGPT-4o Mini 96 (45.28%) cases out of a total of 212 (100%) cases.

Turning to the major category, the data indicated that Copilot yielded the most major responses, surpassing Gemini and ChatGPT-4o Mini. Specifically, Copilot identified 65 (30.66%) cases, Gemini 46 (21.70%) cases, and ChatGPT-4o Mini 21 (9.91%) cases out of 212 (100%) total cases. ChatGPT-4o Mini exhibited a markedly lower number of major interaction identifications compared to Copilot and Gemini. The

major category was ranked second for interaction categorization in Copilot and Gemini, whereas it was ranked third in ChatGPT-4o Mini.

It appears that all AI platforms were in agreement that the moderate category interactions represented the largest percentage when analyzing digoxin and warfarin interactions with other medications, as compared to the other three severity categories, although the platforms differed in their percentage allocations for individual drug interactions, as shown in Table 3.

Table 3. Percentage of Digoxin and Warfarin Interactions Across AI Platforms Based on Severity Level

Platform AI	No Known Interaction (Total)	Minor (Total)	Moderate (Total)	Major (Total)	Total Amount
ChatGPT-4o Mini	15 (7.07%)	80 (37.74%)	96 (45.28%)	21 (9.91%)	212 (100%)
Copilot	8 (3.77%)	35 (16.51%)	104 (49.06%)	65 (30.66%)	212 (100%)
Gemini	6 (2.83%)	42 (19.81%)	118 (55.66%)	46 (21.70%)	212 (100%)

Furthermore, regarding the evaluation of each AI platform's validity in determining digoxin and warfarin interactions with other drugs, which included tests for sensitivity, specificity, positive predictive value, negative predictive value, and accuracy, the results demonstrated that each AI platform exhibited varying capabilities in identifying drug interactions, as shown in Table 4.

Table 4. Sensitivity, Specificity, PPV, NPV, and Accuracy of AI Compared to UpToDate Lexidrug as a Reference

Program	TP	TN	FP	FN	Sensitivity	Specificity	PPV	NPV	Accuracy
ChatGPT-4o Mini	95	55	22	40	0.704	0.714	0.812	0.579	0.708
Copilot	116	24	53	19	0.859	0.312	0.686	0.558	0.660
Gemini	120	33	44	15	0.889	0.429	0.732	0.688	0.722

Table 4 shows that Gemini had the highest sensitivity value in analyzing digoxin and warfarin interactions compared to the other two AI platforms, at 0.889, while Copilot had a sensitivity value of 0.859 and ChatGPT-4o Mini 0.704. These results indicate that Gemini can detect true positive cases better while minimizing false negative errors compared to Copilot and ChatGPT-4o Mini. This suggests that Gemini is suitable for screening digoxin and warfarin interactions with potentially dangerous side effects by minimizing false negatives.

Another study conducted by Al-Ashwal et al. (2023), which evaluated the sensitivity, specificity, and accuracy of ChatGPT-3.5 (now known as ChatGPT-4o Mini), ChatGPT-4, Bing AI (now known as Copilot), and Bard (now known as Gemini) against Micromedex and Drugs.com, using canagliflozin, dapagliflozin, empagliflozin,

clarithromycin, and erythromycin as models paired with 51 of the most commonly prescribed drugs from the "Top 200 of 2020" drug list from the DrugStats Database, found different results compared to our study. This study found that ChatGPT-3.5 (ChatGPT-4o Mini) had a higher sensitivity value compared to Bing (Copilot) and Bard (Gemini) when using Micromedex as a reference, with values of 0.913, 0.872, and 0.830, respectively. Still, in the same study, different results were shown when Drugs.com was used as a reference, indicating that Bing (Copilot) had a higher sensitivity value compared to ChatGPT-3.5 (ChatGPT-4o Mini) and Bard (Gemini), with values of 0.886, 0.823, and 0.823, respectively (Bing and Bard showed the same sensitivity value) [17].

Different results were also shown by the study conducted by Aksoyalp and Erdoğan (2024), which evaluated the sensitivity, specificity, and accuracy of ChatGPT-3.5 (ChatGPT-4o Mini) and Bing (Copilot), using clopidogrel as a model and paired with 78 drugs that interact with clopidogrel according to the UpToDate Lexidrug database, using UpToDate Lexidrug as a reference. The results showed that ChatGPT-3.5 (ChatGPT-4o Mini) had superior sensitivity (0.91) compared to Bing (Copilot) (0.19). However, this study did not include Gemini as one of the AI platforms tested [23].

Furthermore, the highest specificity was observed in ChatGPT-4o Mini, with a value of 0.714, while Gemini and Copilot recorded values of 0.429 and 0.312, respectively. The specificity results for digoxin and warfarin interaction analysis suggest that Gemini and Copilot exhibit low specificity, leading to a higher likelihood of false positive results. In contrast, ChatGPT-4o Mini showed a superior capacity to identify true negative cases, with a specificity value of 0.714.

Similar results were shown by Aksoyalp and Erdoğan (2024) study, which found that ChatGPT-3.5 (ChatGPT-4o Mini) had a higher specificity value compared to Bing (Copilot), with values of 0.97 and 0.78, respectively [23]. In contrast, Al-Ashwal et al. (2023) reported different findings, indicating that Bing (Copilot) had the highest specificity when both Micromedex and Drugs.com were used as references, with values of 0.769 and 0.892, respectively, while the other two AI platforms, Bard (Gemini) and ChatGPT-3.5 (ChatGPT-4o Mini), had specificity values of 0.558 and 0.625 for Bard (Gemini) and 0.375 and 0.392 for ChatGPT-3.5 (ChatGPT-4o Mini) [17].

Regarding the positive predictive value (PPV), the results indicated that ChatGPT-4o Mini had the highest PPV in analyzing digoxin and warfarin interactions compared to Gemini and Copilot, with a value of 0.812, followed by Gemini at 0.732 and Copilot at 0.686. These results suggest that ChatGPT-4o Mini is better at ensuring that all positive predictions are truly positive and minimizing false positive predictions compared to Gemini and Copilot.

Similar findings were reported in the Aksoyalp and Erdoğan (2024) study, which indicated that ChatGPT-3.5 (ChatGPT-4o Mini) had a slightly higher PPV compared to Bing (Copilot), with respective values of 0.97 and 0.93 [23]. Conversely, Al-Ashwal et al. (2023) presented divergent results, revealing that Bing (Copilot) exhibited the highest PPV when both Micromedex and Drugs.com were utilized as references, specifically 0.461 and 0.787, respectively, while the other two AI platforms, Bard (Gemini) and ChatGPT-3.5 (ChatGPT-4o Mini), displayed PPV values of 0.298 and

0.496 for Bard (Gemini), and 0.244 and 0.378 for ChatGPT-3.5 (ChatGPT-4o Mini) [17].

In terms of negative predictive value (NPV), the results indicated that Gemini had the highest NPV in analyzing digoxin and warfarin interactions compared to ChatGPT-4o Mini and Copilot, which had relatively low NPV of 0.579 for ChatGPT-4o Mini and 0.558 for Copilot. Gemini had a higher negative predictive value of 0.688. This suggests that Gemini can better ensure that all negative predictions are truly negative and minimize false negative predictions compared to ChatGPT-4o Mini and Copilot.

The study conducted by Aksoyalp and Erdoğan (2024) indicated that ChatGPT-3.5 (ChatGPT-4o Mini) had a higher negative predictive value (NPV) compared to Bing (Copilot), with respective values of 0.22 and 0.06 [23]. In contrast, Al-Ashwal et al. (2023) presented divergent results, revealing that Bing (Copilot) exhibited the highest NPV when both Micromedex and Drugs.com were utilized as references, specifically 0.964 and 0.946, respectively, while the other two AI platforms, Bard (Gemini) and ChatGPT-3.5 (ChatGPT-4o Mini), displayed NPV values of 0.935 and 0.887 for Bard (Gemini), and 0.951 and 0.831 for ChatGPT-3.5 (ChatGPT-4o Mini) [17].

Regarding accuracy values, the results revealed that Gemini had the highest accuracy, at 0.722, followed by ChatGPT-4o Mini at 0.708 and Copilot at 0.660. Based on these results, it can be stated that Gemini has a greater proportion of correct total predictions (both positive and negative) in analyzing digoxin and warfarin interactions compared to ChatGPT-4o Mini and Copilot. Therefore, Gemini is more reliable for use in clinical settings that require overall consistent predictions regarding drug interactions.

Another study conducted by Al-Ashwal et al. (2023) showed different results. This study found that Bing (Copilot) had a higher accuracy value when using both Micromedex and Drugs.com as references, with accuracy values of 0.788 and 0.890, respectively, followed by Bard (Gemini) with 0.608 and 0.686, and ChatGPT-3.5 (ChatGPT-4o Mini) with 0.472 and 0.525, respectively [17]. In contrast, results from another study by Aksoyalp and Erdoğan (2024) indicated that ChatGPT-3.5 (ChatGPT-4o Mini) had a higher accuracy score, specifically 307, followed by Bing (Copilot) with a score of 197 (maximum score 400), which aligns with our study showing that ChatGPT-3.5 (ChatGPT-4o Mini) had a higher accuracy score compared to Bing (Copilot) [23].

From the research above, it is evident that using different references can yield varying answers or conclusions, thus the performance of each AI may depend on the reference source. This highlights the importance of considering the reference to be used when testing AI validity [17]. Our study utilized UpToDate Lexidrug as a reference for several reasons. First, a study conducted by Kheshti et al. (2016) demonstrated that UpToDate Lexidrug has the highest accuracy and comprehensiveness scores compared to other drug information databases, indicating its superior performance [25]. Second, research by Pinkoh et al. (2023) showed that UpToDate Lexidrug provides the most extensive drug list compared to other drug information databases [27]. Additionally, a study by Patel and Beckett (2016) revealed that UpToDate Lexidrug ranks among the drug information databases with high scope and completeness scores for providing drug interaction information [16].

Furthermore, the selection of drug pairs tested can also influence the results, as using drug pairs with clinically significant interactions will yield higher sensitivity values compared to a random drug list. This aligns with research conducted by Rada Krishnan (2024), which stated that differences in sensitivity and specificity results compared to previous studies are due to the use of highly selective drug pairs (those with clinically significant drug interaction rates), which may yield different results than common drug pairs derived from real-world settings (prescriptions directly obtained from healthcare facilities) [28].

Several key findings emerged from this study. Copilot and Gemini provided source citations for some drug interaction pairs in the form of accessible hyperlinks, whereas ChatGPT-4o Mini did not offer any. Copilot included hyperlinks for 181 out of 212 drug pairs, while Gemini included a smaller number, specifically 50 drug pairs. These findings are illustrated in Fig. 3 and 4.

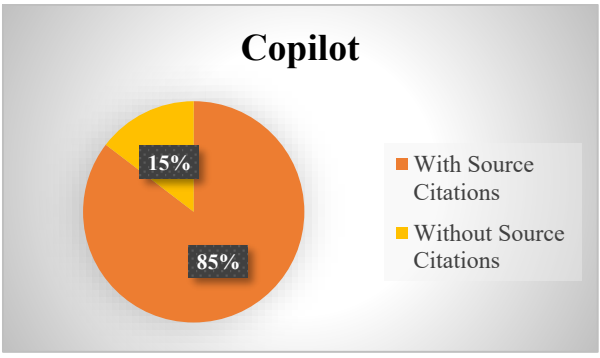

Fig. 3. Percentage of Sources Listed and Not Listed by Copilot in Drug Interaction Responses

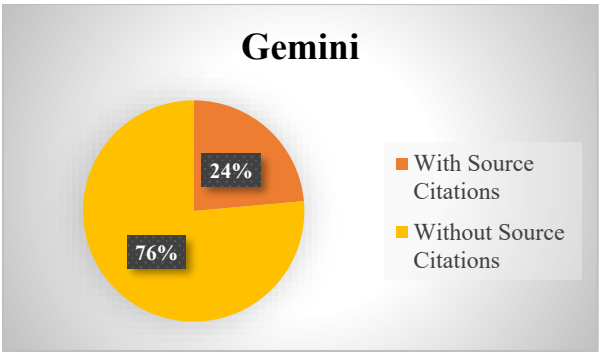

Fig. 4. Percentage of Sources Listed and Not Listed by Gemini in Drug Interaction Responses

The primary sources referenced by Copilot, as indicated by the provided hyperlinks, were Drugs.com, followed by UpToDate, Wikipedia, Springer, HelloPharmacist, DrugBank, GoodRx, RemediRx, WebMD, eHealthMe, Queensland Health

(Queensland Government), Mayo Clinic, and JAMA Network. On the other hand, Gemini predominantly cited PubMed, followed by MedicalHubNews, Drugs.com, ResearchGate, National Health Service (England), Mayo Clinic, and Taylor & Francis. Al-Ashwal et al. (2023) reported that Bing AI (Copilot) frequently cited Drugs.com as a direct reference when addressing drug interaction queries [17]. This observation aligns with our study's findings, where Copilot embedded Drugs.com hyperlinks in 178 of the 212 drug interaction pairs examined.

Finally, within the evolving domain of future healthcare, generative AI platforms present substantial opportunities for enhancing patient care, notably in identifying and mitigating drug-drug interactions. These AI systems have the potential to minimize the risk of adverse drug events through the provision of rapid, thorough analyses of potential interactions, thereby contributing to improved patient safety. Nevertheless, further refinements in their validity are imperative before being applied in real-world settings because some of the outputs were inaccurate or incomplete, which highlights the need for further improvement. Generative AI platforms like these can be helpful, but they should not replace healthcare professionals in making clinical decisions.

4 Conclusion

This study revealed that Gemini demonstrated superiority in accuracy, sensitivity, and negative predictive value, while ChatGPT-4o Mini excelled in specificity and positive predictive value. This indicates that both AI platforms are competitive in predicting drug-drug interactions; however, overall, Gemini was deemed superior compared to the other two platforms. Copilot provided the most references when answering questions, whereas ChatGPT-4o Mini did not provide any references. In general, generative AI platforms present substantial opportunities for enhancing patient care, notably in identifying and mitigating drug-drug interactions. Nevertheless, shortcomings remain, and further refinements in their validity are imperative before being applied in real-world settings. Some of the outputs were inaccurate or incomplete, which highlights the need for further improvement. Generative AI platforms can be helpful, but they should not replace healthcare professionals in making clinical decisions.

Acknowledgments. We would like to express our gratitude to the Faculty of Pharmacy, Universitas Gadjah Mada, for facilitating the research, and to the 2025 Faculty Research Grant for supporting the publication.

References

1. M. Mutalik and D. Sanghavi, Review of Drug Interactions: A Comprehensive Update, *British Journal of Pharmaceutical Research*, **4** (2014).
2. M. Askari *et al.*, Frequency and nature of drug-drug interactions in the intensive care unit, *Pharmacoepidemiology and Drug Safety* **22**, 4 (2013).

3. T. Hammar, S. Hamqvist, M. Zetterholm, P. Jokela, and M. Ferati, Current Knowledge about Providing Drug-Drug Interaction Services for Patients—A Scoping Review, *Pharmacy* **9**, 2 (2021).
4. F. Sjöqvist and Y. Böttiger, Historical perspectives: drug interactions - it all began with cheese, *Journal of Internal Medicine* **268**, 6 (2010).
5. L. M. de Oliveira, J. do A. C. Diel, A. Nunes, and T. da S. D. Pizzol, Prevalence of drug interactions in hospitalised elderly patients: a systematic review, *Eur J Hosp Pharm* **28**, 1 (2021).
6. B. Snyder, T. M. Polasek, and M. P. Doogue, Drug interactions: principles and practice, *Aust Prescr* **35**, 3 (2012).
7. R. T. Scheife *et al.*, Consensus recommendations for systematic evaluation of drug-drug interaction evidence for clinical decision support, *Drug Saf* **38**, 2 (2015).
8. K. Iyer, N. Dilipkumar, S. Vasaya, S. Pawar, and A. Diwan, Comparison of Drug Related Problems Associated with Use of Narrow Therapeutic Index Drugs and Other Drugs in Hospitalized Patients, *JYP* **10**, 3 (2018).
9. M. N. V. David and M. Shetty, *"Digoxin," in StatPearls*, StatPearls Publishing (2023).
10. S. Patel, R. Singh, C. V. Preuss, and N. Patel, *"Warfarin," in StatPearls*, Treasure Island (FL): StatPearls Publishing (2024).
11. K. Wang *et al.*, Drug-drug interactions induced by Linderane based on mechanism-based inactivation of CYP2C9 and the molecular mechanisms, *Bioorg Chem*, **118** (2022).
12. M. F. Crader, T. Johns, and J. K. Arnold, *"Warfarin Drug Interactions," in StatPearls*, Treasure Island (FL): StatPearls Publishing (2024).
13. A. Di Minno *et al.*, Old and new oral anticoagulants: Food, herbal medicines and drug interactions, *Blood Rev* **31**, 4 (2017).
14. M. Onysko, N. Holcomb, and J. Hornecker, Antibiotic interactions: Answers to 4 common questions, *J Fam Pract* **65**, 7 (2016).
15. D. N. Juurlink, M. Mamdani, A. Kopp, A. Laupacis, and D. A. Redelmeier, Drug-drug interactions among elderly patients hospitalized for drug toxicity, *JAMA* **289**, 13 (2003).
16. R. I. Patel and R. D. Beckett, Evaluation of resources for analyzing drug interactions, *J Med Libr Assoc* **104**, 4 (2016).
17. F. Y. Al-Ashwal, M. Zawiah, L. Gharaibeh, R. Abu-Farha, and A. N. Bitar, 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*, **15** (2023).
18. O. Khan, M. Parvez, P. Kumari, S. Parvez, and S. Ahmad, The future of pharmacy: How AI is revolutionizing the industry, *Intelligent Pharmacy* **1**, 1 (2023).
19. L. Liu and V. G. Duffy, Exploring the future development of Artificial Intelligence (AI) Applications in Chatbots: A bibliometric analysis, *Int J of Soc Robotics* **15**, 5 (2023).
20. P. Rivas and L. Zhao, Marketing with ChatGPT: Navigating the Ethical Terrain of GPT-Based Chatbot Technology, *AI* **4**, 2 (2023).
21. D. Giansanti, A Deep Dive into the Nexus between Digital Health and Life Sciences Amidst the COVID-19 Pandemic: An Editorial Expedition, *Life* **13**, 5 (2023).
22. M. Wang *et al.*, Drug-drug interactions with warfarin: A systematic review and meta-analysis, *British Journal of Clinical Pharmacology* **87**, 11 (2021).
23. Z. Ş. Aksoyalp and B. R. Erdoğan, Comparative Evaluation of Artificial Intelligence and Drug Interactions Tools: A Perspective with the Example of Clopidogrel, *J. Fac. Pharm. Ankara* **48**, 3 (2024).

24. L. A. Marcath, J. Xi, E. K. Hoylman, K. M. Kidwell, S. L. Kraft, and D. L. Hertz, Comparison of Nine Tools for Screening Drug-Drug Interactions of Oral Oncolytics, *J Oncol Pract* **14**, 6 (2018).
25. R. Kheshti, M. Aalipour, and S. Namazi, A comparison of five common drug-drug interaction software programs regarding accuracy and comprehensiveness, *J Res Pharm Pract* **5**, 4 (2016).
26. J. B. Bossaer, D. Eskens, and A. Gardner, Sensitivity and specificity of drug interaction databases to detect interactions with recently approved oral antineoplastics, *J Oncol Pharm Pract* **28**, 1 (2022).
27. R. Pinkoh, R. Rodsiri, and S. Wainipitapong, Retrospective cohort observation on psychotropic drug-drug interaction and identification utility from 3 databases: Drugs.com®, Lexicomp®, and Epocrates®, *PLoS One* **18**, 6 (2023).
28. R. P. Radha Krishnan *et al.*, Evaluating the capability of ChatGPT in predicting drug-drug interactions: Real-world evidence using hospitalized patient data, *Br J Clin Pharmacol* **90**, 12 (2024).