Original Paper

Extracting Clinical Guideline Information Using Two Large Language Models: Evaluation Study

Hsing-Yu Hsu[1,2], MS; Lu-Wen Chen[3], MS; Wan-Tseng Hsu[1], PhD; Yow-Wen Hsieh[2,4], PhD; Shih-Sheng Chang[3,5,6], PhD

[1]Graduate Institute of Clinical Pharmacy, College of Medicine, National Taiwan University, Taipei, Taiwan
[2]Department of Pharmacy, China Medical University Hospital, Taichung, Taiwan
[3]Artificial Intelligence Center, China Medical University Hospital, Taichung, Taiwan
[4]School of Pharmacy, College of Pharmacy, China Medical University, Taichung, Taiwan
[5]Division of Cardiovascular Medicine, Department of Medicine, China Medical University Hospital, Taichung, Taiwan
[6]School of Medicine, China Medical University, Taichung, Taiwan

Corresponding Author:

Shih-Sheng Chang, PhD
Artificial Intelligence Center
China Medical University Hospital
2, Yude Road
Taichung 404327
Taiwan
Phone: 886 4-22052121
Email: chssheng@gmail.com

Abstract

Background: The effective implementation of personalized pharmacogenomics (PGx) requires the integration of released clinical guidelines into decision support systems to facilitate clinical applications. Large language models (LLMs) can be valuable tools for automating information extraction and updates.

Objective: This study aimed to assess the effectiveness of repeated cross-comparisons and an agreement-threshold strategy in 2 advanced LLMs as supportive tools for updating information.

Methods: The study evaluated the performance of 2 LLMs, GPT-4o and Gemini-1.5-Pro, in extracting PGx clinical guidelines and comparing their outputs with expert-annotated evaluations. The 2 LLMs classified 385 PGx clinical guidelines, with each recommendation tested 20 times per model. Accuracy was assessed by comparing the results with manually labeled data. Two prospectively defined strategies were used to identify inconsistent predictions. The first involved repeated cross-comparison, flagging discrepancies between the most frequent classifications from each model. The second used a consistency threshold strategy, which designated predictions appearing in less than 60% of the 40 combined outputs as unstable. Cases flagged by either strategy were subjected to manual review. This study also estimated the overall cost of model use and was conducted between October 1 and November 30, 2024.

Results: GPT-4o and Gemini-1.5-Pro yielded reproducibility rates of 97.8% (7534/7700) and 98.9% (7612/7700), respectively, based on the most frequent classification for each query. Compared with expert labels, GPT-4o achieved 93.5% accuracy (Cohen κ=0.90; P<.001) and Gemini-1.5-Pro 92.7% accuracy (Cohen κ=0.89; P<.001). Both models demonstrated high overall performance, with comparable weighted average F_1-scores (GPT-4o: 0.929; Gemini: 0.935). The models generated consistent predictions for 341 of 385 guideline items, reducing the need for manual review by 88.6%. Among these agreed-upon cases, only one (0.3%) diverged from expert labels. Applying a predefined agreement-threshold strategy further reduced the number of priority manual review cases to 2.9% (11/385), although the error rate slightly increased to 0.5% (2/374). The inconsistencies identified through these methods prompted the prioritization of manual review to minimize errors and enhance clinical applicability. The total combined cost of using both LLMs was only US $0.76.

Conclusions: These findings suggest that using 2 LLMs can effectively streamline PGx guideline integration into clinical decision support systems while maintaining high performance and minimal cost. Although selective manual review remains necessary, this approach offers a practical and scalable solution for PGx guideline classification in clinical workflows.

J Med Internet Res 2025;27:e73486; doi: 10.2196/73486

Keywords: large language models; reproducibility; reliability; Guideline Classification; Pharmacogenomics; Clinical Decision Support System

Introduction

As pharmacogenomic (PGx) testing becomes more widely available, enhancing decision-making and knowledge-sharing in clinical genetics is essential for integrating PGx data into routine clinical practice [1]. PGx knowledge is new, complex, and constantly evolving, making it inadequate to rely solely on clinicians' expertise for clinical implementation [2]. We developed a comprehensive database of drug-gene interactions and incorporated it into clinical decision support systems (CDSS) to achieve truly personalized treatment, improve patient outcomes, and optimize treatment strategies.

However, maintaining CDSS knowledge bases is crucial to ensuring alignment with the evolving nature of medical practice and clinical guidelines [3,4]. These labor-intensive activities necessitate manual input and significantly contribute to burnout among pharmacists. Recent research has highlighted the transformative potential of generative artificial intelligence (AI), particularly in leveraging large language models (LLMs) for data analysis and text generation to make complex medical information more accessible to health care providers and patients [5-7]. However, applying PGx guidelines as structured input for AI models remains rare, and even fewer studies have explored how to summarize classification task results effectively [8].

We hypothesize that using cross-comparison methods with 2 state-of-the-art LLMs can facilitate real-time updates to CDSS knowledge bases in health care institutions. Our objective is to enhance the clinical applicability and reliability of medical information while improving the dissemination of PGx knowledge.

Methods

Study Design

This methodological study followed the enhanced Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD)-LLM reporting guidelines [9].

Ethical Considerations

This study did not involve human participants, identifiable personal data, or biological materials. Therefore, the Institutional Review Board of China Medical University Hospital determined that the study did not require ethical review or informed consent. All data used in the study were publicly available or derived from deidentified, nonhuman sources. No interventions, clinical recruitment, or patient-level interactions were conducted.

Model Selection and Data Sources

GPT-4o and Gemini-1.5-Pro were selected for their stable and publicly available cloud-based application programming interfaces (APIs), which support integration into clinical workflows and system deployment. In addition to their practical usability, both models have demonstrated strong performance in medical language understanding and instruction adherence in prior biomedical natural language processing studies [10,11]. Accordingly, we used these 2 LLMs to perform the classification tasks, analyzing drug-gene interaction recommendations derived from 385 clinical guideline annotations in PharmGKB as of October 5, 2024. Details of the retrospective datasets from PharmGKB are provided in Table S1 in Multimedia Appendix 1. All experiments and analyses involving the tested LLMs were conducted in English between October 1 and November 30, 2024.

Classification Process and Prompt Design

Researchers supplied LLMs (GPT-4o and Gemini-1.5-Pro) with pertinent information on drug-gene interactions. Using this data, LLMs were instructed to generate recommendations based on a predefined prompt. To enable rapid adaptation to complex tasks and efficient data processing, we used zero-shot prompting techniques. The prompt used was: " Your task is to summarize the following content: Provide a treatment recommendation categorized into one of the following options: "No action needed," "Consider dosage modification," "Change medication," or "Monitor adverse effects." If none of these categories fit, you may respond with "Other." If both dosage modification and adverse effect monitoring were recommended, priority was given to classifying the recommendation under dosage modification. If both dosage modification and changing medication were recommended, priority was given to classifying the recommendation under changing medication. Please directly give "No action needed," "Consider dosage modification," "Change medication," or "Monitor adverse effects." The classification tasks of these 4 options must comply with the CDSS system update requirements, and there is no need to explain the reasons." For each input query, separate responses from GPT-4o and Gemini-1.5-Pro were recorded and are presented in Tables S2 and S3 in Multimedia Appendix 1.

Evaluation of Reproducibility, Accuracy, and Human Validation

GPT-4o and Gemini-1.5-Pro were used to evaluate 385 drug-gene interaction recommendations. Each recommendation underwent 20 repeated tests, generating 7700 responses per model (15,400 in total). Reproducibility was assessed by calculating the proportion of consistent responses across the 20 repetitions for each recommendation, using the most

frequent response per model as the reference. Accuracy was evaluated by comparing model outputs with human-labeled data (Table S4 in Multimedia Appendix 1).

To further enhance accuracy and mitigate misclassification and hallucination issues, we implemented 2 strategies: (1) repeated cross-comparisons and (2) agreement-threshold strategy.

Repeated Cross-Comparisons

Classifications were first assessed within each model. Recommendations with inconsistent outputs between models were flagged for manual review.

Agreement-Threshold Strategy

For each recommendation, we aggregated 40 outputs from both models. If a classification appeared in ≥60% of responses (at least 24 out of 40 occurrences), it was considered stable and accepted as the final result. Those below the threshold triggered manual review.

Expert Validation

To facilitate expert validation, 5 pharmacists from hospitals of varying sizes participated in reviewing and classifying the guideline content. To establish a reference standard for comparison with LLM-generated outputs, each pharmacist independently reviewed the PGx guidelines and categorized the therapeutic recommendations using the same classification scheme as the LLMs. Recommendations were assigned to the "Other" category if none of the predefined options were applicable. A final consensus meeting was conducted to resolve discrepancies and determine the definitive reference labels.

Technical Implementation and CDSS Integration

PGx guideline information was initially curated by clinical pharmacists and stored in JSON format. A Python-based pipeline was developed to parse drug names, gene symbols, and recommendation summaries, which were then standardized into 385 unique drug-gene pairs. Each pair was processed 20 times by both GPT-4o and Gemini-1.5-Pro using fixed prompts via publicly available cloud-based APIs. The outputs were recorded in a structured tabular format, where each column represented one model repetition, enabling reproducible and scalable classification. To integrate the results into clinical workflows, model outputs were reviewed for consistency. Cases with discrepancies across repetitions or between models were flagged for manual review. Verified results were then automatically formatted for incorporation into the existing CDSS database.

Economic Evaluation

We conducted a preliminary evaluation of cost based on publicly available API pricing data. As of April 17, 2025, according to the pricing overview published on the Azure OpenAI Service website, the cost for GPT-4o is US $2.50 per million tokens for input and US $10.00 per million tokens for output. For Gemini 1.5 Pro, based on the pricing listed on Google Cloud's Vertex AI Generative AI service, the input cost is US $0.3125 per million tokens, and the output cost is US $1.25 per million tokens.

Statistical Analysis

Descriptive statistics were used to report the reproducibility results of the classification tool. We also evaluated the accuracy of GPT-4o and Gemini-1.5-Pro in comparison with human-labeled data. For each LLM, classifications that agreed with human-labeled data were coded as 1, while disagreements were coded as 0. Cohen κ was used to assess the level of agreement. The κ statistic measures the level of agreement between 2 raters, accounting for agreement expected by chance [12]. Cohen κ was computed using the irr package in R (version 4.4.0; R Foundation for Statistical Computing). A P value of less than .05 was considered statistically significant. We calculated category-specific metrics to comprehensively evaluate the performance of the 2 LLM models. These metrics included precision (true positive number / total positive number), recall (sensitivity, true positive number / total actual number), and F_1-score (2× [precision / [precision+ recall]). All statistical analyses were conducted using R.

Results

Among 385 PGx guideline recommendations, the majority were provided by the Clinical Pharmacogenetics Implementation Consortium (265/385, 69%). Most PGx interactions involved *CYP2D6*, *CYP2C9*, and *CYP2C19*. The largest proportion of guidelines pertained to drugs targeting the nervous system (147/385, 38.2%) and the cardiovascular system (115/385, 29.9%). These were followed by antineoplastic and immunomodulating agents (37/385, 9.6%). Drugs targeting the musculoskeletal system, anti-infectives for systemic use, and the alimentary tract and metabolism each accounted for 20 (5.2%). Medications related to the blood and blood-forming organs represented 4.4% (n=17), while those targeting the respiratory system were the least represented, comprising 2.3% (n=9) (Figure 1; Table S1 in Multimedia Appendix 1).

Among 7700 responses, GPT-4o produced 7534 consistent responses (97.8%; 95% CI 0.977-0.982), while Gemini-1.5-Pro generated 7612 consistent responses (98.9%; 95% CI 0.986-0.991) (Tables S2 and S3 in Multimedia Appendix 1). Human labeling consistency was 97.2% (95% CI 0.953-0.987) (Table S4 in Multimedia Appendix 1). A total of 385 drug-gene interaction recommendations were compared with final classifications made by human experts. GPT-4o achieved agreement in 360 recommendations (accuracy: 93.5%; 95% CI 0.906-0.958; Cohen κ=0.90; P<.001), while Gemini-1.5-Pro agreed on 357 recommendations (accuracy: 92.7%; 95% CI 0.897-0.951; Cohen κ=0.89; P<.001) (Figure 2). The F_1-scores of GPT-4o and Gemini ranged from 0.878 to 0.988 across different classes. The weighted average F_1-scores were also very similar between the 2 models (GPT-4o: 0.929; Gemini: 0.935) (Table S5 in Multimedia Appendix 1).

Figure 1. Distribution of pharmacogenomic guidelines by source and gene target. This sunburst chart illustrates the distribution of 385 PGx guideline recommendations across 3 major sources: CPIC, DPWG, and CPNDS. The inner ring represents the guideline sources, while the outer ring details the corresponding genes and the number of guidelines associated with each gene. CPIC: Clinical Pharmacogenetics Implementation Consortium; CPNDS: Canadian Pharmacogenomics Network for Drug Safety; DPWG: Dutch Pharmacogenetics Working Group.

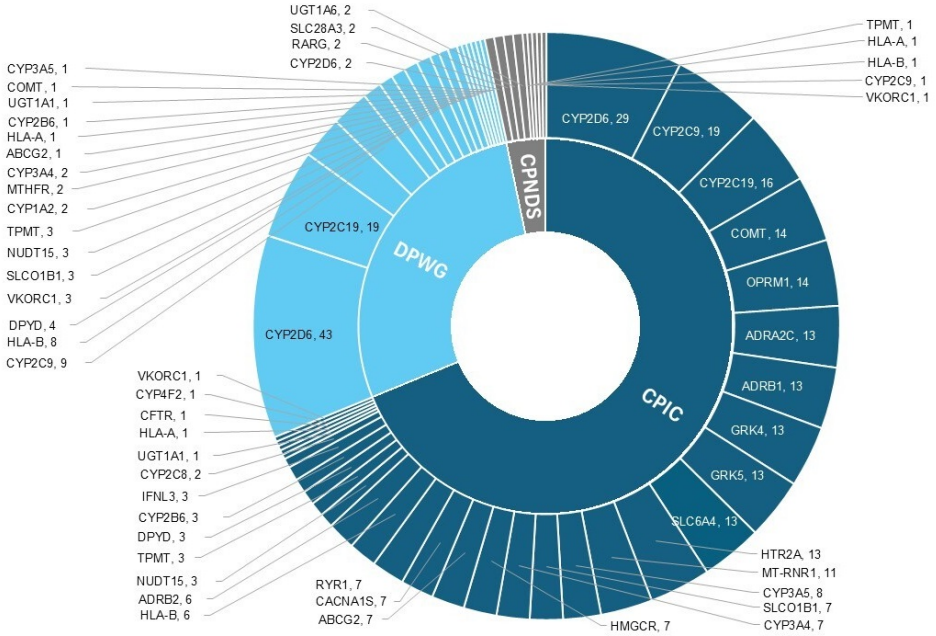

Figure 2. Reproducibility and accuracy rates of pharmacogenomic guideline classifications using GPT-4o, Gemini, and human review. This figure compares the reproducibility and accuracy rates of PGx guideline classifications using GPT-4o, Gemini, and human review. GPT-4o and Gemini showed high reproducibility rates (97.8% and 98.9%, respectively), and slightly lower accuracy rates (93.5% and 92.7%) when compared with human consensus labels. Human annotations, used as the gold standard for evaluation, yielded 100% accuracy. However, initial interrater reproducibility among human reviewers before consensus was 97.2% (see results section).

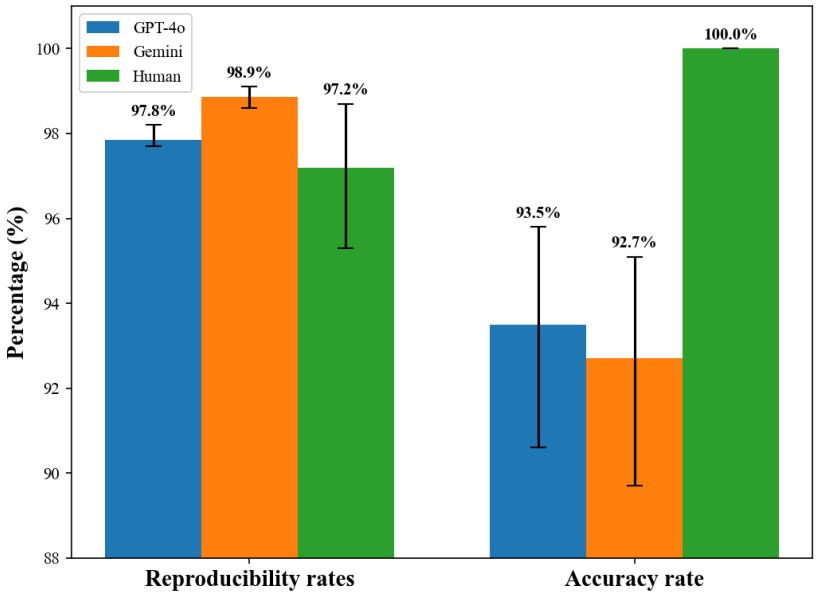

Table 1 shows that the 2 models generated consistent predictions for 341 guideline classifications, while 44 classifications exhibited discrepancies. An analysis of the 44 guideline classifications with discrepancies revealed several recurring patterns. Specifically, 19 guidelines indicated that therapy should vary according to the patient's metabolic type. Nine guidelines involved narrow therapeutic index drugs, for which dose adjustment should be carefully considered. Another 9 guidelines recommended monitoring for adverse events, even when the therapy was not explicitly advised. Five guidelines lacked sufficient clinical evidence to support a clear recommendation, and 2 provided dual options,

advising avoidance of the drug or, if necessary, adjusting the dosage accordingly (Table S6 in Multimedia Appendix 1).

Table 1. Predictive performance and classification error/hallucination rates for repeated cross-comparisons and agreement-threshold strategy.

Variables	Repeated cross-comparisons (n=385), n (%)	Agreement-threshold strategy (n=385), n (%)
Predictive performance		
Classifiable guidelines	341 (88.6)	374 (97.1)
Unclassifiable guidelines (manual review rate)[a]	44 (11.4)	11 (2.9)
Classification error/hallucination rate		
Correctly classified	340 (99.7)	372 (99.5)
Incorrectly classified (hallucinations)	1 (0.3)	2 (0.5)

[a]Manual review rate (%) = (number of unclassifiable cases / total number of PGx guidelines)×100.

Applying a cross-comparison approach between the models reduced the proportion of guideline classifications requiring manual review to 11.4% (44/385). Furthermore, applying the predefined 60% agreement threshold (ie, at least 24 out of 40 classifications aligning with a single guideline question) further reduced the number of cases requiring prioritized manual review to 2.9% (11/385). However, this approach slightly increased the error rate, from 0.3% (1/341) to 0.5% (2/374). Table 2 lists the guideline questions where hallucinations persisted despite both AI approaches. For the 385 PGx clinical guideline entries, the total number of tokens processed in GPT-4o was 84,098 for input and 21,574 for output. In comparison, Gemini 1.5-Pro processed 446,284 input characters and 154,880 output characters. The combined total cost for running both LLMs was US $0.76 (Table 3).

Table 2. Pharmacogenomic interaction guidelines associated with error rates or hallucination generation. This table displays pharmacogenomic interaction guidelines linked to error rates or hallucination generation during classification. It details the specific guidelines, artificial intelligence (AI) tool thresholds used, large language model (LLM) answers, manual review results, and the reasons for discrepancies with human labels. Discrepancies were mainly due to conflicting classifications for different genotypes or model-generated information beyond the scope of the guidelines.

Pharmacogenomic interaction guidelines	AI tool threshold used	LLM answers	Manual review result	Guideline content	Reasons for discrepancies with human labels (hallucinations)
No.254 trimipramine - *CYP2D6*	Repeated cross-comparisons	Consider dosage modification	Change medication	• CYP2D6 UM/PMs, CYP2C19 UM/RM/PMs: Use an alternative drug. • CYP2D6 or CYP2C19 PMs (if used): Reduce dose by 50%. • CYP2D6 IMs: Reduce dose by 25%.	Guideline prompts conflicting classifications for different genotypes
No.211 imipramine - *CYP2D6*	Agreement-threshold strategy	Consider dosage modification	Change medication	• CYP2D6 UM/PMs, CYP2C19 UM/RM/PMs: Use an alternative drug. • CYP2D6 or CYP2C19 PMs (if used):	Guideline prompts conflicting classifications for different genotypes

Pharmacogenomic interaction guidelines	AI tool threshold used	LLM answers	Manual review result	Guideline content	Reasons for discrepancies with human labels (hallucinations)
No.278 daunorubicin - *SLC28A3*	Agreement-threshold strategy	Consider dosage modification	Other	Reduce dose by 50%. • CYP2D6 IMs: Reduce dose by 25%. • Perform pharmacogenomic testing for RARG rs2229774, SLC28A3 rs7853758, and UGT1A6*4 (rs17863783)	Model-generated extraneous information beyond guideline content

Table 3. Cost comparison of GPT-4o and Gemini 1.5-Pro based on token and character use.

Cost type[a]	GPT-4o	Gemini 1.5-Pro
Input cost (US $)	0.21 (total input tokens: 84,098)	0.14 (total input characters: 446,284)
Output cost (US $)	0.22 (total output tokens: 21,574)	0.19 (total output characters: 154,880)
Total cost (US $)	0.43	0.33

[a]GPT-4o pricing is based on token use, whereas Gemini 1.5-Pro pricing is based on character count. Token-to-character conversions may vary depending on language structure.

Discussion

Principal Findings

To our knowledge, this is the first study to evaluate the accuracy and reproducibility of PGx guideline classification using 2 advanced LLMs simultaneously. Our approach was integrated into the existing CDSS workflow via automated guideline updates, supplemented by a manual review of inconsistent classifications. This reduces documentation burden and supports real-time updates. Although prior research suggests that LLM reproducibility may decline over time [13], our findings demonstrate that GPT-4o and Gemini-1.5-Pro consistently generated stable outputs across repeated classifications at the same time point, indicating high reliability. For simplified PGx recommendations, both models produced outputs comparable to expert annotations, particularly in classifying drug-gene interaction recommendations. The simultaneous use of 2 LLMs and repeated testing improved classification accuracy and reduced low-frequency hallucinations.

Comparison to Prior Work

Research on LLM applications in medical genetics is limited. One study explored genetic education for BRCA1-related cancer syndrome, MLH1-related cancer syndrome, and HFE-related hemochromatosis [14]. Other studies have used chatbots to educate high-risk patients, provide cancer risk assessments, and promote preventive genetic testing [15,16]. Our study applies LLMs to clinical management, classifying PGx guidelines for integration into CDSS, enhancing health care education, and PGx application. Optimizing medication management requires understanding gene-drug interactions [17]. PGx-CDSS, a key clinical decision support tool, bridges knowledge gaps for health care providers and patients [18]. However, careful implementation is crucial to prevent alert fatigue and risks of liability [19].

Restricting output classifications also minimizes medical risks and enhances clinical decision support. Another potential application of this technology is enabling health care professionals to guide patients in precise pharmacotherapy and reduce adverse drug reactions. A study evaluated various publicly available LLMs for their ability to answer cancer-related questions. No clinically significant hallucinations were observed in simple queries [20]. Similar research fine-tuned Bidirectional Encoder Representations from Transformers models for classifying pharmacy publications to enhance clinical workflow applicability and reduce bias and hallucinations [21]. Our study confirms that leveraging zero-shot performance from 2 advanced LLMs provides a promising classification tool for simple tasks.

The classification system we adopted for PGx-drug interactions, “no action needed”, “consider dosage modification,” “change medication,” and “monitor adverse effects” was inspired by the structure of established drug-drug interaction databases, such as UpToDate Lexidrug, which typically provide a single recommended action for each

interaction. It is worth noting that in certain clinical scenarios, multiple management strategies, such as dose adjustment and adverse effect monitoring, may both be appropriate. However, we prioritized recommendations involving active therapeutic changes (eg, dosage modification or medication substitution) over monitoring alone. This prioritization assumes that monitoring should naturally follow any therapeutic intervention, whereas recommending monitoring as the sole action may lead clinicians to overlook the need for proactive treatment adjustments. This complexity reflects real-world clinical decision-making, where nuanced judgment is often required. By clearly defining and prioritizing actionable strategies within our classification system, we aimed to enhance the clarity of both model training and output interpretation.

In this study, discrepancies in 44 guideline classifications between the 2 LLMs were used to identify cases warranting prioritized manual review. As described in Table S6 in Multimedia Appendix 1, an analysis of these inconsistencies reveals several recurring patterns. In some instances, treatment recommendations varied according to the patient's metabolizer phenotype. In others, guidelines presented dual options, such as advising complete drug avoidance or, alternatively, recommending dose adjustment if use was deemed necessary. Certain guidelines did not offer explicit therapeutic recommendations but suggested monitoring for potential adverse effects. Dose adjustments were also frequently recommended for drugs with a narrow therapeutic index. Last, some discrepancies resulted from insufficient supporting evidence, leading to recommendations favoring alternative therapies. These patterns underscore the inherent complexity of guideline interpretation, particularly in contexts involving conditional recommendations or limited clinical data.

After verification, the manual review workload was reduced by 88.57%. A further examination of 341 consistent results from both LLMs revealed one (0.29%) discrepancy with human-labeled data, primarily due to a guideline prompting 2 different classification decisions for distinct genotypes, leading to inconsistent LLM responses. Many studies evaluating the capabilities of AI chatbots use a threshold of >60% as a benchmark for adequate knowledge and reasoning performance [22,23]. Notably, when applying an LLM agreement threshold of >60%, the manual review workload was reduced by 97.14% compared with human-reviewed guidelines. Among 374 guidelines, 2 (0.53%) showed discrepancies with human-labeled data. Interestingly, while the daunorubicin-SLC28A guideline recommends PGx testing for RARG rs2229774, SLC28A3 rs7853758, and UGT1A6*4 (rs17863783), it lacks specific treatment recommendations. Nonetheless, LLM responses suggested dose adjustments. Studies link hENT1 expression to increased Ara-C activity and altered drug metabolism, potentially affecting pharmacokinetics, though the evidence remains inconclusive [24]. Despite these inconsistencies, these errors likely do not compromise medical quality, as each classification partially aligns with guideline recommendations. Moreover, clinicians can review the original guideline recommendations if they have concerns about the classification summary. In deployment settings, guideline classifications with inconsistent outputs, either across repeated model runs or between models, would be automatically flagged and subjected to pharmacist validation before CDSS integration.

The challenge lies in the fact that even when the same prompts and models are used, slight randomness may lead to minor variations in the generated results [25]. Nevertheless, both LLMs demonstrated high accuracy and high reproducibility, suggesting that the observed discrepancies might merely reflect random fluctuations and that a majority-based approach considering the statistical distribution of outputs could be used. Simplifying text and classifying it did not compromise the quality of health information, as the LLM was explicitly directed to produce a particular output [26]. Further comparison revealed that both GPT-4o and Gemini performed well across the 3 major categories—no action needed, considered dosage modification, and changed medication—which together accounted for 97.7% of all samples. When considering class-specific metrics and class distribution, both models continued to demonstrate performance levels approaching that of expert pharmacists. In fact, these findings align with previous research, indicating that the quality of a model's responses depends on the content of its input. A prior study showed that using an electronic health record–integrated generative AI chatbot to automatically draft responses to patient messages can streamline workflows and reduce burnout [27].

Economic Estimation

We conducted a preliminary evaluation of cost and response time based on publicly available API pricing and average latency data. The economic estimation indicated that the total combined cost for running both LLMs was only US $0.76. This low cost is primarily attributable to the controlled volume of input and output tokens or characters. When managed appropriately, the combined use of both models demonstrates strong feasibility for real-world implementation, offering an excellent cost-performance balance. Both GPT-4o and Gemini 1.5-Pro operate through cloud-based APIs, with response times averaging between 2 and 4 seconds. The current results suggest that using repeated cross-comparisons and an agreement-threshold strategy can effectively narrow the scope of manual review. This approach may be a viable strategy and provides an important reference for the practical implementation of PGx CDSS in clinical settings.

Limitations

This study has several limitations. First, the PGx guidelines we used were designed for personalized pharmacotherapy and may differ in structure or language from guidelines in other medical fields, potentially limiting generalizability. To mitigate this, we focused on clearly structured guideline recommendations and standardized the prompt format to ensure interpretability and consistency. Second, LLMs generate output based on probabilistic distributions, which can produce biased responses. However, repeated inputs

and constrained target outputs can approximate the most likely distribution of responses. Third, determining the correct classification can be challenging. In some clinical scenarios, multiple management strategies, such as dose adjustment and adverse effect monitoring, may all be appropriate. However, we prioritized recommendations involving active therapeutic changes. Finally, we observed an inter-rater agreement of only 97.2% among human reviewers, highlighting the inherent heterogeneity in manual classification. Discrepancies were resolved through consensus discussion, ensuring a standardized reference for comparison with LLM-generated outputs. Future research should focus on how to automate further the integration of more complex and diverse clinical guideline updates into real-world clinical workflows.

Conclusions

These results suggest that using 2 LLMs simultaneously can simplify PGx guidelines and generate classification outputs that enhance CDSS databases. However, manual review and cautious application remain necessary, as the models are not entirely infallible. Repeated cross-comparisons and an agreement-threshold strategy offer a promising approach for updating CDSS guidelines.

Acknowledgments

We would like to thank our participants and colleagues for generously dedicating their time to the classification of pharmacogenomic guidelines. This study would not have been possible without their contributions. Special thanks to Pharmacist Chih-Kang Lin (Department of Pharmacy, Cheng Ching Hospital, Taichung, Taiwan), Pharmacist Chih-Yun Liu (Department of Pharmacy, Asia University Hospital, Taichung, Taiwan), Pharmacist Ching-Jung Wu (Division of Traditional Chinese Medicine, Department of Pharmacy, China Medical University Hospital, Taichung, Taiwan), and Pharmacist Li-Hui Lee (Division of Clinical Pharmacy, Department of Pharmacy, China Medical University Hospital, Taichung, Taiwan). Their time and efforts were provided without compensation.

Data Availability

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

Authors' Contributions

H-YH led the conceptualization, methodology design, data curation, formal analysis, and drafting of the manuscript. L-WC contributed to model implementation and statistical validation. W-TH supervised the technical framework of the study and reviewed the model evaluation. Y-WH contributed to project administration and funding acquisition. SSC provided clinical oversight, interpreted the results from a medical perspective, and critically reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Online Supplements for " Extracting Clinical Guideline Information Using Two Large Language Models: An Evaluation Study Table S1 to Table S6.
[DOCX File (Microsoft Word File), 455 KB-Multimedia Appendix 1]

References

1. Krebs K, Milani L. Translating pharmacogenomics into clinical decisions: do not let the perfect be the enemy of the good. Hum Genomics. Aug 27, 2019;13(1):39. [doi: 10.1186/s40246-019-0229-z] [Medline: 31455423]
2. Relling MV, Klein TE. CPIC: clinical pharmacogenetics implementation consortium of the pharmacogenomics research network. Clin Pharmacol Ther. Mar 2011;89(3):464-467. [doi: 10.1038/clpt.2010.279] [Medline: 21270786]
3. Caraballo PJ, Bielinski SJ, St Sauver JL, Weinshilboum RM. Electronic medical record-integrated pharmacogenomics and related clinical decision support concepts. Clin Pharmacol Ther. Aug 2017;102(2):254-264. [doi: 10.1002/cpt.707] [Medline: 28390138]
4. Caraballo PJ, Hodge LS, Bielinski SJ, et al. Multidisciplinary model to implement pharmacogenomics at the point of care. Genet Med. Apr 2017;19(4):421-429. [doi: 10.1038/gim.2016.120] [Medline: 27657685]
5. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. Mar 30, 2023;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
6. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature New Biol. Aug 2023;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
7. Steimetz E, Minkowitz J, Gabutan EC, et al. Use of artificial intelligence chatbots in interpretation of pathology reports. JAMA Netw Open. May 1, 2024;7(5):e2412767. [doi: 10.1001/jamanetworkopen.2024.12767] [Medline: 38776080]
8. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA. Jan 28, 2025;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]

9. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. Jan 2025;31(1):60-69. [doi: 10.1038/s41591-024-03425-5] [Medline: 39779929]
10. Lin C, Kuo CF. Roles and potential of large language models in healthcare: a comprehensive review. Biomed J. Apr 29, 2025;29(100868):100868. [doi: 10.1016/j.bj.2025.100868] [Medline: 40311872]
11. Neveditsin N, Lingras P, Mago V. Clinical insights: a comprehensive review of language models in medicine. PLOS Digit Health. May 2025;4(5):e0000800. [doi: 10.1371/journal.pdig.0000800] [Medline: 40338967]
12. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-282. [Medline: 23092060]
13. Kamminga NCW, Kievits JEC, Plaisier PW, et al. Do large language model chatbots perform better than established patient information resources in answering patient questions? a comparative study on melanoma. Br J Dermatol. Jan 24, 2025;192(2):306-315. [doi: 10.1093/bjd/ljae377] [Medline: 39365602]
14. Walton N, Gracefo S, Sutherland N, Kozel BA, Danford CJ, McGrath SP. Evaluating ChatGPT as an agent for providing genetic education. bioRxiv. Oct 29, 2023;PMID:38076902. [doi: 10.1101/2023.10.25.564074] [Medline: 38076902]
15. Nazareth S, Hayward L, Simmons E, et al. Hereditary cancer risk using a genetic chatbot before routine care visits. Obstet Gynecol. Dec 1, 2021;138(6):860-870. [doi: 10.1097/AOG.0000000000004596] [Medline: 34735417]
16. Siglen E, Vetti HH, Lunde ABF, et al. Ask rosa - the making of a digital genetic conversation tool, a chatbot, about hereditary breast and ovarian cancer. Patient Educ Couns. Jun 2022;105(6):1488-1494. [doi: 10.1016/j.pec.2021.09.027] [Medline: 34649750]
17. Skryabin V, Rozochkin I, Zastrozhin M, et al. Meta-analysis of pharmacogenetic clinical decision support systems for the treatment of major depressive disorder. Pharmacogenomics J. May 2023;23(2-3):45-49. [doi: 10.1038/s41397-022-00295-3] [Medline: 36273107]
18. Jarvis JP, Peter AP, Keogh M, et al. Real-world impact of a pharmacogenomics-enriched comprehensive medication management program. J Pers Med. Mar 8, 2022;12(3):421. [doi: 10.3390/jpm12030421] [Medline: 35330421]
19. Hinderer M, Boeker M, Wagner SA, et al. Integrating clinical decision support systems for pharmacogenomic testing into clinical routine - a scoping review of designs of user-system interactions in recent system development. BMC Med Inform Decis Mak. Jun 6, 2017;17(1):81. [doi: 10.1186/s12911-017-0480-y] [Medline: 28587608]
20. Menz BD, Modi ND, Abuhelwa AY, et al. Generative AI chatbots for reliable cancer information: evaluating web-search, multilingual, and reference capabilities of emerging large language models. Eur J Cancer. Mar 11, 2025;218:115274. [doi: 10.1016/j.ejca.2025.115274] [Medline: 39922126]
21. Adeosun SO, Faibille AB, Qadir AN, Mutwol JT, McMannen T. A deep neural network model for classifying pharmacy practice publications into research domains. Res Social Adm Pharm. Feb 2025;21(2):85-93. [doi: 10.1016/j.sapharm.2024.10.009] [Medline: 39523144]
22. Şahin MF, Doğan Ç, Topkaç EC, Şeramet S, Tuncer FB, Yazıcı CM. Which current chatbot is more competent in urological theoretical knowledge? a comparative analysis by the European board of urology in-service assessment. World J Urol. Feb 11, 2025;43(1):116. [doi: 10.1007/s00345-025-05499-3] [Medline: 39932577]
23. Nicikowski J, Szczepański M, Miedziaszczyk M, Kudliński B. The potential of ChatGPT in medicine: an example analysis of nephrology specialty exams in Poland. Clin Kidney J. Aug 2024;17(8):sfae193. [doi: 10.1093/ckj/sfae193] [Medline: 39099569]
24. Jaramillo AC, Hubeek I, Broekhuizen R, et al. Expression of the nucleoside transporters hENT1 (SLC29) and hCNT1 (SLC28) in pediatric acute myeloid leukemia. Nucleosides Nucleotides Nucleic Acids. 2020;39(10-12):1379-1388. [doi: 10.1080/15257770.2020.1746803] [Medline: 32312148]
25. Lu Y, Aleta A, Du C, Shi L, Moreno Y. LLMs and generative agent-based models for complex systems research. Phys Life Rev. Dec 2024;51(283-93):283-293. [doi: 10.1016/j.plrev.2024.10.013] [Medline: 39486377]
26. Huang J, Yang DM, Rong R, et al. A critical assessment of using ChatGPT for extracting structured data from clinical notes. NPJ Digit Med. May 1, 2024;7(1):106. [doi: 10.1038/s41746-024-01079-8] [Medline: 38693429]
27. Garcia P, Ma SP, Shah S, et al. Artificial intelligence-generated draft replies to patient inbox messages. JAMA Netw Open. Mar 4, 2024;7(3):e243201. [doi: 10.1001/jamanetworkopen.2024.3201] [Medline: 38506805]

Abbreviations

AI: artificial intelligence
API: application programming interface
CDSS: clinical decision support system
LLM: large language model
PGx: pharmacogenomic
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

Edited by Andrew Coristine; peer-reviewed by Ki-Seong Park, Mohan Krishna Ghanta, Zhen Hou; submitted 05.03.2025; final revised version received 16.06.2025; accepted 16.06.2025; published 05.09.2025

Please cite as:
Hsu HY, Chen LW, Hsu WT, Hsieh YW, Chang SS
Extracting Clinical Guideline Information Using Two Large Language Models: Evaluation Study
J Med Internet Res 2025;27:e73486
URL: https://www.jmir.org/2025/1/e73486
doi: 10.2196/73486