Computers in Biology and Medicine 200 (2026) 111323

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Hierarchical RAG enhances a pharmacogenomic AI assistant in guideline related queries

Yaejin Jeon [a,1], Mi Seon Youn [a,1], Sunghoon Kang [a], Jonghyung Park [b], Eun Sil Kim [b], Juyoung Kim [b], Ju Han Kim [a,*]

[a] Seoul National University Biomedical Informatics (SNUBI), Div. of Biomedical Informatics, Seoul National University College of Medicine, Seoul, 03080, Republic of Korea
[b] Meninblox Inc., Gwangju, 61008, Republic of Korea

ARTICLE INFO

Dataset link: https://cpicpgx.org/, https://www.knmp.nl/dossiers/farmacogenetica, https://github.com/KarlKeat/PGxQA.git

Keywords:
Pharmacogenomics
Large language models
Hierarchical retrieval-augmented generation
AI assistant

ABSTRACT

Background: Implementing pharmacogenomics (PGx) in the healthcare system faces several challenges. These include limited education among healthcare professionals and restricted patient awareness of the benefits of genetic testing. Generative AI offers a promising solution by generating tailored content. It also provides interactive clinical decision support to help bridge the knowledge gap.
Methods: This study introduces a Hierarchical Retrieval-Augmented Generation (HRAG) framework for anti-cancer drugs (5-fluorouracil, capecitabine, and tamoxifen) to reflect the relationships between PGx documents. HRAG organizes the PGx guidelines into hierarchical tree structures, treating each guideline at the same level. Documents were chunked into passages, which were represented as leaf nodes. We evaluated the model using PGxQA dataset, based on RAGAS scores and human evaluation.
Results: HRAG and RAG models showed lower performance in PGxQA categories requiring precise numerical or entity matching, such as allele definition. They performed better in categories focused on textual reasoning, such as phenotype-to-guideline tasks. In contrast, HRAG significantly outperformed RAG in guideline-related tasks, demonstrating higher context precision, context recall, and accuracy (F1). Specifically, in the phenotype-to-guideline category, HRAG achieved an F1 score of 0.89, while RAG scored 0.80 (p-value $= 9 \times 10^{-4}$).
Conclusions: These findings suggest that the HRAG framework could contribute to the development of a PGx AI assistant. It may help narrow the knowledge gap and facilitate the broader adoption of PGx.

1. Introduction

Pharmacogenomics (PGx) is part of precision medicine, which examines how genetic factors affect the way individuals response to medications [1]. This field aims to enhance treatment efficacy and reduce adverse drug reactions (ADR). As ADR imposed a significant burden on public health [2,3], the importance of personalized medicine and PGx has been increasingly recognized.

Clinical Pharmacogenetics Implementation Consortium (CPIC) [4] was established by the National Institute of Health's Pharmacogenomics Research Network [5] and the Pharmacogenomics Knowledge Base (PharmGKB) [6]. It provides peer-reviewed and freely accessible evidence-based guidelines to facilitate the adoption of PGx in clinical practice. Similarly, the Dutch Pharmacogenetics Working Group (DPWG) [7,8], founded by the Royal Dutch Pharmacists Association (KNMP) in 2005, issues PGx recommendations to improve drug safety and efficacy. However, the integration of genetic and clinical knowledge in PGx presents challenges for healthcare professionals to apply these guidelines in practice.

Artificial intelligence (AI) has recently emerged as a promising tool in medicine. Generative AI, in particular, has been applied to medical education and clinical decision support [9,10]. In education, models are increasingly used for self-directed learning, simulation-based training, and writing assistance, though concerns remain regarding hallucinations and factual accuracy [11]. Smart healthcare, proposed by Tian et al. [12], emphasizes the dynamic and intelligent flow of medical information to assist decision-making. For example, NLP-based system such as GatorTron was trained to interpret and analyze electronic health records to support diagnostic reasoning [13]. These advances

---

[*] Corresponding author.
E-mail address: juhan@snu.ac.kr (J.H. Kim).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.compbiomed.2025.111323
Received 10 June 2025; Received in revised form 13 November 2025; Accepted 16 November 2025
0010-4825/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

demonstrate the potential of AI to improve the flow of information in multidisciplinary fields such as PGx.

Despite continuous efforts from initiatives like CPIC and DPWG, the adoption of PGx in clinical practice has been delayed. Barriers include insufficient education for healthcare professionals and the complexity of interpreting genomics data [14,15]. Recent efforts such as PGx4Statins [16] explored the use of retrieval-augmented generation (RAG) and prompt engineering to support PGx-informed decision making. However, it relies on retrieving semantically related chunks from documents, without reflecting the relationships between different guidelines. This can lead to imbalanced information when relevant chunks come from a single guideline, especially if there are conflicts between two guidelines. These limitations underscore a critical gap in the ability to provide recommendations based on the full spectrum of PGx guidelines.

These limitations motivate the present study, which aims to address the adoption of PGx into clinical practice. Our objective is to develop an PGx assistant that can provide real-time decision and educational support. This research aims to make the PGx guidelines more accessible, actionable, and efficient for clinicians, contributing to the broader adoption of personalized medicine.

In this study, we introduced a Hierarchical Retrieval-Augmented Generation (HRAG) framework designed to capture semantic structures and enhance information retrieval. This is a novel approach that reflect the structure and relationships among PGx documents. Contributions of this work are as follows: First, we used a structured evaluation approach for assessing the effectiveness of large language models (LLMs) in PGx application [17]. Second, we systematically evaluated the AI's performance, revealing variations across different query types. Additionally, this research contributes to the application of AI in medical education and clinical decision support. Finally, we demonstrate the benefits of using hierarchical document structures in supporting interpretation and application of PGx guidelines.

2. Method

2.1. Data source

The datasets were obtained from CPIC and DPWG, two widely recognized and rigorously curated PGx knowledge bases. Free-text contextual sources included CPIC guidelines [18,19] and supplements, as well as DPWG recommendations [20]. To integrate tabular information from the guidelines into the RAG framework, we converted tables into descriptive free-text formats compatible with unstructured data. Additionally, structured datasets such as the CPIC diplotype–phenotype translation tables, allele definition tables, functionality tables, and population-specific frequency tables [21,22], were transformed into descriptive text to capture detailed PGx nuances. To investigate the performance of HRAG in the context of PGx4Statins, we evaluated its responses using statin PGx documents from the PGx4Statins dataset. Whereas anticancer drugs have PGx guidelines for a single gene, statin guidelines involve three genes: SLCO1B1, ABCG2, and CYP2C9. The dataset consisted of CPIC, DPWG guidelines, RNPGX, FDA labels, and diplotype-to-phenotype mapping CSV files. These detailed datasets are available on GitHub [23].

To evaluate model performance, we used the PGxQA [24,25] dataset. PGxQA is a publicly available PGx queries with reference answers developed to address the lack of evaluation resources in PGx chatbot. Unlike previous study, which could only be tested on a limited number of questions for a single drug, PGxQA enables a large-scale evaluation. The questions were constructed based on CPIC Level A guidelines that are considered to have strong clinical significance. They cover 10 categories, including translating genotypes into phenotypes, identifying relevant dbSNP variants, and deriving clinical recommendations. The question set was created through both automated methods, using the psycopg2 package to query CPIC's PostgreSQL database, and manual curation by experts.

In this study, we used categories from the PGxQA dataset, including Allele definition, Allele frequency, Allele function, Diplotype to phenotype, Phenotype to guideline, Drugs to genes, Genes to drugs, and Phenotype to category (Supplementary Table A.1). We filtered the PGxQA dataset to include questions related to the genes CYP2D6 and DPYD, and the drugs fluorouracil, capecitabine, and tamoxifen, resulting in a total of 170 questions. We also assessed questions related to the genes SLCO1B1, ABCG2, and CYP2C9, and the drugs atorvastatin, fluvastatin, lovastatin, pitavastatin, pravastatin, rosuvastatin, and simvastatin, yielding 259 questions. We excluded questions categorized as Adversarial (refusal) because our primary objective was not to evaluate the model's ability to detect contradictory queries. Additionally, we excluded questions categorized as External, as they did not contain queries related to our target drugs, and were therefore outside the scope of our current evaluation. Accordingly, our evaluation focused on verifying the retrieval accuracy and reliability of the HRAG and RAG model.

2.2. Hierarchical RAG

We developed the HRAG framework, which leverages a hierarchically organized structure of documents represented as trees (Fig. 1). To construct document trees for retrieval, we used the llama-index Python package [26].

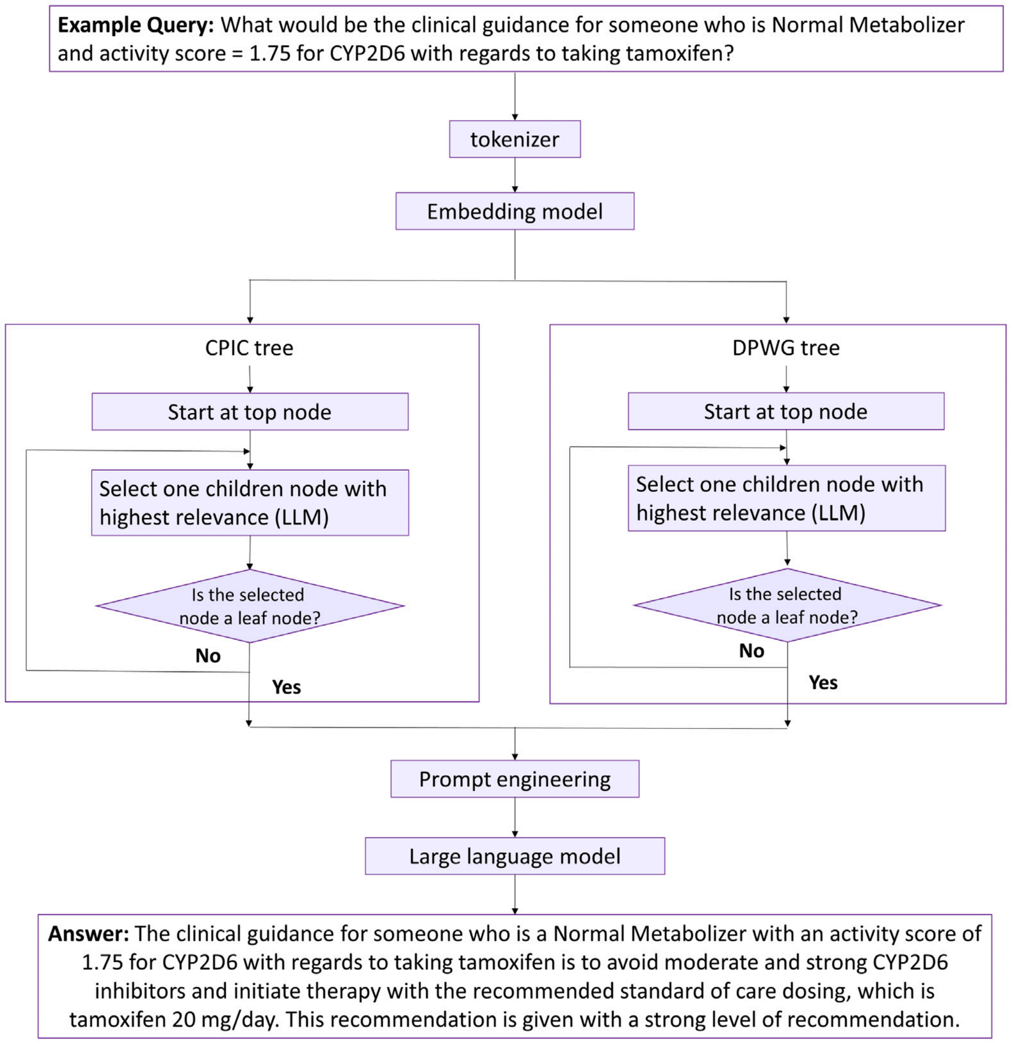
We constructed the CPIC and DPWG trees independently at the same hierarchical level because each guideline was developed by an independent consortium. This design prevents bias toward one guideline. During preprocessing, the CPIC and DPWG files were manually labeled according to the consortium names in their filenames (e.g., CPIC or DPWG). The labeled files were then grouped accordingly, and each group was constructed into a tree structure. Documents were chunked into passages with a chunk size of 1000, consistent with prior work (PGx4Statin). Each passage was then assigned to the leaf nodes of its respective tree. Higher-level nodes were then created by summarizing the content of the leaf nodes.

Each tree has its own root node, and no parent–child relationship exists between them. During retrieval, HRAG searches both trees in parallel and integrates the results. Within each tree, the model recursively identifies the most semantically similar child node, ultimately retrieving the leaf node as the final outcome. We utilized the GPT-4o model with zero temperature to generate responses to user queries. The passages were embedded by OpenAI's text-embedding-ada-002 model [27].

2.3. Prompt engineering

To enable a fair and unbiased comparison between HRAG and RAG, we minimized the prompt for evaluation (Fig. 2a). The prompt contained only essential instructions: to answer a given question using the provided context and to explicitly state "UNKNOWN" if the answer could not be inferred from the given content. This design was intended to reduce hallucinated responses and enforce strict adherence to evidence-based reasoning.

In addition, to compare the responses of HRAG and RAG in practical clinical scenario, we designed another prompt (Fig. 2b). It was designed to reflect situations in which healthcare professionals utilize PGx information in decision-making. The prompt included a role description (AI trained for PGx support), clear objectives (e.g., assist clinical decision making with accurate and concise information) and constraints (e.g., do not infer beyond the given context, distinguish between CPIC and DPWG when applicable). It was also instructed to acknowledge uncertainty when information was insufficient to make a decision, using phrases such as "Insufficient information to provide a specific response". In cases where CPIC and DPWG guidelines provided



Example Query: What would be the clinical guidance for someone who is Normal Metabolizer and activity score = 1.75 for CYP2D6 with regards to taking tamoxifen?

tokenizer
Embedding model
CPIC tree
Start at top node
Select one children node with highest relevance (LLM)
Is the selected node a leaf node?
No
Yes
DPWG tree
Start at top node
Select one children node with highest relevance (LLM)
Is the selected node a leaf node?
No
Yes
Prompt engineering
Large language model

Answer: The clinical guidance for someone who is a Normal Metabolizer with an activity score of 1.75 for CYP2D6 with regards to taking tamoxifen is to avoid moderate and strong CYP2D6 inhibitors and initiate therapy with the recommended standard of care dosing, which is tamoxifen 20 mg/day. This recommendation is given with a strong level of recommendation.

Fig. 1. Flowchart of Hierarchical Retrieval-Augmented Generation (HRAG) Framework for a Pharmacogenomics (PGx) AI Assistant.

conflicting recommendations, the prompt was engineered to provide the LLM with both sets of guideline information. This reflected real-world clinical practice where final decisions are made by a qualified clinician. We compared their responses to a sample question: “A patient with a CYP2D6 *4/4 is being prescribed tamoxifen for breast cancer treatment. What is the recommended course of action for tamoxifen dosage or alternative therapy?”

2.4. *Evaluation*

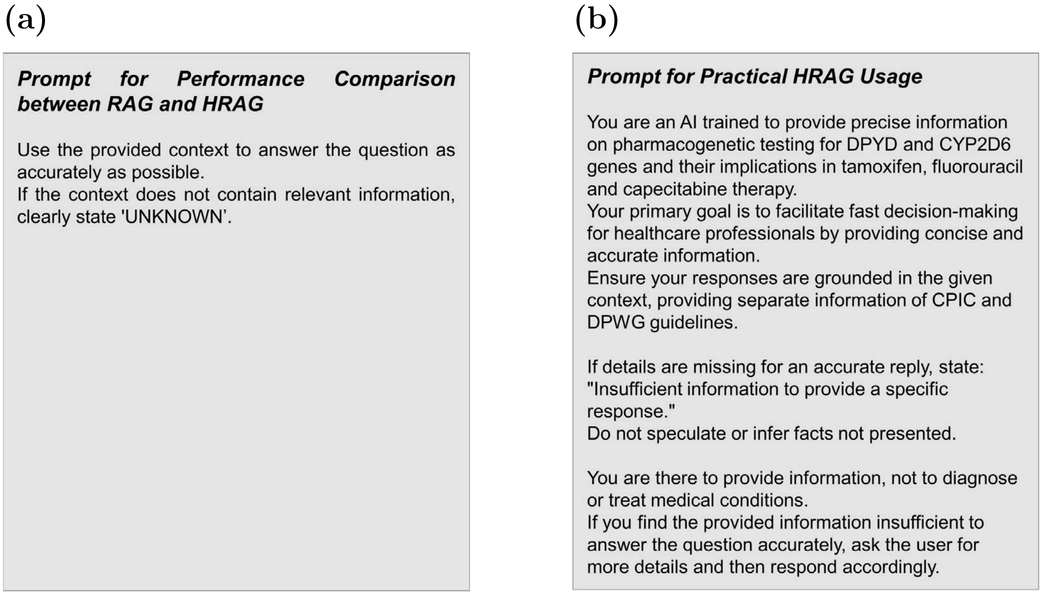
To compare HRAG with RAG (PGx4Statins), we calculated the differences in evaluation scores for each PGxQA query. Wilcoxon signed-rank test was conducted to examine whether the distribution of differences was symmetrically centered around zero. Statistical analyses were performed using the SciPy (version 1.15.2) [28]. The test yielded the Wilcoxon statistic and the corresponding p-values. In addition, Cohen’s d was calculated to quantify the effect size and the practical magnitude of the observed differences. However, in some cases, p-values or Cohen’s d could not be computed because the Wilcoxon statistic was not available under those conditions. The Wilcoxon signed-rank test requires at least two valid pairs, and when this requirement was not met, the test statistic could not be calculated (Supplementary Table A.3).

2.4.1. *Automated evaluation metrics*

For evaluation, we used an RAG-specific evaluation framework, RAGAS [29]. RAGAS allows us to assess not only the generated responses but also the retrieved contexts. We used following evaluation metrics: context precision, context recall, context entity recall, noise sensitivity, response relevancy, and faithfulness(Table 1). The metrics are computed by prompting an LLM, and the detailed prompts used for these evaluations can be found in the original RAGAS publication.

For example, given a query *‘I want to give my patient tamoxifen. What genes should I include in a pharmacogenetics panel?’*, HRAG generated response *CYP2D6*. The ground-truth reference answer was also *CYP2D6*. According to the equations summarized in Table 1, the RAGAS scores were computed as follows: Context Precision = 1, Context Recall = 1, Context Entity Recall = 1, Noise Sensitivity = 0, Answer Relevancy = 0.75, and Faithfulness = 1.

Furthermore, we used F1 and BERTScore, which are widely used metrics for natural language processing (NLP) tasks, to further quantify the precision and recall of the generated responses. The F1 score was used to assess the responses of multiple choice questions. BERTScore [30], which can measure semantic similarity between generated responses and references, was used to evaluate short-answer questions.



(a)

Prompt for Performance Comparison between RAG and HRAG

Use the provided context to answer the question as accurately as possible.
If the context does not contain relevant information, clearly state 'UNKNOWN'.

(b)

Prompt for Practical HRAG Usage

You are an AI trained to provide precise information on pharmacogenetic testing for DPYD and CYP2D6 genes and their implications in tamoxifen, fluorouracil and capecitabine therapy.
Your primary goal is to facilitate fast decision-making for healthcare professionals by providing concise and accurate information.
Ensure your responses are grounded in the given context, providing separate information of CPIC and DPWG guidelines.

If details are missing for an accurate reply, state:
"Insufficient information to provide a specific response."
Do not speculate or infer facts not presented.

You are there to provide information, not to diagnose or treat medical conditions.
If you find the provided information insufficient to answer the question accurately, ask the user for more details and then respond accordingly.

Fig. 2. Two distinct prompts were used for different purposes: (a) A prompt for performance evaluation between HRAG and RAG, and (b) A prompt for practical usage. Prompt (a) was designed in a minimalistic manner to directly compare the inherent performance of the two methods, whereas prompt (b) was constructed with domain-specific instructions to simulate practical usage scenarios.

Table 1
Evaluation with RAGAS framework.

Score	Measurement	Equation
Context Precision	Measures a proportion of retrieved contexts that are relevant to the reference.	$Context\ Precision@K = \frac{\sum_{k=1}^{K}(\text{Precision}@k \times v_k)}{\text{Total number of relevant items in top } K}$ $Precision@k = \frac{\text{true positives}@k}{\text{true positives}@k+\text{false positives}@k}$
Context Recall	Measures a extent to which all relevant information was successfully retrieved.	$Context\ Recall = \frac{\text{Number of claims in the reference supported by the retrieved context}}{\text{Total number of claims in the reference}}$
Context Entity Recall	Evaluates whether key entities found in the reference are also present in the retrieved context.	$Context\ Entity\ Recall = \frac{\text{Number of common entities between RCE and RE}}{\text{Total number of entities in RE}}$
Noise Sensitivity	Assesses the model's tendency to produce incorrect answers when exposed to either relevant or irrelevant retrieved content.	$Noise\ Sensitivity = \frac{\text{Total number of incorrect claims in response}}{\text{Total number of claims in the response}}$
Response Relevancy	Determines how closely the generated response corresponds to the user query.	$Answer\ Relevancy = \frac{1}{N}\sum_{i=1}^{N} \text{cosine similarity}(E_{g_i}, E_o)$
Faithfulness	Assesses whether the model's response stays true to the retrieved evidence.	$Faithfulness = \frac{\text{Number of claims in the response supported by the retrieved context}}{\text{Total number of claims in the response}}$

Note: Item = retrieved text chunks, Claim = verifiable factual statement, Entity = named entities extracted from text, RCE = The set of entities in the retrieved contexts, RE = The set of entities in the reference, E_{g_i} = Embedding of the ith generated question, Eo = Embedding of the user input, N = Number of generated questions based on the response(default = 3).

2.4.2. Human evaluation

Human evaluation was conducted by two PGx experts (Y.J., M.S.Y.), both with research experience in PGx. A total of 170 question–context–answer triplets were evaluated based on six criteria: Accuracy, Relevancy, Completeness, Hallucination mitigation, RAG accuracy, and RAG relevancy. Each of the six criteria was scored using a four-point Likert scale (1–4), where 4 indicated 'strongly agree', 3 'agree', 2 'disagree', and 1 'strongly disagree'. This approach eliminates the neutral midpoint (e.g., "3" in a five-point scale), thereby preventing ambiguous or neutral responses and ensuring a binary evaluation [31]. The final score for each response was computed as the average of both experts' ratings. Among the six criteria, Accuracy, Relevancy, Completeness, and Hallucination Mitigation were used to evaluate the final generated responses. In contrast, RAG Accuracy and RAG Relevancy were designed to assess the retrieval performance of each method.

- **Accuracy**: Measures how well the generated responses aligned with the reference, ensuring clinical reliability.
- **Relevancy**: Evaluates the appropriateness and directness of the response in addressing the given question.
- **Completeness**: Whether all essential components were included in the response.
- **Hallucination mitigation**: Whether the model avoided generating fabricated, misleading information.
- **RAG accuracy**: Measured how closely the retrieved context matched the reference.
- **RAG relevancy**: Whether the retrieved context was appropriately aligned with the input query and presented in a suitable format.

2.5. Performance in statins

To ensure comparability, we evaluated HRAG and RAG (PGx4Statin) using the Statin guidelines employed in the previous PGx4Statin study (Supplementary Figure A.1). The evaluation procedure was the same as that used for the anticancer drugs. In addition, to examine the impact of prompt design, we generated responses using both the prompt (a) in Fig. 2 and PGx4Statin prompt, and compared the results. This setup resulted in four distinct RAG-prompt combinations: prompt(a) with the RAG, prompt(a) with the HRAG, PGx4Statin prompt with the RAG, PGx4Statin prompt with HRAG. To ensure consistency and reduce variability in response generation, we used GPT-4o as the language model with a temperature setting of 0. By evaluating both RAG and HRAG across all prompt conditions, we were able to disentangle the influence of prompt design and attribute the remaining performance differences to the retrieval method itself.

The inclusion of prompt comparisons led to a fourfold increase in the number of evaluation items. To efficiently manage the human evaluation process while maintaining consistency and objectivity, we adopted stratified sampling. To ensure unbiased assessment, each of the four subsets was anonymized before evaluation. From each anonymized subset, we randomly extracted 10% of the questions to form validation sets. Excluding the validation set, the remaining questions were divided using category-based stratified sampling. Two experts independently evaluated the responses, and to verify consistency between evaluators, we calculated the inter-rater agreement using Cohen's kappa coefficient. The agreement rate confirmed reliable consistency between the two evaluators. After individual assessments, we combined the results from both experts to create a unified evaluation dataset.

3. Results

3.1. Structure of tree index

Since genetic information is exclusively available in CPIC, the CPIC tree was constructed with two subtrees: a genetic tree and a guideline tree. The genetic tree contains documents related to allele definition, allele function, allele frequency, and diplotype-to-phenotype provided by CPIC. In contrast, the guideline tree includes the official CPIC clinical guideline PDFs. Specifically, the guideline tree consists of two root nodes and reaches a maximum depth of three, whereas the genetic tree has three root nodes with a maximum depth of four (depth is defined as the length of the longest path from the root to a leaf node). As a result, the HRAG framework searches the most relevant passage from both the guideline and genetic information subtrees.

On the other hand, the DPWG tree includes only the guideline-related hierarchy, as the DPWG database does not provide structured genetic information. It is organized under five root nodes: three corresponding to fluorouracil and capecitabine, and two to tamoxifen. Each root node reaches a maximum depth of two, with varying numbers of child nodes. These structural characteristics, including the number of roots and the maximum depth, were determined by the default settings of the llama-index according to the volume and organization of the source documents.

3.2. Qualitative evaluation of responses from HRAG and RAG

We compared the RAG and HRAG approaches in practical usage by testing with a sample question (Fig. 3). Both methods could recommend alternative medications for managing the risk of adverse effects associated with tamoxifen prescriptions; however, HRAG more accurately reflected the guideline recommendations by not only suggesting alternative drugs but also providing detailed dosage adjustment information when alternatives are unsuitable. This represents a key advantage of HRAG over RAG. Furthermore, while both models generated summaries based on CPIC guidelines and DPWG recommendations, HRAG clearly separated and presented the content of individual documents, thereby improving the overall readability of the information.

3.3. Performance evaluation

3.3.1. Hierarchical RAG outperforms in guideline-related queries

The complete set of anticancer drug-related PGxQA queries, retrieved contexts, and generated responses from HRAG and RAG is provided in Supplementary Table A.4. This table serves as the primary dataset for automated evaluation and practical evidence supporting the reported outcomes. For each PGxQA category, we calculated the average evaluation score and visualized the results in Fig. 4.

HRAG consistently outperformed RAG in evaluation metrics, including Context precision, Context recall, Context Entity Recall, Response relevancy, Faithfulness, and F1 score, in the four categories related to guidelines; Phenotype to guideline, Drugs to genes, Genes to drugs, and Phenotype to category (Fig. 4, (e)–(h)). For example, F1 score for Phenotype to guideline was 0.89 in HRAG compared to 0.80 in RAG, indicating a substantial improvement that was statistically significant ($p < 0.001$, Fig. 4, (e)). In addition, HRAG also achieved higher scores in Context precision, Context recall across Phenotype to guideline and Drugs to genes categories ($p < 0.05$, Fig. 4, (e)–(f)). This results indicate that HRAG retrieved more relevant documents and fewer relevant documents were left out.

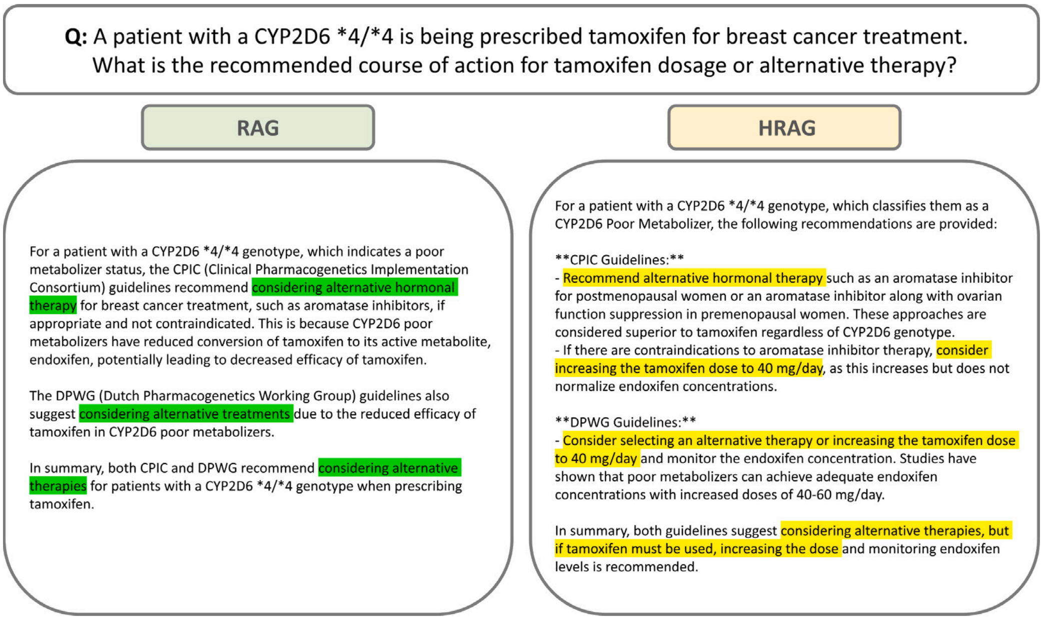
Categories related to genetic information retrieval, including Allele definition, Allele frequency, Allele function, and Diplotype to phenotype, showed lower performance for both models (Fig. 4, (a)–(d)). Overall, RAG outperformed HRAG in genotype-related categories. For instance, in the Allele frequency category, RAG achieved F1 score of 0.19, significantly outperforming HRAG, which scored 0.04 ($p < 0.001$, Fig. 4, (b)). These categories require precise matching of identifiers or numerical values (e.g., rsIDs, allele frequency), which are challenging for LLMs due to their limited numerical reasoning and difficulty in exact matching of numbers.

3.3.2. Human evaluation results

Human evaluation showed that HRAG outperformed RAG across guideline-related categories such as Phenotype to guideline, Drugs to genes, Genes to Drugs, and Phenotype to category. In the Phenotype to guideline category, for instance, HRAG achieved an Accuracy score of 3.917, significantly higher than the 1.250 scored by RAG ($p < 0.001$, Fig. 5, (e)). In addition to Accuracy, HRAG also showed statistically significant improvements across all other evaluation metrics in this category, including Relevancy, Completeness, Hallucination mitigation, RAG accuracy, and RAG relevancy. This suggests that HRAG retrieved context that is factually consistent with reference materials. It also delivered responses that were better aligned with the user's intent, enhancing both the reliability and clinical utility of its outputs.

A similar trend was observed in the Drug to genes category, where HRAG received a perfect score of 4.0, significantly outperforming RAG's 1.83 ($p < 0.001$, Fig. 5, (f)). According to the PGxQA benchmark description (Supplementary Table A.1), this category explicitly evaluates a model's ability to infer actionable gene targets for a given drug based on CPIC and DPWG guidelines. Therefore, the superior performance of HRAG in this task demonstrates its enhanced capacity to captures pharmacogenomic relationships between drugs and genes. Notably, HRAG achieved this perfect score not only in Accuracy, but also in Relevancy, Completeness, and Hallucination mitigation, reflecting its consistent performance across multiple evaluation criteria. This finding aligns with the automated evaluation results (Fig. 4, (f)).

In contrast, for genotype-related questions (Allele definition, Allele frequency, Allele function, and Diplotype to phenotype), RAG showed better scores (Fig. 5,(a)–(d)). Consistent with the automated evaluation results, the human evaluation also showed lower performance for both models in these categories, highlighting the difficulty of precise matching of numeric values. This suggests that guideline-related content may benefit more from HRAG's architecture than genotype-level queries.



Q: A patient with a CYP2D6 *4/*4 is being prescribed tamoxifen for breast cancer treatment. What is the recommended course of action for tamoxifen dosage or alternative therapy?

RAG

For a patient with a CYP2D6 *4/*4 genotype, which indicates a poor metabolizer status, the CPIC (Clinical Pharmacogenetics Implementation Consortium) guidelines recommend considering alternative hormonal therapy for breast cancer treatment, such as aromatase inhibitors, if appropriate and not contraindicated. This is because CYP2D6 poor metabolizers have reduced conversion of tamoxifen to its active metabolite, endoxifen, potentially leading to decreased efficacy of tamoxifen.

The DPWG (Dutch Pharmacogenetics Working Group) guidelines also suggest considering alternative treatments due to the reduced efficacy of tamoxifen in CYP2D6 poor metabolizers.

In summary, both CPIC and DPWG recommend considering alternative therapies for patients with a CYP2D6 *4/*4 genotype when prescribing tamoxifen.

HRAG

For a patient with a CYP2D6 *4/*4 genotype, which classifies them as a CYP2D6 Poor Metabolizer, the following recommendations are provided:

CPIC Guidelines:
- Recommend alternative hormonal therapy such as an aromatase inhibitor for postmenopausal women or an aromatase inhibitor along with ovarian function suppression in premenopausal women. These approaches are considered superior to tamoxifen regardless of CYP2D6 genotype.
- If there are contraindications to aromatase inhibitor therapy, consider increasing the tamoxifen dose to 40 mg/day, as this increases but does not normalize endoxifen concentrations.

DPWG Guidelines:
- Consider selecting an alternative therapy or increasing the tamoxifen dose to 40 mg/day and monitor the endoxifen concentration. Studies have shown that poor metabolizers can achieve adequate endoxifen concentrations with increased doses of 40-60 mg/day.

In summary, both guidelines suggest considering alternative therapies, but if tamoxifen must be used, increasing the dose and monitoring endoxifen levels is recommended.

Fig. 3. Comparison of responses from RAG and HRAG using the practical usage prompt. Both models were asked to recommend an appropriate course of action for a CYP2D6 *4/*4 patient prescribed tamoxifen. The highlighted phrases in each response indicate actionable recommendations.

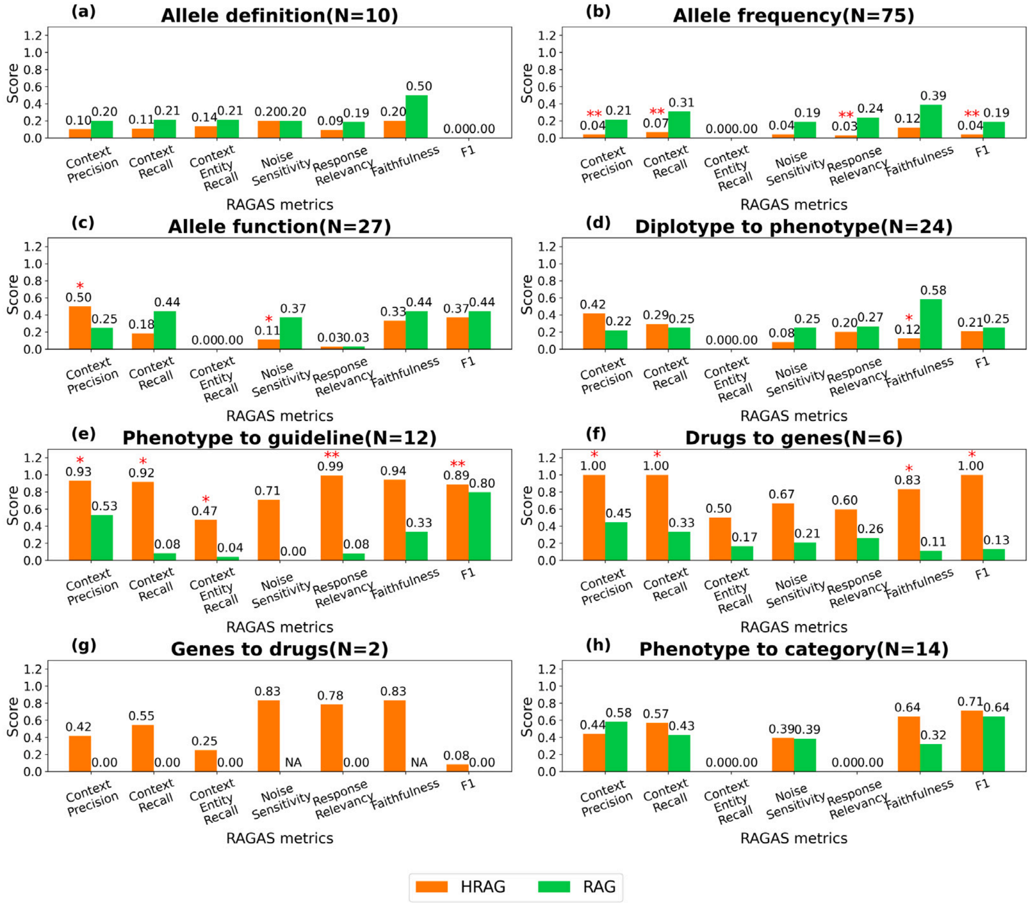


Fig. 4. Performance comparison of HRAG and RAG based on automated evaluation metrics, RAGAS. (a)–(h) subplots correspond to a PGxQA question category(N = the number of queries). Metrics include Context precision/recall/entity recall, Noise sensitivity, Response relevancy, Faithfulness, and F1 or BERTScore. For multiple-choice questions, the F1 score was used, while BERTScore was applied to short-answer questions, such as the 'Phenotype to guideline' category. The numeric labels above the bars represent the score values for each RAGAS metric. A single asterisk (*) indicates statistical significance at $p < 0.05$, and a double asterisk (**) indicates significance at $p < 0.001$.

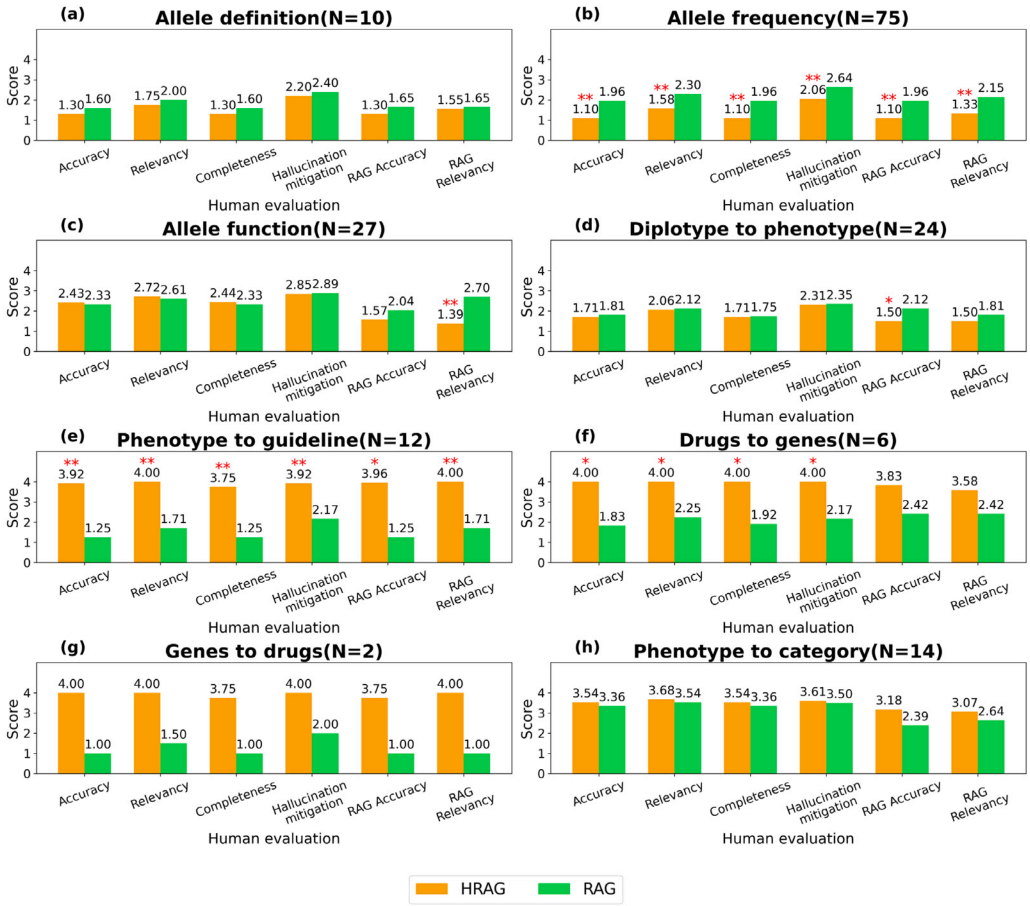


Fig. 5. Performance comparison of HRAG and RAG based on human evaluation metrics. (a)–(h) subplots represent a PGxQA question category (N = the number of queries). Six evaluation criteria were rated by two PGx experts on a 4-point Likert scale (1 = strongly disagree, 4 = strongly agree): Accuracy, Relevancy, Completeness, Hallucination mitigation, RAG Accuracy, and RAG Relevancy. The numeric labels above the bars represent the score values for each evaluation metric. A single asterisk (*) indicates statistical significance at $p < 0.05$, and a double asterisk (**) indicates significance at $p < 0.001$.

Table 2
Performance comparison of HRAG and PGx4Statins on Statin. This table presents the mean scores from automated evaluation based on RAGAS metrics. HRAG_1: HRAG with prompt(a), HRAG_2: HRAG with PGx4Statin prompt, PGx4Statin_1: PGx4Statin with prompt(a), PGx4Statin_2: PGx4Statin with PGx4Statin prompt.

Model	Context precision	Context recall	Context entity recall	Noise sensitivity	Answer relevancy	Faithfulness	F1
HRAG_1	0.38	0.39	0.20	0.19	0.23	0.35	0.40
HRAG_2	0.35	0.41	0.18	0.08	0.11	0.30	0.23
PGx4Statin_1	0.20	0.30	0.05	0.29	0.10	0.37	0.29
PGx4Statin_2	0.23	0.29	0.05	0.20	0.13	0.31	0.24

3.3.3. Performance comparison on statin

We compared the PGx4Statin and HRAG on statin using identical datasets and evaluation criteria (Supplementary Figure A.1). The average performance across RAGAS metrics for statin-related PGxQA queries showed that HRAG with prompt (a) achieved the highest performance in context precision, context entity recall, answer relevancy, and F1 score (Table 2). When the mean scores were analyzed by category, the performance of both HRAG and PGx4Statin showed overall low accuracies in the Allele Definition, Allele Frequency, Allele Function, and Diplotype to Phenotype categories (Supplementary Figure A.2 and A.3). These results suggest that the inclusion of numerical nomenclature in star-allele (e.g., CYP2C19*2, *3) posed inherent challenges for both PGx4Statin and HRAG. Additionally, PGx4Statin's data scope contributed to this results, as it included diplotype-to-phenotype mapping data but lacked other genetic data sources, such as allele frequency and allele definition.

In the Drug to Gene and Gene to Drug categories, HRAG demonstrated a significant advantage ($p < 0.001$), underscoring the strength of its hierarchical retrieval in inferring drug–gene relationships (Supplementary Figure A.2 and A.3). Whereas PGx4Statin retrieved diplotype-related documents, HRAG retrieved guideline-related chunks through summary-based searches. Therefore, HRAG was able to capture bidirectional relationships between drugs and genetic variants, highlighting its versatility across PGx contexts.

In the Phenotype to guideline and Phenotype to category tasks, PGx4Statin outperformed HRAG in specific evaluation metrics, such as context precision ($p < 0.05$, Supplementary Figure A.2 and A.3). This was attributed to HRAG's difficulty in accurately structuring tables extracted from PDFs, as confirmed by additional evaluation using preprocessed tables(Supplementary Table A.2). Notably, in the phenotype to category task, a significantly low HRAG_2 score of Noise Sensitivity was observed. This difference was primarily due to the exceptionally low performance of HRAG with the PGx4Statin prompt, suggesting that the choice of template significantly influenced the results in this specific case.

4. Discussion

In this study, we introduced a HRAG framework specialized in PGx and applied it to three anticancer drugs which have PGx guidelines; 5-fluorouracil, capecitabine, and tamoxifen. We integrated the HRAG with an LLM to construct a PGx AI assistant capable of handling CPIC and DPWG documents on the same level of importance and granularity. As a result, HRAG outperformed the previous RAG framework in cases where both guideline and genomic information are involved. This improvement led to more accurate and clinically relevant PGx support in the guideline-related queries.

HRAG framework is the first to apply a hierarchical retrieval framework specifically to the PGx domain. Prior studies in the text summarization field proposed hierarchical encoding frameworks, which captured both horizontal relationships among tokens and vertical relationships across different levels of textual granularity [32,33]. HiRAG applied a hierarchical RAG framework to general-domain questions by using two layers of retrieval mechanisms [34]. More recently, a benchmarking study evaluated the performance of LLMs in generating PGx-based recommendations [35]. Whereas that study covered the entire spectrum of PGx genes, the present work focused on anticancer drugs and their associated PGx genes. Therefore, this study can be considered as a case study demonstrating how hierarchical retrieval can be applied to domain-specific medical guidelines.

Automated evaluation metrics and human evaluation showed that HRAG was effective in addressing guideline-related queries across four categories: Phenotype to guideline, Drugs to genes, Genes to drugs, and Phenotype to category. In particular, HRAG improved automated evaluation scores for Context precision, Context recall, Context entity recall, Response relevancy, Faithfulness, and F1 in the Phenotype to guideline category. The improved performance of HRAG in guideline-related queries is supported by previous HiRAG research, which highlighted the limitations of RAG. Specifically, dense retrieval systems often divide entity-related information into separate textual segments, disrupting the contextual integrity of the content and hindering accurate retrieval. Prior work demonstrated that retrieving document chunks on a per-entity basis was more effective under these conditions. Our findings reinforce this by showing that HRAG's structured document database enables more accurate retrieval in specific cases.

However, HRAG showed higher noise sensitivity scores, indicating more incorrect claims in the responses. This can be attributed to RAG frequently returning "UNKNOWN" as a response, resulting in shorter or minimal outputs. Since the noise sensitivity metric uses the number of claims in the response as the denominator, systems that generate shorter responses—or avoid making claims altogether—may appear to have lower noise sensitivity. Notably, this difference was not statistically significant. In the Genes to drugs category, Context recall and Response relevancy were calculated as "NA". Such a result was observed because the metrics use the total number of claims in the response as the denominator. In cases where the RAG model's response was "UNKNOWN", the denominator became zero, resulting score computation infeasible. These cases highlight a limitation in the RAGAS framework when models choose not to generate substantive answers.

Among the eight PGxQA categories evaluated in this study, Allele definition, Allele frequency, Allele function, and Diplotype to Phenotype focus primarily on retrieving precise genetic information, where accurate numerical and entity matching (e.g., correct rsIDs or allele labels) is crucial. For example, in the Allele definition category, the ability to correctly identify and explain the rsIDs of a given star-allele is essential for an accurate response. However, in these four categories, both HRAG and RAG showed lower performance compared to other QA categories. This observation aligns with a well-known limitation of current LLMs: difficulty in performing precise numerical reasoning and entity grounding.[36] Our findings reinforce this limitation in the biomedical domain, where precise numeric retrieval is critical. To address this issue, future models may require improved numeric-aware retrieval frameworks to enhance reliability in such contexts.

In healthcare system, utilizing PGx AI assistants has the potential to adopt PGx in clinical workflows by delivering real-time, guideline-concordant drug recommendations tailored to a patient's genetic profile. This could improve treatment efficacy, and accelerate the implementation of precision medicine. Beyond PGx, HRAG approach holds promise for broader medical applications. For example, many areas of medicine, such as disease classification systems (e.g., ICD codes), clinical pathways, and diagnostic frameworks, have tree-like structures. Our findings indicate that hierarchical indexing strategies may be applicable in structuring and navigating complex medical knowledge.

In a clinical context, the strong performance of HRAG in guideline-related categories carries important implications for clinical decision support and medical education. Accurate retrieval and interpretation of PGx guidelines are critical, as previous studies have shown that the clinical implementation of PGx reduces inappropriate prescriptions and prevents ADRs [37–39]. However, clinicians often face substantial workload and time constraints, which hinder the routine use of PGx information in clinical practice [40–46]. LLM-based systems can help alleviate this burden by enabling faster access to accurate guideline information [47]. By reducing clinicians' workload, such tools may facilitate the broader adoption of PGx-guided prescribing, ultimately contributing to safer pharmacotherapy and fewer ADRs. Conversely, HRAG's limitations in precise numerical reasoning and entity grounding extend beyond technical challenges. Errors in retrieving genotype–phenotype relationships may result in misdirected guideline recommendations, potentially compromising patient safety and therapeutic efficacy. These observations underscore the necessity of developing retrieval frameworks that are both numerically robust and semantically precise, ensuring accuracy, safety, and clinical utility in PGx decision support.

This study has several limitations. First, HRAG framework was evaluated only on three anticancer drugs. To generalize the approach to the broader scope of PGx-guided therapies, future research is needed to determine what hierarchical structures would be most appropriate for various classes of drugs beyond oncology. While deeper hierarchies improve granularity, excessive layering can introduce inefficiencies and increase retrieval latency. Therefore, striking an optimal balance between hierarchical depth and retrieval efficiency is critical to maximize the effectiveness of the approach. Second, we did not incorporate additional PGx databases, such as PharmGKB and the FDA table, which contain further relevant information [48]. Integrating these resources is challenging because of differences in how they define and standardize key terms [49]. For example, CPIC and PharmGKB employ different terminologies for "phenotype", emphasizing the need for standardization prior to integration. This lack of standardization also affected our evaluation process: when CPIC and DPWG guidelines presented differing recommendations, the LLM responses tended to aggregate toward CPIC content. Although both guidelines were provided in the practical prompts, resolving these discrepancies is crucial for effective data integration and represents an important subject for future research.

In conclusion, our study proposed a domain-specific HRAG framework for PGx application, particularly focused on anticancer drugs. To the best of our knowledge, no previous work has applied hierarchical retrieval to PGx, making our approach a novel and impactful contribution. HRAG showed significantly improved performance compared to existing RAG approaches, especially in guideline-related queries. We hope that our findings will contribute to the development of more robust and reliable AI assistants that support healthcare professionals in delivering precision medicine.

CRediT authorship contribution statement

Yaejin Jeon: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Mi Seon Youn:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sunghoon Kang:** Writing – review & editing, Writing – original draft, Validation, Software, Investigation, Formal analysis, Data curation. **Jonghyung Park:** Resources, Funding acquisition. **Eun Sil Kim:** Resources, Funding acquisition. **Juyoung Kim:** Software, Resources. **Ju Han Kim:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Funding acquisition, Conceptualization.

Ethics statement

This study did not involve human participants, human data or tissues, or animals, and thus did not require ethical approval or informed consent. All data analyzed in this work are publicly available. The authors declare that they have no competing interests.

Declaration of competing interest

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

Acknowledgments

This research was supported by a grant of the Korean Cancer Survivors Healthcare R&D Project through the National Cancer Center, funded by the Ministry of Health & Welfare, Republic of Korea (RS-2023-CC139850) and J.K. was partly supported by the Education and Research Encouragement Fund of Seoul National University Hospital. The funding sources had no involvement in any aspect of the study, including its design, data collection, analysis, interpretation, manuscript preparation, or the decision to submit for publication.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2025.111323.

Data availability

The datasets used in this study are publicly available. The CPIC guidelines can be accessed at https://cpicpgx.org/, and the DPWG guidelines are available via the Dutch Pharmacogenetics Working Group's website at https://www.knmp.nl/dossiers/farmacogenetica. The PGxQA benchmark dataset is publicly available at https://github.com/KarlKeat/PGxQA.git.

References

[1] Kefyalew Ayalew Getahun, Dessie Abebaw Angaw, Mezgebu Silamsaw Asres, Wubayehu Kahaliw, Zelalem Petros, Solomon Mequanente Abay, Getnet Yimer, Nega Berhane, The role of pharmacogenomics studies for precision medicine among ethiopian patients and their clinical implications: a scoping review, Pharmacogenomics Pers. Med. (2024) 347–361.

[2] Jamie J. Coleman, Sarah K. Pontefract, Adverse drug reactions, Clin. Med. 16 (5) (2016) 481–485.

[3] Nicholas J. Schork, Personalized medicine: time for one-person trials, Nature 520 (7549) (2015) 609–611.

[4] M.V. Relling, T.E. Klein, CPIC: clinical pharmacogenetics implementation consortium of the pharmacogenomics research network, Clin. Pharmacol. Ther. 89 (3) (2011) 464–467.

[5] National Institutes of Health's Pharmacogenomics Research Networ, URL https://www.pgrn.org.

[6] Julia M. Barbarino, Michelle Whirl-Carrillo, Russ B. Altman, Teri E. Klein, PharmGKB: a worldwide resource for pharmacogenomic information, Wiley Interdiscip. Rev.: Syst. Biology Med. 10 (4) (2018) e1417.

[7] Michelle Whirl-Carrillo, Rachel Huddart, Li Gong, Katrin Sangkuhl, Caroline F. Thorn, Ryan Whaley, Teri E. Klein, An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine, Clin. Pharmacol. Ther. 110 (3) (2021) 563–572.

[8] Carin A.T.C. Lunenburg, Cathelijne H. van der Wouden, Marga Nijenhuis, Mandy H. Crommentuijn-van Rhenen, Nienke J. de Boer-Veger, Anne Marie Buunk, Elisa J.F. Houwink, Hans Mulder, Gerard A. Rongen, Ron H.N. van Schaik, et al., Dutch pharmacogenetics working group (DPWG) guideline for the gene–drug interaction of DPYD and fluoropyrimidines, Eur. J. Human Genet. 28 (4) (2020) 508–517.

[9] Maad M. Mijwil, Mostafa Abotaleb, Guma Ali, Klodian Dhoska, Assigning medical professionals: ChatGPT's contributions to medical education and health prediction, Mesopotamian J. Artif. Intell. Heal. 2024 (2024) 76–83.

[10] Shiva Maleki Varnosfaderani, Mohamad Forouzanfar, The role of AI in hospitals and clinics: transforming healthcare in the 21st century, Bioengineering 11 (4) (2024) 337.

[11] Carl Preiksaitis, Christian Rose, Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review, JMIR Med. Educ. 9 (2023) e48785.

[12] Shuo Tian, Wenbo Yang, Jehane Michael Le Grange, Peng Wang, Wei Huang, Zhewei Ye, Smart healthcare: making medical care more intelligent, Glob. Health J. 3 (3) (2019) 62–65.

[13] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, et al., A large language model for electronic health records, NPJ Digit. Med. 5 (1) (2022) 194.

[14] Michelle E. Klein, Md Masud Parvez, Jae-Gook Shin, Clinical implementation of pharmacogenomics for personalized precision medicine: Barriers and solutions, J. Pharm. Sci. 106 (9) (2017) 2368–2379.

[15] Meghan J. Arwood, Supatat Chumnumwat, Larisa H. Cavallari, Edith A. Nutescu, Julio D. Duarte, Implementing pharmacogenomics at your institution: establishment and overcoming implementation challenges, Clin. Transl. Sci. 9 (5) (2016) 233.

[16] Mullai Murugan, Bo Yuan, Eric Venner, Christie M. Ballantyne, Katherine M. Robinson, James C. Coons, Liwen Wang, Philip E. Empey, Richard A. Gibbs, Empowering personalized pharmacogenomics with generative AI solutions, J. Am. Med. Informatics Assoc. 31 (6) (2024) 1356–1366.

[17] Malik Sallam, Roaa Khalil, Mohammed Sallam, Benchmarking generative AI: A call for establishing a comprehensive framework and a generative AIQ test, Mesopotamian J. Artif. Intell. Heal. 2024 (2024) 69–75.

[18] Ursula Amstutz, Linda M. Henricks, Steven M. Offer, Julia Barbarino, Jan H.M. Schellens, Jesse J. Swen, Teri E. Klein, Howard L. McLeod, Kelly E. Caudle, Robert B. Diasio, et al., Clinical pharmacogenetics implementation consortium (CPIC) guideline for dihydropyrimidine dehydrogenase genotype and fluoropyrimidine dosing: 2017 update, Clin. Pharmacol. Ther. 103 (2) (2018) 210–216.

[19] Matthew P. Goetz, Katrin Sangkuhl, Henk-Jan Guchelaar, Matthias Schwab, Michael Province, Michelle Whirl-Carrillo, W. Fraser Symmans, Howard L. McLeod, Mark J. Ratain, Hitoshi Zembutsu, et al., Clinical pharmacogenetics implementation consortium (CPIC) guideline for CYP2D6 and tamoxifen therapy, Clin. Pharmacol. Ther. 103 (5) (2018) 770–777.

[20] the Dutch Pharmacogenetic Working Group, Pharmacogenetics (information in english), 2021, https://www.knmp.nl/sites/default/files/2023-11/Recommendation_text_pharmacogenetics_20211109.pdf. (Accessed: 01 April 2025).

[21] The Clinical Pharmacogenetics Implementation Consortium, Cpic® guideline for fluoropyrimidines and DPYD, 2017, https://cpicpgx.org/guidelines/guideline-for-fluoropyrimidines-and-dpyd/. (Accessed: 02 April 2025).

[22] The Clinical Pharmacogenetics Implementation Consortium, Cpic® guideline for tamoxifen based on CYP2d6 genotype, 2018, https://cpicpgx.org/guidelines/cpic-guideline-for-tamoxifen-based-on-cyp2d6-genotype/. (Accessed: 02 April 2025).

[23] Liwen Wang, Mullai Murugan, Github-PGx4Statins-AI-Assistant, 2024, URL https://github.com/BCM-HGSC/PGx4Statins-AI-Assistant/tree/main.

[24] Karl Keat, Rasika Venkatesh, Yidi Huang, Rachit Kumar, Sony Tuteja, Katrin Sangkuhl, Binglan Li, Li Gong, Michelle Whirl-Carrillo, Teri E. Klein, Marylyn D. Ritchie, Dokyoon Kim, PGxQA: A resource for evaluating LLM performance for pharmacogenomic QA tasks, in: Biocomputing 2025, pp. 229–246.

[25] Karl Keat, Yidi Huang, Rachit Kumar, Rasika Venkatesh, Github-PGxQA, 2025, URL https://github.com/KarlKeat/PGxQA.git.

[26] Jerry Liu, LlamaIndex, 2022.

[27] OpenAI, text-embedding-ada-002, 2022, URL https://platform.openai.com/docs/guides/embeddings.

[28] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R.J. Nelson, Eric Jones, Robert Kern, Eric Larson, C.J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E.A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, SciPy 1.0 Contributors, SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nature Methods 17 (2020) 261–272.

[29] Shahul Es, Jithin James, Luis Espinosa Anke, Steven Schockaert, Ragas: Automated evaluation of retrieval augmented generation, in: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 2024, pp. 150–158.

[30] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi, Bertscore: Evaluating text generation with BERT, in: International Conference on Learning Representations.

[31] Natacha Borgers, Dirk Sikkel, Joop Hox, Response effects in surveys on children and adolescents: The effect of number of response options, negative wording, and neutral mid-point, Qual. Quant. 38 (2004) 17–33.

[32] Qian Ruan, Malte Ostendorff, Georg Rehm, HiStruct+: Improving extractive text summarization with hierarchical structure information, in: Findings of the Association for Computational Linguistics: ACL 2022, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 1292–1308.

[33] Xingxing Zhang, Furu Wei, Ming Zhou, HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 5059–5069.

[34] Xiaoming Zhang, Ming Wang, Xiaocui Yang, Daling Wang, Shi Feng, Yifei Zhang, Hierarchical retrieval-augmented generation model with rethink for multi-hop question answering, 2024, arXiv preprint arXiv:2408.11875.

[35] Mike Zack, Ioan Slobodchikov, Danil Stupichev, Alex Moore, David Sokolov, Igor Trifonov, Allan Gobbs, Benchmarking large language models for replication of guideline-based PGx recommendations, Pharmacogenomics J. 25 (4) (2025) 23.

[36] Haotong Yang, Yi Hu, Shijia Kang, Zhouchen Lin, Muhan Zhang, Number cookbook: Number understanding of language models and how to improve it, Int. Conf. Learn. Represent. (2025).

[37] Kathryn A. Phillips, David L. Veenstra, Eyal Oren, Jane K. Lee, Wolfgang Sadee, Potential role of pharmacogenomics in reducing adverse drug reactions: a systematic review, Jama 286 (18) (2001) 2270–2279.

[38] Giovanni Severino, Maria Del Zompo, Adverse drug reactions: role of pharmacogenomics, Pharmacol. Res. 49 (4) (2004) 363–373.

[39] Laurent Becquemont, Pharmacogenomics of adverse drug reactions: practical applications and perspectives, Pharmacogenomics 10 (6) (2009) 961–969.

[40] Erika N. Dreikorn, Seth Sacchi, Katherine Riden, Philip E. Empey, Mylynda B. Massart, Lucas A. Berenbrok, Understanding clinician support preferences related to the use of pharmacogenomic results in electronic health records: A qualitative study in an academic health system, J. Am. Pharm. Assoc. 65 (6) (2025).

[41] Susanne B. Haga, Wylie Burke, Geoffrey S. Ginsburg, Rachel Mills, Robert Agans, Primary care physicians' knowledge of and experience with pharmacogenetic testing, Clin. Genet. 82 (4) (2012) 388–394.

[42] Lucas A. Berenbrok, Kristin M. Hart, Stephanie Harriman McGrath, Kim C. Coley, Melissa A. Somma McGivney, Philip E. Empey, Community pharmacists' educational needs for implementing clinical pharmacogenomic services, J. Am. Pharm. Assoc. 59 (4) (2019) 539–544.

[43] I. Rafi, I. Crinson, M. Dawes, D. Rafi, M. Pirmohamed, F.M. Walter, The implementation of pharmacogenomics into UK general practice: a qualitative study exploring barriers, challenges and opportunities, J. Community Genet. 11 (3) (2020) 269–277.

[44] Rachel Writer, Christine Barthen, Brandon Antinopoulos, Ryley Uber, James M. Stevenson, Lucas A. Berenbrok, How community pharmacists envision using pharmacogenomic data: A qualitative analysis, J. Am. Pharm. Assoc. 61 (5) (2021) e64–e70.

[45] Angela Pearce, Bronwyn Terrill, Jan-Willem Alffenaar, Asad E. Patanwala, Sarah Kummerfeld, Richard Day, Mary-Anne Young, Sophie L. Stocker, Pharmacogenomic testing: Perception of clinical utility, enablers and barriers to adoption in Australian hospitals, Intern. Med. J. 52 (7) (2022) 1135–1143.

[46] Sarah A. Shue, Elizabeth Rowe, Lauren A. Bell, Teresa Damush, Alexis DeLong, Tayler Gowan, Todd Skaar, David Haggstrom, Pharmacogenomics implementation across multiple clinic settings: a qualitative evaluation, Pharmacogenomics 24 (17) (2023) 881–893.

[47] Annisa Fatharani, Ali Alsayegh, Pharmacogenomics meets generative AI: transforming clinical trial design with large language models, J. Pharmacol. Pharmacother. (2025) 0976500X251321885.

[48] Mohammad A. Alshabeeb, Mesnad Alyabsi, Mohammad A. Aziz, Salah Abohelaika, Pharmacogenes that demonstrate high association evidence according to CPIC, DPWG, and PharmGKB, Front. Med. 9 (2022) 1001876.

[49] Mi Seon Youn, Se Hwan Ahn, Ju Han Kim, Pharmacogenomic profiling of the South Korean population: Insights and implications for personalized medicine, Front. Pharmacol. 15 (2024) 1476765.