

MED-COPILOT: A Medical Assistant Powered by GraphRAG and Similar Patient Case Retrieval

Shuheng Chen, Namratha Patil, Haonan Pan, Angel Hsing-Chi Hwang,
Yao Du, Ruishan Liu, Jieyu Zhao
University of Southern California

Abstract

Clinical decision-making requires synthesizing heterogeneous evidence, including patient histories, clinical guidelines, and trajectories of comparable cases. While large language models (LLMs) offer strong reasoning capabilities, they remain prone to hallucinations and struggle to integrate long, structured medical documents. We present MED-COPILOT, an interactive research prototype for evidence-aware clinical reasoning, designed to help clinicians and medical trainees inspect guideline-level and patient-level evidence. MED-COPILOT combines guideline-grounded GraphRAG retrieval with hybrid semantic-keyword similar-patient retrieval to support transparent and evidence-aware clinical reasoning. The system builds a structured knowledge graph from WHO and NICE guidelines, applies community-level summarization for efficient retrieval, and maintains a 36,000-case similar-patient database derived from SOAP-normalized MIMIC-IV notes and Synthea-generated records.

We evaluate our framework on clinical note completion and medical question answering, and demonstrate that it consistently outperforms parametric LLM baselines and standard RAG, improving generation fidelity and benchmark QA accuracy. The full system is available at <https://huggingface.co/spaces/shuhengc/MED-COPILOT>, enabling users to inspect retrieved evidence, visualize token-level similarity contributions, and conduct guided follow-up analysis. Our results suggest a practical and interpretable approach to integrating structured guideline knowledge with patient-level analogical evidence for clinical LLMs.

1 Introduction

Clinical reasoning requires integrating heterogeneous evidence sources, including patient histories, population-level clinical guidelines, and outcomes from comparable prior cases (Esteva et al., 2019;

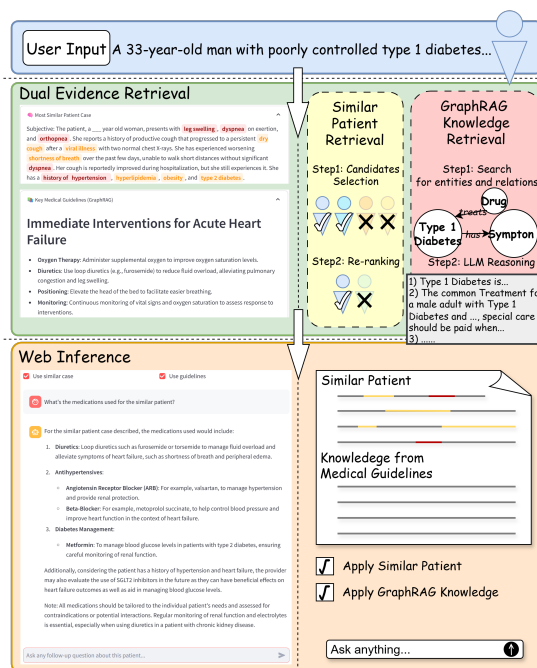


Figure 1: Overview of the proposed system, which performs dual evidence retrieval from similar patient cases and graph-structured clinical guidelines to support clinical inference.

Rajkomar et al., 2019). While large language models (LLMs) exhibit strong open-domain reasoning capabilities, their use in medical settings remains risky due to hallucinations, confidence miscalibration, and a lack of verifiable evidence grounding (Pandit et al., 2025; Kim et al., 2025; Huang et al., 2025). Consequently, there is growing consensus that safe and trustworthy medical AI must rely on explicit retrieval of external clinical evidence rather than purely parametric knowledge (Nazi and Peng, 2024; Asgari et al., 2025), particularly for clinical decision-support systems intended to assist clinicians in real-world reasoning workflows.

Retrieval-Augmented Generation (RAG) frameworks partially address this need by grounding model outputs in retrieved text (Lewis et al., 2020;

Izacard and Grave, 2021). However, conventional RAG operates over flat, unstructured corpora and struggles to model the relational and hierarchical structure essential for clinical reasoning (Lewis et al., 2020; Zhang et al., 2025). Recent work has therefore emphasized graph-based retrieval, such as knowledge-graph-augmented RAG and GraphRAG, to support multi-step reasoning over structured clinical knowledge (Liang et al., 2025; Sanmartin, 2024; Wu et al., 2025).

Despite these advances, a critical aspect of clinical reasoning remains largely unsupported: patient-level analogical reasoning. In clinical practice, decisions are frequently informed by comparisons to clinically similar patients with shared symptoms, comorbidities, or disease trajectories (Beam and Kohane, 2018; Rajkomar et al., 2018). Existing clinical RAG systems primarily focus on guideline or document retrieval and rarely incorporate patient-level comparators into the reasoning loop (Wu et al., 2025; Zhang et al., 2025; Lu et al., 2024), leaving an important gap in case-centric decision.

To bridge these limitations, we propose a unified and interpretable framework that integrates guideline-grounded GraphRAG retrieval, hybrid similar-patient retrieval, and LLM-based reasoning within a single pipeline. The system constructs a structured knowledge graph from authoritative WHO and NICE clinical guidelines (Organization, 2024; for Health and Excellence, 2024) and pairs it with a large-scale similar-patient database derived from SOAP-normalized MIMIC-IV ICU records and Synthea-generated synthetic trajectories (Johnson et al., 2023; Walonoski et al., 2018). A hybrid similarity function combines clinically weighted keyword matching with semantic embedding similarity, enabling flexible and transparent evidence selection during inference.

We evaluate our framework on both generative and discriminative medical reasoning tasks, including clinical note completion on MIMIC-IV and standardized medical question answering benchmarks such as MedQA and the clinical subset of MMLU (Jin et al., 2021; Hendrycks et al., 2021). Across settings, the proposed approach consistently outperforms LLM-only and conventional RAG baselines, demonstrating the complementary value of structured guideline knowledge and patient-level analogical evidence. We additionally release an interactive system to support evidence inspection and reproducible clinical reasoning workflows.

Our focus is to build an end-to-end interactive

workflow that makes complementary clinical evidence sources inspectable, controllable, and usable within a unified interface.

Our contributions are threefold: (1) we present an end-to-end interactive system that jointly exposes graph-structured guideline retrieval and similar-patient retrieval for evidence-aware clinical reasoning; (2) we construct standardized clinical evidence resources spanning authoritative guidelines and patient trajectories; and (3) we demonstrate consistent empirical gains across medical reasoning tasks and provide a public demo interface for evidence inspection. Figure 1 provides an overview of the dual-evidence retrieval workflow.

2 Related Work

2.1 Retrieval-Augmented Clinical Reasoning

RAG enhances large language models by grounding generation in externally retrieved evidence, and has become a widely adopted paradigm for medical question answering and clinical reasoning (Lewis et al., 2020; Izacard and Grave, 2021; Xiong et al., 2024a). Domain-specific variants adapt RAG to biomedical corpora such as PubMed, clinical guidelines, and electronic health records, demonstrating substantial improvements over LLM-only or chain-of-thought baselines on different benchmarks including MedQA (Xiong et al., 2024a,b).

However, prior studies consistently report that text-based RAG remains sensitive to retrieval noise and degrades under long-context and multi-hop reasoning scenarios, limiting its reliability for complex clinical decision-making (Tang and Yang, 2024; Sun et al., 2025; Gupta et al., 2024). To address these limitations, recent work incorporates structured knowledge graphs into retrieval pipelines, enabling explicit modeling of biomedical entities and relations for more interpretable multi-step reasoning (Yasunaga et al., 2021; Sohn et al., 2025). GraphRAG-style systems further organize retrieved evidence into graph-structured representations and community-level summaries, improving provenance tracking and reasoning consistency in medical QA and diagnostic tasks (Wu et al., 2025; Sekar et al., 2025; Yang et al., 2024). Despite these advances, existing systems primarily focus on guideline- or document-level evidence and largely overlook patient-level analogical reasoning.

2.2 Similar-Patient Retrieval

Similar-patient retrieval has been extensively studied in intensive care analytics and personalized risk modeling. Early work demonstrates that trajectory-based similarity enables cohort-level outcome estimation and interpretable subgroup analysis in ICU populations (Lee et al., 2015; Alcaide and Aerts, 2021). More recent representation-learning approaches model temporal dynamics and relational structures among patients to improve predictive performance (Sun et al., 2022; Ma et al., 2023).

Despite its clinical relevance, prior work typically treats patient similarity as a standalone analytical or predictive tool, rather than integrating similar-patient evidence into retrieval-augmented or LLM-based reasoning pipelines (Sharafoddini et al., 2017; Zhao et al., 2023). As a result, patient-level analogical evidence remains underutilized in current clinical RAG systems, motivating the urgent need for unified frameworks that combine structured guideline knowledge with clinically comparable patient trajectories.

3 Methodology

This section describes the design of a unified retrieval framework that integrates graph-augmented clinical guidelines and similar-patient evidence. We focus on how authoritative medical knowledge is represented, indexed, and retrieved to support interactive and transparent clinical reasoning.

3.1 System Overview

The proposed system supports clinical queries in the form of either SOAP-formatted patient cases or free-text questions. Given an input, it retrieves two complementary forms of evidence: (i) graph-structured guideline knowledge encoding multi-step clinical logic, and (ii) clinically similar patient cases capturing trajectory-level analogical patterns. Retrieved evidence is aggregated and presented with explicit provenance to support downstream reasoning and user inspection.

3.2 Graph-Augmented Guideline Retrieval

Guideline Sources. To ensure evidential reliability and clinical safety, all guideline knowledge is sourced exclusively from internationally sanctioned authorities. Specifically, we curate clinical practice guidelines from the World Health Organization (WHO) and the National Institute for Health and Care Excellence (NICE), both of which pro-

vide consensus-based, evidence-graded recommendations with transparent revision and disclosure procedures (Organization, 2024; for Health and Excellence, 2024).

After normalization and de-duplication, the resulting corpus comprises 118 WHO guidelines and 525 NICE guidelines, totaling over 40 million tokens of clinically actionable recommendations. These guidelines explicitly articulate eligibility criteria, contraindications, risk boundaries, and escalation pathways, forming a reliable substrate for structured retrieval and multi-hop reasoning.

Graph Construction. Clinical guidelines are long, hierarchically structured documents whose logic cannot be fully captured through flat text segmentation alone. While segmenting guidelines into semantically coherent units supports fine-grained retrieval, it fails to expose relational structures such as conditional applicability, treatment dependencies, and escalation logic.

To address this limitation, each guideline is first segmented into semantically coherent *TextUnits*, where each unit corresponds to a single recommendation, decision step, or well-defined clinical scenario. Segmentation respects the native guideline hierarchy rather than arbitrary token windows. Each *TextUnit* is annotated with metadata derived from the original guideline document, including its source, section title, and character span, enabling transparent provenance tracking and citation.

We then construct a clinical knowledge graph by extracting clinically meaningful entities (e.g., diseases, symptoms, medications, procedures) and relations (e.g., indication, contraindication, monitoring, escalation) from the segmented *TextUnits*. Entity mentions are identified using medical named-entity recognition and domain-adapted LLM prompting, and normalized to standard biomedical vocabularies (e.g., SNOMED-CT, ICD-10, ATC). Relations and constraint qualifiers (such as age limits or escalation triggers) are encoded explicitly, and all graph elements retain links to their originating guideline sections. This graph representation enables structured navigation of guideline logic beyond only what is accessible through text-only retrieval.

Indexing. To support efficient retrieval across both fine-grained recommendations and higher-level guideline logic, we adopt a hybrid indexing strategy. Guideline *TextUnits*, graph community summaries, and entity descriptions are embedded

using BioClinicalBERT and stored in a local vector database (LanceDB). Indexing is organized by evidence modality, preserving the structural hierarchy of the guideline graph while enabling flexible retrieval at different levels of granularity with full provenance traceability.

3.3 Similar-Patient Retrieval

Patient Data Sources. Similar-patient retrieval operates over a curated repository of 36,000 structured patient cases. To balance clinical realism and controlled coverage, the repository integrates two complementary sources. First, 18,000 intensive care encounters are sampled from MIMIC-IV, a large-scale de-identified critical care database containing comprehensive EHR-derived clinical documentation (Johnson et al., 2023). Each admission is converted into a standardized SOAP representation using a rule-guided conversion agent, preserving temporal evolution and diagnostic intent while normalizing stylistic variation.

Second, 18,000 additional patient trajectories are generated using Synthea, an agent-based synthetic patient simulator designed to model longitudinal disease courses and care pathways (Walonoski et al., 2018). These synthetic records are rewritten under the same SOAP specification to align structure and semantic granularity with the MIMIC-derived cohort. The combined repository maintains structural uniformity and broadens coverage across etiologies, acuity levels, and disease trajectories, while acknowledging that synthetic records may not fully reproduce the complexity of real-world clinical documentation.

Hybrid Similarity Scoring. Identifying clinically similar patients requires balancing strict clinical constraints with trajectory-level similarity, motivating a hybrid retrieval formulation that combines keyword alignment with semantic embeddings.

To identify clinically comparable cases, we rank candidate patients using a hybrid similarity formulation that combines keyword-conditioned alignment with embedding-based semantic proximity. The keyword component emphasizes discrete clinical signals such as diagnoses, comorbidities, and key interventions, while the semantic component captures graded similarity in longitudinal trajectories and response patterns. This dual formulation allows retrieval to satisfy eligibility-level constraints while remaining robust to surface-level lexical variation. For conciseness, we present only the high-

level formulation here.

Importantly, we do not assume that retrieved patients are clinically identical to the query case. Similar-patient retrieval is used to surface partial analogs that share salient diagnoses, symptoms, comorbidities, laboratory patterns, or intervention histories. Retrieved cases are therefore interpreted as comparative evidence rather than direct clinical precedents. To reduce over-interpretation of weak matches, the interface highlights the specific clinical attributes contributing to each retrieved case.

3.4 Unified Retrieval Workflow

At inference time, a query case or question is embedded and used to retrieve semantically relevant guideline artifacts and patient cases from the vector index. In parallel, graph-conditioned retrieval identifies guideline communities whose relational structure aligns with the query, enabling access to multi-hop clinical logic such as contraindications, escalation pathways, and applicability constraints.

Retrieved guideline evidence and similar patient cases are then aggregated into a unified evidence set comprising ranked guideline communities with supporting nodes and ranked patient cases with aligned attributes. All retrieved items are presented with explicit source links, allowing users to inspect provenance and understand the basis of the system’s recommendations.

Details of the GraphRAG construction and retrieval pipeline are provided in Appendix B.

4 Evaluation and Demonstration

We evaluate the proposed framework on clinical note generation and medical question answering (QA), and additionally present an interactive system demonstration. Our evaluation is intended to assess retrieval-grounded generation and reasoning behavior rather than to establish clinical effectiveness in deployment. Accordingly, we interpret the benchmark results as evidence of complementary retrieval benefits, while leaving clinician-centered validation of usability, cognitive load, and workflow impact to future work.

4.1 Tasks and Setup

Clinical Note Generation. We sample 1,000 de-identified ICU notes from MIMIC-IV and convert each into standardized SOAP format (Weed, 1968). The model receives the Subjective (S), Objective (O), and Assessment (A) sections and generates the corresponding Plan (P).

Medical QA. We evaluate on MedQA (Jin et al., 2021) and the clinical subset of MMLU (Hendrycks et al., 2021) using the same retrieval-augmented pipeline as in the clinical setting.

The similarity mixing weight λ is tuned on a 100-case validation set and fixed across all experiments.

4.2 Metrics and Baselines

For clinical note generation, we report standard lexical and semantic similarity metrics, including BLEU, METEOR, ROUGE-1/2/L, and BERTScore-F1. For QA, we report accuracy.

We compare against strong parametric-only LLM baselines (DeepSeek-Chat, GPT-4.1-mini, and Gemini 2.5) under identical prompting and decoding settings.

4.3 Main Results

Table 1 summarizes performance across tasks. The full MED-COPILOT framework consistently improves over the no-retrieval baselines across both generation and QA tasks, and generally outperforms standard RAG across backbone models. Combining similar-patient retrieval and GraphRAG yields the strongest overall results.

On clinical note generation, the full system improves ROUGE-L over the corresponding no-retrieval baselines for all three backbone models, with GPT-4.1-mini increasing from 0.144 to 0.291 and Gemini 2.5 increasing from 0.147 to 0.300. The gains are also reflected in lexical and semantic metrics such as BLEU, METEOR, ROUGE-1/2, and BERTScore-F1, suggesting that retrieved evidence helps the models produce plans that better match the reference notes under automatic evaluation. Compared with standard RAG, the dual-evidence system further improves most generation metrics, indicating that flat document retrieval alone does not capture all useful case- and guideline-level signals. For medical QA, the largest gains appear on MMLU-clinical, where GPT-4.1-mini improves from 0.868 without retrieval and 0.887 with standard RAG to 0.971 with the full system. MedQA accuracy also improves consistently, with Gemini 2.5 increasing from 0.867 without retrieval to 0.937 with the full system. These patterns suggest that structured guideline evidence and patient-level analogical evidence provide complementary context for both generation and knowledge-intensive reasoning benchmarks.

4.4 Ablation

Using GPT-4.1-mini as the backbone model, Table 2 isolates the contribution of each retrieval component. Both similar-patient retrieval and GraphRAG knowledge retrieval independently improve performance, with the combined framework achieving the largest gains.

For note generation, similar-patient retrieval (SPR) alone substantially improves BLEU from 0.012 to 0.028 and METEOR from 0.282 to 0.315, suggesting that similar-patient cases contribute useful case-level context for plan completion. GraphRAG alone yields larger gains on ROUGE-L, increasing from 0.144 to 0.258, which is consistent with guideline retrieval helping recover structured treatment and management content. Combining both sources achieves the best scores across all generation metrics, including ROUGE-L of 0.291 and BERTScore-F1 of 0.851. For QA, GraphRAG provides stronger gains than SPR alone, improving MedQA from 0.846 to 0.870 and MMLU-clinical from 0.868 to 0.952, whereas SPR improves them to 0.851 and 0.914. This trend suggests that guideline structure is especially useful for knowledge-intensive questions, while patient analogs provide additional complementary signal. The combined setting reaches the highest QA accuracy on both MedQA (0.912) and MMLU-clinical (0.971), supporting the system-level integration of the two evidence sources.

4.5 Interactive System

We deploy the proposed framework as a public HuggingFace Space.¹ The interface provides end-to-end interactive clinical exploration grounded in dual evidence retrieval, enabling clinicians to inspect, compare, and reason over guideline-level and patient-level evidence within a unified evidence-inspection workflow.

Dual Evidence Control. Given a free-text clinical case, the system retrieves (i) similar patient records from the case repository and (ii) guideline-level evidence via the GraphRAG pipeline. Users may selectively enable or disable each evidence source through explicit interface controls, allowing controllable reasoning under patient-level evidence, guideline-level knowledge, or their combination. The selected evidence is automatically packaged

¹<https://huggingface.co/spaces/shuhengc/MED-COPILOT>

Model	Text Generation						QA	
	MIMIC-IV Clinical Notes						MedQA	MMLU-clinical
	BLEU	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore-F1	Acc	Acc
Baselines without Retrieval								
DeepSeek-Chat	0.006	0.251	0.205	0.042	0.120	0.809	0.809	0.857
GPT-4.1-mini	0.012	0.282	0.301	0.064	0.144	0.825	0.846	0.868
Gemini 2.5	0.014	0.289	0.295	0.070	0.147	0.829	0.867	0.875
with RAG								
DeepSeek-Chat	0.015	0.272	0.229	0.057	0.157	0.799	0.851	0.875
GPT-4.1-mini	0.019	0.291	0.351	0.081	0.221	0.831	0.885	0.887
Gemini 2.5	0.020	0.316	0.368	0.099	0.214	0.833	0.892	0.905
Our Framework								
DeepSeek-Chat	0.029	0.287	0.340	0.097	0.203	0.839	0.864	0.910
GPT-4.1-mini	0.035	0.321	0.380	0.124	0.291	0.851	0.912	0.971
Gemini 2.5	0.042	0.334	0.419	0.155	0.300	0.862	0.937	0.971

Table 1: Performance under different retrieval configurations on clinical note generation and QA benchmarks.

Setting	Text Generation						QA	
	MIMIC-IV Clinical Notes						MedQA	MMLU-clinical
	BLEU	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore-F1	Acc	Acc
Baseline	0.012	0.282	0.301	0.064	0.144	0.825	0.846	0.868
+ SPR	0.028	0.315	0.324	0.099	0.159	0.828	0.851	0.914
+ GraphRAG	0.031	0.310	0.352	0.115	0.258	0.840	0.870	0.952
+ Both	0.035	0.321	0.380	0.124	0.291	0.851	0.912	0.971

Table 2: Ablation results using GPT-4.1-mini as the backbone model, isolating the contribution of similar-patient retrieval (SPR) and GraphRAG-based guideline reasoning.

with the query context for downstream generation and question answering.

Query-Conditioned Saliency Highlighting. To improve interpretability during evidence inspection, the similar-patient panel incorporates a lightweight query-conditioned saliency overlay. Each extracted clinical concept (e.g., symptoms, diagnoses, comorbidities) is assigned an importance score based on its semantic similarity to the input query embedding.

The score is discretized into two levels and visually encoded (yellow: important; red: highly important), enabling users to quickly identify case attributes most relevant to the current clinical question.

These highlights are intended as navigation aids rather than validated clinical importance labels. They indicate query-conditioned textual and semantic relevance, and should not be interpreted as clinician-verified saliency or causal importance.

Figure 2 illustrates interactive question answering with controllable evidence selection. Figure 3 shows the similar-patient view with query-conditioned saliency highlighting.

Use similar case
 Use guidelines

What's the medications used for the similar patient?

For the similar patient case described, the medications used would include:

1. **Diuretics:** Loop diuretics such as furosemide or torsemide to manage fluid overload and alleviate symptoms of heart failure, such as shortness of breath and peripheral edema.
2. **Antihypertensives:**
 - **Angiotensin Receptor Blocker (ARB):** For example, valsartan, to manage hypertension and provide renal protection.
 - **Beta-Blocker:** For example, metoprolol succinate, to help control blood pressure and improve heart function in the context of heart failure.
3. **Diabetes Management:**
 - **Metformin:** To manage blood glucose levels in patients with type 2 diabetes, ensuring careful monitoring of renal function.

Additionally, considering the patient has a history of hypertension and heart failure, the provider may also evaluate the use of SGLT2 inhibitors in the future as they can have beneficial effects on heart failure outcomes as well as aid in managing blood glucose levels.

Note: All medications should be tailored to the individual patient's needs and assessed for contraindications or potential interactions. Regular monitoring of renal function and electrolytes is essential, especially when using diuretics in a patient with chronic kidney disease.

Ask any follow-up question about this patient... ➤

Figure 2: Interactive question answering with dual evidence control.

Step 2 — Review similar case and guidelines

Most Similar Patient Case

Subjective: The patient is a female with a history of **heart failure** with reduced ejection fraction (HFrEF, EF 25%), **coronary artery disease** (CAD), **atrial fibrillation**, **type 2 diabetes mellitus** (T2DM), and **hypertension** (HTN), who presents with **worsening dyspnea** and **chest pain**. She reports increased tiredness, decreased appetite, and progressively worsening lower extremity edema over the last few days. This morning at 9 AM, she experienced **sudden onset chest pain**, described as non-radiating, which has somewhat decreased but is still present. She has had worsening **shortness of breath**, **orthopnea**, and a **dry cough**, and denies fever, chills, nausea, or vomiting. The patient was seen in the ED two days ago for **left hand pain** attributed to gout, where she was prescribed colchicine. Her family history includes diabetes and **heart disease**, and she has **multiple past medical conditions** including gout and Non-Hodgkin's lymphoma.

Objective: Admission vitals show 98 ____, HR 18, O2 saturation 94% on 2L NC. Weight: 85.7 kg. Elevated JVP to jawline. Physical exam reveals irregular heart rhythm, but normal S1 and S2 without murmurs or gallops. Lung examination shows diffuse crackles bilaterally. Extremities show 1+ pitting edema to knees bilaterally. Admission labs indicate elevated proBNP (8312), glucose (302), creatinine (1.5), and potassium (7.1). CXR suggests **moderate pulmonary edema** and possible pneumonia. Echocardiogram shows LVEF 25% with severe LV hypokinesis and moderate to severe mitral regurgitation. Discharge labs show creatinine at 1.7, glucose at 79, and weight at 82.1 kg. Patient was diuresed effectively with Lasix and transitioned to Torsemide.

Assessment:

1. Acute on **chronic systolic heart failure** secondary to multivessel CAD **exacerbation**, likely triggered by **fluid overload**.
2. **Chest pain** likely secondary to type 2 NSTEMI related to CHF **exacerbation** without ongoing **ischemic** changes.
3. Gout flare contributing to wrist pain.
4. **Atrial fibrillation** managed with apixaban.
5. **Poorly controlled diabetes** (I/DDM).
6. **Chronic kidney disease** (CKD), likely multifactorial.

Plan:

1. Medications adjusted to include Torsemide 40 mg BID, Imdur 30 mg daily, Hydralazine 25 mg TID, and Spironolactone 25 mg daily. 2. Continue apixaban, atorvastatin, and metoprolol with increased dosage to 75 mg BID. 3. Start colchicine 0.6 mg BID for gout with plans to consider **uric acid lowering therapy** after **acute flare** resolves.
4. Monitor diabetes management with decreased insulin doses.
5. Arrange follow-up with **nephrology** for CKD.
6. Discharge with rolling walker and education on monitoring weight.
7. Schedule follow-up appointments as needed.

Figure 3: Similar-patient panel with query-conditioned saliency highlighting. Clinical concepts are color-coded according to their semantic relevance to the input query (yellow: important; red: highly important).

Additional screenshots and a step-by-step walk-through are provided in Appendix A.

5 Limitations

The proposed system has several limitations. First, MED-COPILOT remains a research prototype and has not been clinically validated in real-world workflows. Our evaluation is benchmark-based and does not include a clinician user study; therefore, we cannot yet assess whether the retrieved evidence, saliency highlights, and interface controls reduce cognitive load or improve workflow-level clinical decisions. Second, the similar-patient repository combines SOAP-normalized MIMIC-IV records with Synthea-generated synthetic cases. Although this design improves coverage and structural consistency, synthetic cases may not fully capture the temporal irregularity, missingness, documentation noise, and comorbidity structure of real EHR data. We also do not separately quantify retrieval performance over MIMIC-IV-derived and Synthea-derived cases. Third, patient similarity should be interpreted as partial analogical similarity rather than

case equivalence. Retrieved patients may share key symptoms, diagnoses, or treatment patterns with the query case, but no retrieved case should be treated as a direct precedent for clinical action. Fourth, the guideline graph abstracts narrative recommendations into structured entities and relations, which may omit contextual nuances from the original guideline text. For high-stakes interpretation, users should inspect the linked source guideline sections. Finally, the system depends on the quality of its preprocessing pipeline, including SOAP normalization, entity and relation extraction, and hybrid similarity scoring; errors in these components may propagate to retrieved evidence and downstream generation. No gold-standard dataset currently exists for assessing patient-level similarity retrieval.

Future work will explore clinician-centered evaluation, source-stratified retrieval analysis, expanded guideline coverage, uncertainty-aware reasoning, and deployment considerations in realistic clinical workflows.

6 Conclusion

We presented MED-COPILOT, an interactive research prototype designed to support clinicians and trainees in transparent and evidence-aware reasoning. By integrating guideline-grounded GraphRAG retrieval with similar-patient evidence within a unified and interpretable pipeline, the system enables clinicians to inspect, compare, and reason over multi-source clinical evidence. Experimental results suggest that combining structured guideline knowledge with patient-level analogical retrieval provides complementary benefits across clinical note generation and medical QA tasks, highlighting the value of controllable and inspectable evidence integration for clinical LLM systems. Future work will focus on expanding guideline coverage, improving uncertainty-aware reasoning, and evaluating the system in realistic clinical workflows.

Ethical Statement

This work uses the MIMIC-IV database under its approved data-use agreement, and all authors completed the required credentialing. Synthetic cases generated with Synthea do not correspond to real patients. The system is currently a research prototype and should not be used for clinical diagnosis or treatment directly.

References

- Daniel Alcaide and Jan Aerts. 2021. A visual analytic approach for the identification of icu patient subpopulations using icd diagnostic codes. *PeerJ Computer Science*, 7:e430.
- Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. 2025. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *npj Digital Medicine*, 8(1):274.
- Andrew L Beam and Isaac S Kohane. 2018. Big data and machine learning in health care. *Jama*, 319(13):1317–1318.
- Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. 2019. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29.
- National Institute for Health and Care Excellence. 2024. Nice clinical guidelines and knowledge summaries. Available from: <https://www.nice.org.uk/guidance>.
- Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. 2024. A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions. *arXiv preprint arXiv:2410.12837*.
- Dan Hendrycks and 1 others. 2021. Measuring massive multitask language understanding. *ICLR*.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.
- Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, pages 874–880.
- Di Jin and 1 others. 2021. What disease does this patient have? a large-scale open-domain medical qa dataset. In *EMNLP*.
- Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.
- Jonathan Kim, Anna Podlasek, Kie Shidara, Feng Liu, Ahmed Alaa, and Danilo Bernardo. 2025. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. *Scientific reports*, 15(1):39426.
- Joon Lee, David M Maslove, and Joel A Dubin. 2015. Personalized mortality prediction driven by electronic medical data and a patient similarity metric. *PLoS one*, 10(5):e0127428.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Lei Liang, Zhongpu Bo, Zhengke Gui, Zhongshu Zhu, Ling Zhong, Peilong Zhao, Mengshu Sun, Zhiqiang Zhang, Jun Zhou, Wenguang Chen, and 1 others. 2025. Kag: Boosting llms in professional domains via knowledge augmented generation. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 334–343.
- Yuxing Lu, Xukai Zhao, and Jinzhuo Wang. 2024. Clinicalrag: Enhancing clinical decision support through heterogeneous knowledge retrieval. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 64–68.
- Manfu Ma, Penghui Sun, Yong Li, and Weilong Huo. 2023. Predicting the risk of mortality in icu patients based on dynamic graph attention network of patient similarity. *Mathematical Biosciences and Engineering*, 20(8):15326–15345.
- Zabir Al Nazi and Wei Peng. 2024. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, page 57. MDPI.
- World Health Organization. 2024. Who guidelines repository. Available from: <https://www.who.int/publications/who-guidelines>.
- Shrey Pandit, Jiawei Xu, Junyuan Hong, Zhangyang Wang, Tianlong Chen, Kaidi Xu, and Ying Ding. 2025. Medhallu: A comprehensive benchmark for detecting medical hallucinations in large language models. *arXiv preprint arXiv:2502.14302*.
- Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. 2019. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358.
- Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, and 1 others. 2018. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18.
- Diego Sanmartin. 2024. Kg-rag: Bridging the gap between knowledge and creativity. *arXiv preprint arXiv:2405.12035*.
- Tharun Sekar, Kushal, Supprethaa Shankar, Sabah Mohammed, and Jinan Fiaidhi. 2025. Investigations on using evidence-based graphrag pipeline using llm tailored for usmle style questions. *medRxiv*, pages 2025–05.

- Anis Sharafoddini, Joel A Dubin, and Joon Lee. 2017. Patient similarity in prediction models based on health data: a scoping review. *JMIR medical informatics*, 5(1):e6730.
- Jiwoong Sohn, Yein Park, Chanwoong Yoon, Sihyeon Park, Hyeon Hwang, Mujeen Sung, Hyunjae Kim, and Jaewoo Kang. 2025. Rationale-guided retrieval augmented generation for medical question answering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12739–12753.
- Duolin Sun, Dan Yang, Yue Shen, Yihan Jiao, Zhehao Tan, Jie Feng, Lianzhen Zhong, Jian Wang, Peng Wei, and Jinjie Gu. 2025. Hanrag: Heuristic accurate noise-resistant retrieval-augmented generation for multi-hop question answering. *arXiv preprint arXiv:2509.09713*.
- Zhaohong Sun, Xudong Lu, Huilong Duan, and Haomin Li. 2022. Deep dynamic patient similarity analysis: Model development and validation in icu. *Computer Methods and Programs in Biomedicine*, 225:107033.
- Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391*.
- Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. 2018. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238.
- Lawrence Weed. 1968. Medical records, medical education, and patient care. *The Problem-Oriented Medical Record*.
- Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. 2025. Medical graph RAG: Evidence-based medical large language model via graph retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28443–28467, Vienna, Austria. Association for Computational Linguistics.
- Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024a. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251.
- Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2024b. Improving retrieval-augmented generation in medicine with iterative follow-up questions. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pages 199–214. World Scientific.
- Rui Yang, Haoran Liu, Edison Marrese-Taylor, Qingcheng Zeng, Yuhe Ke, Wanxin Li, Lechao Cheng, Qingyu Chen, James Caverlee, Yutaka Matsuo, and 1 others. 2024. Kg-rank: Enhancing large language models for medical qa with knowledge graphs and ranking techniques. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 155–166.
- Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. *arXiv preprint arXiv:2104.06378*.
- Qinggong Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Hao Chen, Yilin Xiao, Chuang Zhou, Junnan Dong, and 1 others. 2025. A survey of graph retrieval-augmented generation for customized large language models. *arXiv preprint arXiv:2501.13958*.
- Zhengyun Zhao, Qiao Jin, Fangyuan Chen, Tuorui Peng, and Sheng Yu. 2023. A large-scale dataset of patient summaries for retrieval-based clinical decision support systems. *Scientific data*, 10(1):909.

A Demo Walkthrough

This appendix provides a step-by-step walkthrough of the interactive demo interface.

A.1 Case Input and Query Locking

Medical GraphRAG Assistant

Patient Case Chat / Q&A

Step 1 — Enter and lock the patient case

Patient case / HPI / key findings:

A 72-year-old woman with a history of hypertension, type 2 diabetes, and chronic kidney disease presented to the emergency department with progressively worsening shortness of breath and bilateral leg swelling over the past three days. On arrival, her blood pressure was 165/90 mmHg, heart rate was 110 beats per minute, respiratory rate was 22 breaths per minute, and oxygen saturation was 88% on room air. Bibasilar crackles were noted on lung auscultation.

Lock / Update Patient Case

Clear Locked Case

Patient case locked and initial evaluation completed!

Figure 4: Patient case input and query locking interface.

A.2 Similar Patient Retrieval

Step 2 — Review similar case and guidelines

Most Similar Patient Case

Subjective: The patient is a female with a history of **heart failure** with reduced ejection fraction (HFrEF, EF 25%), **coronary artery disease** (CAD), **atrial fibrillation**, **type 2 diabetes mellitus** (T2DM), and **hypertension** (HTN), who presents with **worsening dyspnea** and **chest pain**. She reports increased tiredness, decreased appetite, and progressively worsening lower extremity edema over the last few days. This morning at 9 AM, she experienced **sudden onset chest pain**, described as non-radiating, which has somewhat decreased but is still present. She has had worsening **shortness of breath**, **orthopnea**, and a **dry cough**, and denies fever, chills, nausea, or vomiting. The patient was seen in the ED two days ago for **left hand pain** attributed to gout, where she was prescribed colchicine. Her family history includes diabetes and **heart disease**, and she has **multiple past medical conditions** including gout and Non-Hodgkin's lymphoma.

Objective: Admission vitals show 98 ____, HR 18, O2 saturation 94% on 2L NC. Weight: 85.7 kg. Elevated JVP to jawline. Physical exam reveals irregular heart rhythm, but normal S1 and S2 without murmurs or gallops. Lung examination shows diffuse crackles bilaterally. Extremities show 1+ pitting edema to knees bilaterally. Admission labs indicate elevated proBNP (8312), glucose (302), creatinine (1.5), and potassium (7.1). CXR suggests **moderate pulmonary edema** and possible pneumonia. Echocardiogram shows LVEF 25% with severe LV hypokinesis and moderate to severe mitral regurgitation. Discharge labs show creatinine at 1.7, glucose at 79, and weight at 82.1 kg. Patient was diuresed effectively with Lasix and transitioned to Torsemide.

Assessment:

1. Acute on **chronic systolic heart failure** secondary to multivessel CAD **exacerbation**, likely triggered by **fluid overload**.
2. **Chest pain** likely secondary to type 2 NSTEMI related to CHF **exacerbation** without ongoing **ischemic** changes.
3. Gout flare contributing to wrist pain.
4. **Atrial fibrillation** managed with apixaban.
5. **Poorly controlled diabetes** (IDDM).
6. **Chronic Kidney disease** (CKD), likely multifactorial.

Plan:

1. Medications adjusted to include Torsemide 40 mg BID, Imdur 30 mg daily, Hydralazine 25 mg TID, and Spironolactone 25 mg daily.
2. Continue apixaban, atorvastatin, and metoprolol with increased dosage to 75 mg BID.
3. Start colchicine 0.6 mg BID for gout with plans to consider **uric acid lowering therapy** after **acute flare** resolves.
4. Monitor diabetes management with decreased insulin doses.
5. Arrange follow-up with **nephrology** for CKD.
6. Discharge with rolling walker and education on monitoring weight.
7. Schedule follow-up appointments as needed.

Figure 5: Retrieval of the most similar patient case from the curated repository.

Following case locking, the system retrieves clinically similar patient records from the 36,000-case repository using the proposed hybrid similarity scoring function. Figure 5 shows the most similar patient identified for the input case. Clinically salient attributes, including symptoms, comorbidities, diagnostic findings, and interventions, are highlighted to support rapid comparison and interpretability. This retrieved case provides experiential, case-based context grounded in prior patient trajectories.

A.3 Guideline Retrieval via GraphRAG

Key Medical Guidelines (GraphRAG)

Pharmacological Interventions for Shortness of Breath

- **Heart Failure Management:**
 - **Diuretics:** To manage fluid overload, especially in patients with heart failure. Commonly used diuretics include furosemide (Lasix), typically starting at 20-40 mg IV or orally, with adjustments based on response.
 - **ACE Inhibitors:** Such as lisinopril, starting at 5-10 mg daily, to reduce blood pressure and improve heart function.
 - **Beta-Blockers:** Initiate with low doses (e.g., metoprolol 12.5-25 mg daily) to manage heart rate and improve cardiac output.
 - **ARNIs:** Consider switching to an angiotensin receptor-neprilysin inhibitor (e.g., sacubitril/valsartan) if the patient is stable and not experiencing side effects from ACE inhibitors.

Strategies for Alleviating Bilateral Leg Swelling

- **Diuretics:** Increase the dose of diuretics to manage fluid overload effectively.
- **Fluid Restriction:** Implement a fluid intake restriction to help reduce swelling.
- **Compression Stockings:** Use of compression garments to improve venous return and reduce edema.
- **Elevate Legs:** Encourage the patient to elevate her legs to promote fluid drainage.

Monitoring Parameters for Respiratory Status and Heart Function

- **Vital Signs:** Regular monitoring of blood pressure, heart rate, and respiratory rate.
- **Oxygen Saturation:** Continuous pulse oximetry to assess oxygen levels.
- **Weight Monitoring:** Daily weight checks to detect fluid retention.
- **Laboratory Tests:** Regular assessment of renal function (e.g., serum creatinine, electrolytes) and BNP levels to evaluate heart failure status.
- **Spirometry:** If feasible, to assess lung function and response to treatment.

Non-Pharmacological Measures to Improve Oxygen Saturation

- **Positioning:** Encourage sitting upright or in a semi-Fowler's position to enhance lung expansion.
- **Breathing Exercises:** Teach pursed-lip breathing and diaphragmatic breathing techniques to improve ventilation.
- **Oxygen Therapy:** Use supplemental oxygen as needed to maintain saturation above 92%.
- **Environmental Modifications:** Ensure good air quality at home, avoiding smoke and allergens that could exacerbate respiratory distress.

Figure 6: Guideline-grounded evidence retrieval through the GraphRAG pipeline.

In parallel with similar-patient retrieval, the system performs guideline-grounded retrieval using the GraphRAG pipeline. As illustrated in Figure 6, relevant clinical recommendations are identified and

summarized from authoritative WHO and NICE guidelines. Retrieved guideline content exposes structured evidence such as treatment options, contraindications, and escalation criteria.

A.4 Interactive Evidence-Aware Question Answering

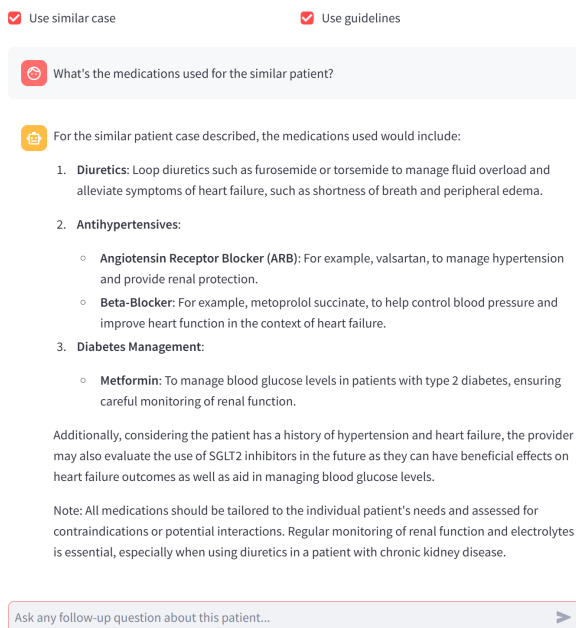


Figure 7: Interactive question answering with user-controlled evidence selection.

Figure 7 demonstrates the interactive question-answering interface. Users may selectively choose whether to incorporate similar patient cases, guideline summaries, or both into downstream reasoning. Based on the selected evidence sources, the system dynamically constructs a structured prompt combining the locked patient case and retrieved evidence, enabling transparent and controllable clinical exploration. Generated responses explicitly reflect the chosen evidence context, allowing users to examine how different information sources influence the model’s reasoning.

B Appendix B: GraphRAG Construction Details

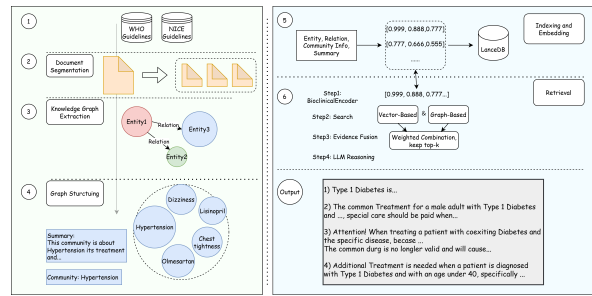


Figure 8: Detailed construction and retrieval pipeline of the GraphRAG module, including segmentation, graph extraction, indexing, and hybrid retrieval.