



Knowledge Graph Augmented Retrieval Applications in Healthcare

C. Anil Cogalan

AnalyticaHouse
Istanbul, Türkiye
anil.cogalan@analyticahouse.com

Ayca Kumluca Topalli

Department of Electrical and Electronics Engineering
Izmir University of Economics
Izmir, Türkiye
ayca.topalli@ieu.edu.tr

Abstract—This paper reviews recent advances in knowledge graph-augmented retrieval for generation in healthcare. The focus is on how structured knowledge graphs can be integrated into retrieval pipelines to improve factual grounding, transparency, and clinical applicability. By linking clinical entities and relations to verifiable sources, knowledge graph-based retrieval reduces hallucinations, improves multi-hop reasoning, and enables auditability through explicit evidence paths and citations. Reported benefits include higher accuracy in laboratory test interpretation, more precise diagnostic reasoning, and improved personalization for under-represented patient groups. Several representative systems are compared to illustrate consistent design patterns, such as short graph paths and provenance tracking, which enhance stability and safety. The review highlights measurable improvements in accuracy, explainability, and equity across diverse clinical use cases. Overall, knowledge graph-guided retrieval augmented generation emerges as a promising approach for supporting decision-making, ensuring reliable evidence citation, and fostering safer adoption of generative AI in healthcare.

Keywords—knowledge graphs; retrieval augmented generation; healthcare AI; clinical decision support; explainability; safety

I. INTRODUCTION

Generative artificial intelligence is increasingly applied to clinical information access, decision support, and patient communication.

In safety-critical use, three recurrent risks are observed. Bias refers to systematic error where performance differs across patient groups or care settings; for example, evidence for rare conditions or under-represented populations may be under-retrieved, leading to skewed suggestions.

Hallucination denotes plausible-sounding statements that are not supported by verifiable sources, typically produced when gaps in knowledge are filled by patterns rather than evidence.

Limited auditability means the provenance of facts cannot be traced, so it is unclear which guideline, study, or reasoning path produced a given statement.

Retrieval augmented generation (RAG) mitigates these risks by retrieving external evidence at query time and conditioning

the generator on that evidence [1]. When the evidence source is a curated knowledge graph (KG), a structured retrieval pipeline is formed that improves factual grounding and traceability.

In parallel, large language models benefit from KG context because nodes and typed relations constrain free-text generation with biomedical semantics.

This study examines and compares recent advancements in healthcare applications employing knowledge graph-augmented retrieval, with a focus on their applicability within the domain of clinical engineering.

II. KNOWLEDGE GRAPH-AUGMENTED RETRIEVAL (KG-RAG)

In retrieval augmented generation, relevant passages or facts are first retrieved from trusted sources and then used as context for answer generation, as depicted in Fig. 1. When a curated KG supplies the evidence, the information is organized as nodes (clinical entities such as *conditions*, *tests*, *findings*, *drugs*, and *guideline sections*) and relations (typed links such as *causes*, *associated_with*, *indicates*, *contraindicated_in*). This structure allows retrieval to focus on specific concepts and clinically meaningful links rather than broad text alone, so the model receives evidence that is both concise and verifiable.

A typical workflow proceeds in three simple steps. First, the clinical question and any patient context are mapped to KG concepts, so synonyms and abbreviations resolve to the same entity (for example, “ALP” and “alkaline phosphatase”). Second, candidate evidence is gathered by following short graph paths and by pulling passages that are already linked to those nodes, such as guideline snippets or study abstracts. Chains of one to three steps are favored, for example:

symptom → *lab abnormality* → *mechanism* → *candidate diagnosis*,

because longer paths tend to add noise. Third, the generator produces an answer that cites the sources and shows the path through the graph that supports the conclusion.

This design helps in several practical ways. Citations and visible paths make each assertion inspectable during clinical audit, reducing unsupported statements. Because entities and relations can be tailored to local practice patterns and

Knowledge Graph Augmented RAG (KG-RAG) Workflow

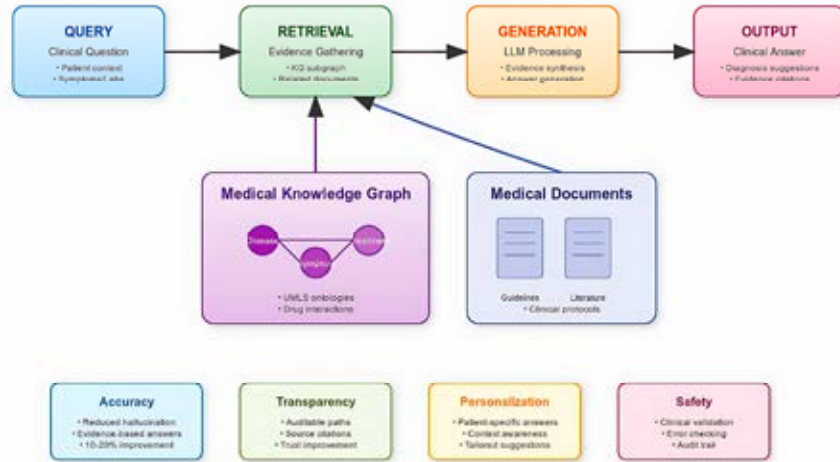


Fig. 1. End-to-end KG-RAG pipeline.

phenotypes, retrieval can surface evidence that better covers under-represented groups and rare conditions. Ambiguous terms are disambiguated early by linking to the intended KG node, which reduces off-target context before generation and improves stability in multi-hop reasoning (Fig. 2).



Fig. 2. An example of multi-hop path.

Evidence selection remains simple enough for routine use. Short graph neighborhoods and a small number of path patterns are usually sufficient, such as linking a test finding to a small set of conditions or filtering drug recommendations by contraindications and comorbidities. When documents are needed, only the most relevant, source-linked snippets are retrieved so that the generator receives compact context. Confidence can be increased by preferring evidence supported by multiple, independent sources and by de-prioritizing paths that rely on weak or outdated links.

From an operational perspective, provenance and versioning are recorded at the node and document level, so it is clear which guideline edition or study underlies a statement. Routine updates can add new entities or replace links as literature changes, while retired concepts can be preserved with time-stamps to maintain traceability. In settings where patient data are available, limited attributes (age range, key comorbidities, critical labs, etc.) can be aligned with KG entities to personalize retrieval without exposing full records.

In summary, KG-based retrieval provides a concise evidence set that is easy to audit: concepts are normalized, short paths explain why an answer follows, and citations point to verifiable sources. This combination reduces hallucination, improves coverage for overlooked subgroups, and supports multi-step reasoning with a level of transparency suitable for clinical settings.

As shown in Fig. 3, with KG context, the same question yields a more accurate and auditable answer; citations and short rationale support clinical review.

III. HEALTHCARE APPLICATIONS ENABLED BY KG-RAG

This section summarizes recent systems to illustrate clinical use, how KGs are used during retrieval, and the reported improvements.

In KRAGEN [2], complex biomedical questions are decomposed into smaller sub-questions that are answered with KG evidence and composed into a final answer with a visible reasoning graph. By embedding the domain KG and retrieving the closest nodes and relations for each sub-question, accuracy increases on multi-step multiple-choice tasks.

Example question: “In Alzheimer disease pathways, which gene interactions support mechanism X?”

The KG guidance keeps the reasoning on clinically relevant paths.

In LAB-KG [3], abnormal laboratory reports are interpreted by linking test findings to candidate conditions through a medical KG, followed by generation of case-specific explanations. Retrieval filters candidates and thresholds confidence before generation.

Example question: “Which conditions best explain elevated alkaline phosphatase with normal bilirubin?”

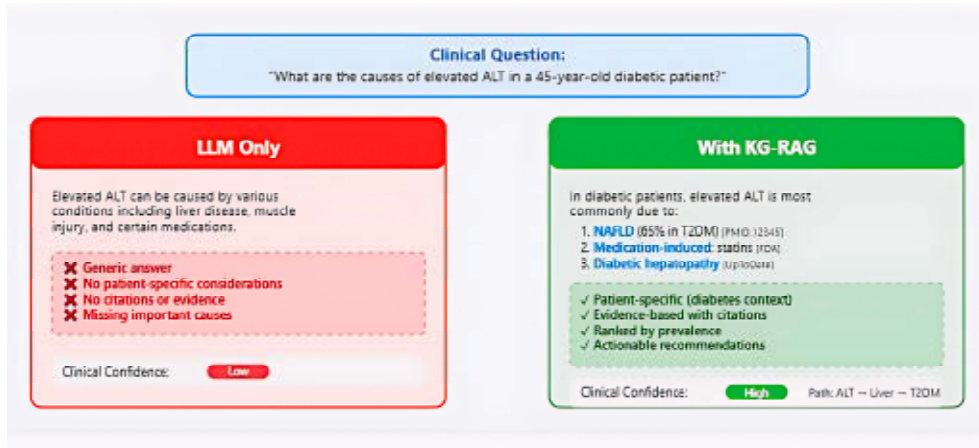


Fig. 3. LLM-only vs KG-RAG.

The KG narrows to plausible causes and supports an auditable explanation.

In Dr. Knows [4], diagnosis prediction from clinical notes is supported by retrieving knowledge paths from a UMLS (Unified Medical Language System)-based KG. Path-ranking methods identify short knowledge chains that connect patient findings to candidate diagnoses, improving concept-level F_1 and the quality of diagnostic summaries.

Example question: “Given these note snippets, which diagnoses should be considered first and why?”

The retrieved paths make the rationale inspectable.

In Medical Graph RAG [5], a graph links case entities to cited literature and standard medical definitions. Retrieval first narrows to a relevant subgraph, then refines the answer with higher-level context, improving question-answering accuracy and citation precision.

Example question: “Does drug A reduce mortality in condition B for adults over 65?”

Citations and graph links make verification straightforward.

In MedRAG [6], a hierarchical diagnostic KG is combined with similar case retrieval. When information is insufficient, targeted follow-up questions are proposed. Fine-grained diagnosis accuracy is increased, especially for smaller model sizes.

Example question: “Which clarifying questions would reduce uncertainty between diagnoses C and D?”

The KG focuses questions on discriminative findings.

Taken together, KRAGEN, LAB-KG, Dr. Knows, Medical Graph RAG, and MedRAG point to a consistent operating pattern: retrieval focuses evidence into short, meaningful graph paths; generation summarizes with citations; and the result is easier to audit in busy clinical settings. Gains are most visible in constrained tasks where ground truth is clear, which matches how such systems are typically introduced in practice.

In the same spirit, two further reviewed applications published after 2022 extend these observations in adjacent clinical settings.

In RareDxGPT [7], in rare disease cases, candidate diagnoses that general models may overlook are surfaced by targeted retrieval over a curated rare-disease knowledge resource. For each query, the most relevant entries (typical findings, associated features, discriminating clues, etc.) are provided as evidence; the generated answer is kept short and shows which facts support the conclusion. Evaluations reported higher accuracy than a general model when explanations were requested, reducing unsupported statements while keeping citations visible for clinical review.

Example question: “For an adolescent with episodic muscle weakness, what are the elevated list of CK likely rare neuromuscular disorders?”

In UMLS-guided medical dialog [8], in telemedicine dialogues, to avoid generic or vague replies, symptom–disease–test triples are retrieved from a UMLS knowledge graph and used during generation. This guidance yields more specific and auditable patient-facing responses; critical medical entities (for example, a key symptom, a likely diagnosis, a required test) are visible alongside the answer. The published study reported improvements over prior systems on MedDialog and an extended CovidDialog set in both automated metrics and expert ratings.

Example question: “For the case of a persistent cough in a smoker with low grade fever state two red flags that warrant urgent assessment and cite the relevant entity links.”

These additional applications reinforce the pattern observed above: short, auditable knowledge paths steady multi-step reasoning while keeping answers concise and citable. Benefits are concentrated in focused clinical tasks where evidence can be reviewed quickly, supporting safer adoption and simpler governance.

Table I summarizes these works in terms of their performance results.

TABLE I. PERFORMANCE COMPARISON OF RECENT KG-RAG SYSTEMS IN HEALTHCARE

System	Primary task and metric	Baseline	KG-RAG
KRAGEN [2]	Two-step multiple-choice accuracy (%)	53.1	71.8
LAB-KG [3]	Condition extraction F_1	0.56	0.76
Dr. Knows [4]	Diagnosis CUI-level F_1	20.96	26.02
Medical Graph RAG [5]	Question-answering accuracy (%)	94.0	96.4
MedRAG [6]	Fine-grained diagnosis accuracy (%)	55.0	68.0
RareDxGPT [7]	Rare-disease case vignette diagnosis accuracy (%)	23.0	43.0

IV. DISCUSSION

Evidence indicates that combining knowledge graphs with retrieval augmented generation can make answers more accurate, auditable, and usable in routine care. The strongest early results tend to appear in narrow, well-defined tasks where ground truth is clear (for example, interpretation of a small set of laboratory patterns or differential diagnosis within a single specialty). Starting with such focused use cases allows simple acceptance criteria and faster feedback from clinicians. Practical operation depends on how evidence is handled. Each statement should be tied to a cited source and, when possible, to a short path in the graph that explains why the conclusion follows. Recording the provenance (the source and its version) keeps outputs inspectable over time, and scheduled updates ensure that guideline changes are reflected quickly. When several sources disagree, both views can be presented with their citations and the confidence indicated as lower. Terms with more than one meaning are linked to the intended concept before generation so that off-target text is reduced.

Governance and safety require simple checks that fit clinical workflow. Basic bias screens can examine whether retrieved subgraphs or suggested answers differ across age, sex, or comorbidity groups. Access controls and audit logs protect sensitive data, while a short, source-linked rationale next to each answer enables quick review in busy clinics. Clear fallbacks are useful, if evidence is weak, the system can ask a clarifying question, defer to a guideline section, or decline with an explanation. Evaluation should go beyond offline accuracy. Useful operational measures include time to a usable answer, the proportion of answers with complete citations, the rate of corrections during chart review, and the stability of recommendations as new literature appears. Equity can be monitored by checking performance across under-represented cohorts. Post-deployment monitoring can detect drift, changes in data or practice that alter behavior and trigger re-review. Limitations are also straightforward. Maintaining a medical graph and keeping links current requires expert time; semi-automated curation, where candidate changes are proposed by models and approved by reviewers, can reduce this burden. Long, complex paths tend to add noise and should be avoided in favor of short, clinically meaningful chains. Finally, user trust depends on consistent interface behavior: brief answers with citations and a compact rationale generally outperform long narratives in time-constrained settings.

V. CONCLUSION

Combining a curated knowledge graph with retrieval augmented generation yields measurable gains across healthcare tasks, including higher accuracy in multi-hop question answering, clinically grounded laboratory test explanations, and more precise differential diagnosis. Beyond metrics, transparency improves through path-level rationales and citations, which supports clinical audit and helps curb hallucinations. Personalization benefits when patient context is aligned with graph semantics, while coverage for under-represented groups can improve as retrieval is guided by structured clinical entities and relations. In practice, successful adoption depends on disciplined scoping and operations. Narrow, well-defined use cases should be prioritized first, with acceptance criteria that are meaningful to clinicians and feasible to measure. Provenance and versioning of sources must be recorded so that each assertion remains inspectable. Post-deployment monitoring is needed to detect drift in content or performance, with clear escalation paths when confidence is low. Interfaces should present short, source-linked rationales rather than long free text, and uncertainty signals should be communicated in a way that fits time-constrained clinical workflows. Integration with existing systems is facilitated by standards such as Fast Healthcare Interoperability Resources (FHIR), while change-management and brief training improve trust and consistent use. Looking ahead, several directions appear promising within routine care. Multimodal knowledge graphs that connect text with imaging and genomics may improve difficult differentials. Dynamic, versioned updates can keep systems aligned with emerging literature and guidelines. Privacy-aware collaboration across sites may expand coverage without exposing sensitive data. Semi-automated curation workflows, where models propose candidates and human experts validate changes, can reduce maintenance burden. Finally, evaluation should extend beyond offline metrics to prospective and post-deployment studies that track clinical usability, error profiles, turnaround time, and equity of performance across patient subgroups.

REFERENCES

- [1] R. Yang *et al.*, "Retrieval augmented generation for generative artificial intelligence in health care," *npj Health Systems*, vol. 2, p. 2, 2025.
- [2] N. Matsumoto *et al.*, "KRAGEN: A knowledge graph-enhanced RAG framework for biomedical problem solving using large language models," *Bioinformatics*, vol. 40, no. 6, 2024.
- [3] R. Guo *et al.*, "LAB-KG: A retrieval augmented generation method with knowledge graphs for medical lab test interpretation," in *Bridging Neurons and Symbols @ COLING 2025*, 2025, pp. 40–50.
- [4] Y. Gao *et al.*, "Leveraging medical knowledge graphs into large language models for diagnosis prediction," *JMIR AI*, vol. 4, e58670, 2025.
- [5] J. Wu *et al.*, "Medical Graph RAG: Towards safe medical large language models via graph retrieval augmented generation," *arXiv preprint*, 2024.
- [6] X. Zhao *et al.*, "MedRAG: Enhancing retrieval augmented generation with knowledge graph elicited reasoning for healthcare copilot," *arXiv preprint*, 2025.
- [7] C. Zelin, W. K. Chung, M. Jeanne, G. Zhang, and C. Weng, "Rare disease diagnosis using knowledge guided retrieval augmentation for ChatGPT," *Journal of Biomedical Informatics*, vol. 157, 104702, 2024.
- [8] D. Varshney, A. Zafar, N. K. Behera, and A. Ekbal, "Knowledge graph assisted end-to-end medical dialog generation," *Artificial Intelligence in Medicine*, vol. 139, 102535, 2023.