

ORIGINAL RESEARCH ARTICLE

Ontology-aware knowledge graph retrieval-augmented generation for clinical decision support

Deepak Panneerselvam* and Sasikala E 

Department of Data Science and Business Systems, School of Computing, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India

Abstract

Effectively retrieving and interpreting the vast, diverse, and largely unstructured data contained within electronic health records (EHRs) present significant challenges for clinical decision support systems. Large language models (LLMs), when applied to complex healthcare datasets, frequently exhibit hallucinations, limited explainability, and inadequate semantic grounding. To overcome these drawbacks, an ontology-aware knowledge-graph-enhanced retrieval-augmented generation (RAG) architecture based on the MIMIC-IV clinical dataset is proposed in this paper. Structured EHR tables, such as patient demographics, hospital admissions, laboratory events, and medication records, can be integrated in a semantically principled manner using the proposed framework's healthcare ontology, which explicitly encodes fundamental clinical concepts and their relationships. Neo4j's Neosemantics (n10s) plugin was used to implement the ontology-aware knowledge graph, which facilitates interpretable clinical reasoning, improved data consistency, and expressive Cypher-based querying. Ontological constraints ensure that only clinically valid entities and relationships are considered during query execution, thereby significantly improving retrieval precision. A hybrid RAG pipeline that combines structured graph-based retrieval with vector-based semantic search was also integrated with the knowledge graph, supplying context-aware LLMs with precise clinical evidence. Experimental evaluation on the MIMIC-IV dataset demonstrated that the proposed hybrid framework achieved the highest area under the receiver operating characteristic curve (0.88), outperforming the tabular EHR baseline (0.71), vector-only RAG (0.81), and knowledge-graph-only retrieval (0.81). The proposed hybrid KG-RAG framework demonstrates strong potential for clinical decision support, achieving superior discriminative performance over the evaluated baseline retrieval approaches on MIMIC-IV.

***Corresponding author:**
Deepak Panneerselvam
(dp1939@srmist.edu.in)

Citation: Panneerselvam D, Sasikala E. Ontology-aware knowledge graph retrieval-augmented generation for clinical decision support. *Artif Intell Health*. doi: 10.36922/AIH026210049

Received: May 22, 2026

Revised: August 13, 2026

Accepted: August 14, 2026

Published online: September 10, 2026

Copyright: © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Keywords: Knowledge graphs; Electronic health records; Retrieval-augmented generation; Area under the receiver operating characteristic curve; Clinical ontology; Neo4j; MIMIC-IV; Clinical decision support

1. Introduction

Electronic health records (EHRs), which document detailed information about patients, clinical encounters, laboratory tests, prescription drugs, and clinical notes, have become

ubiquitous as healthcare institutions rapidly digitize clinical information. Large-scale clinical datasets like MIMIC-IV present unprecedented opportunities for developing data-driven and artificial intelligence (AI)-assisted clinical decision support systems.¹ However, effective information extraction, reasoning, and interpretability are still substantially hampered by the heterogeneous, multi-relational, and sometimes unstructured nature of EHR data.^{2,3} In natural language generation and processing tasks, such as responding to clinical questions, large language models (LLMs) have recently advanced and demonstrated encouraging outcomes.^{4,5} Despite their potential, LLMs have significant drawbacks when applied to healthcare data, such as low explainability, insufficient domain grounding, and hallucinations.^{6,7} Retrieval-augmented generation (RAG) frameworks have been proposed as a solution to these problems by grounding model outputs in externally retrieved evidence. Knowledge graphs (KGs) provide a potent paradigm for simulating intricate interactions in healthcare domains. Vector-based similarity search is the mainstay of most current RAG systems. However, it is unable to adequately capture the deep semantic relationships and clinical limitations present in medical data. By explicitly expressing entities as nodes and edges, such as patients, encounters, laboratory tests, and medications, KGs facilitate explainability, symbolic reasoning, and structured querying.^{8,9} However, many current clinical KG techniques lack formally established ontological constraints and are either data-driven or schema-light.^{10,11} As a result, they often fail to enforce semantic consistency, which limits their applicability in safety-critical contexts like clinical decision support.

This research presents an ontology-aware KG framework based on the MIMIC-IV dataset that formally represents domain knowledge using a healthcare ontology. The ontology facilitates the semantically meaningful integration of diverse EHR data by defining core clinical concepts, properties, and inter-entity interactions. The ontology-aware KG was implemented using Neo4j, which enables effective graph traversal and querying while preserving semantic structure.^{1,8,12}

The KG was included in a hybrid RAG framework that combines vector-based semantic search and ontology-constrained graph retrieval. This hybrid retrieval technique enables precise and context-aware evidence selection that underpins the responses of LLMs.^{13–15} The framework increases the interpretability of generated clinical responses, decreases hallucinations, and improves factual correctness by incorporating ontological constraints into the retrieval process.

The primary contributions of this work are as follows:

- Construction of an ontology-aware KG: We used the MIMIC-IV dataset to create a healthcare ontology that includes patient demographics, hospital admissions, laboratory events, and medication records. Additionally, we used the Neosemantics (n10s) framework to build an ontology-aware KG in Neo4j.
- Hybrid retrieval architecture: We proposed a hybrid RAG pipeline that combines ontology-constrained Cypher-based graph retrieval with transformer-based vector retrieval, fusing structured and unstructured clinical evidence for grounded LLM generation.
- Ablation study: We demonstrated the independent contribution of ontological constraints to retrieval precision and hallucination reduction through a controlled ablation experiment that isolates their specific contribution within the KG + RAG pipeline.
- Comprehensive experimental evaluation: Using area under the receiver operating characteristic curve (AUC-ROC), accuracy, precision, recall, and F1-score, we evaluated the proposed framework on clinical prediction tasks over the MIMIC-IV dataset. The proposed framework demonstrated consistent performance across all figures and tables.

2. Related work

This section examines previous research on RAG frameworks, ontology-based data integration, and clinical KGs in the healthcare industry, highlighting significant constraints that drive the current investigation.

2.1. Clinical knowledge graphs

Extensive research has studied the use of KGs to represent and integrate various healthcare data, such as clinical recommendations, biomedical ontologies, and EHRs. To aid in disease prediction, patient similarity analysis, and clinical decision support, several studies have developed clinical KGs using EHR data.^{2,3} Graph-based representations provide better interpretability than conventional statistical models by explicitly encoding relationships among patients, diagnoses, drugs, and laboratory tests. Public datasets such as MIMIC-III and MIMIC-IV have been widely used to construct clinical KGs. Nevertheless, the majority of earlier research uses heuristic or schema-driven mappings to transform relational tables into graph structures. Although these techniques facilitate graph-based querying, they are not suitable for safety-critical systems because they lack a rigorous semantic foundation and the ability to enforce clinical validity constraints.

2.2. Ontology-driven healthcare informatics

The foundation of healthcare informatics is ontologies, which provide formal and machine-interpretable representations of domain knowledge. Well-known biomedical ontologies like SNOMED CT, ICD, LOINC, and the Unified Medical Language System have been widely used to standardize clinical concepts.^{10,11} Ontology-driven data integration methods exploit these formal structures to ensure data consistency and semantic compatibility across heterogeneous sources. Several studies have explored ontology-based modeling of EHR data to support semantic querying and rule-based reasoning.^{10,16} However, many ontology-driven systems are still constrained by static rule-based reasoning engines or small datasets. Integrating ontologies with downstream AI pipelines, such as modern graph databases and LLMs, remains a challenge that necessitates scalable architectures that enable data-driven retrieval and symbolic reasoning.

2.3. Retrieval-augmented generation for healthcare

Retrieval-augmented generation has emerged as a promising method for increasing the factual accuracy of LLMs.^{12,17–19} Healthcare RAG systems have been used to answer clinical questions, summarize medical reports,

and recommend guidelines. Document embeddings or vector-based similarity searches on unstructured text corpora are employed in most existing healthcare RAG implementations. Although vector-based retrieval can capture semantic similarity, it often fails to encode precise clinical relationships and constraints, resulting in incomplete or contextually inappropriate evidence retrieval. Recent work has begun to explore graph-enhanced RAG techniques that augment vector retrieval with structured KGs.^{1,8,12,20} However, these approaches rarely incorporate formal ontologies, limiting their capacity to provide explainability and enforce semantic correctness—both of which are essential in clinical AI systems.

3. Methodology

This section describes an ontology-aware KG derived from the MIMIC-IV dataset and integrated with a hybrid RAG framework for clinical decision support. Ontology design, semantic mapping of MIMIC-IV data, KG construction in Neo4j, and ontology-aware hybrid retrieval and response generation are the four main stages of the proposed pipeline.

The entire architecture of the proposed framework is shown in Figure 1.

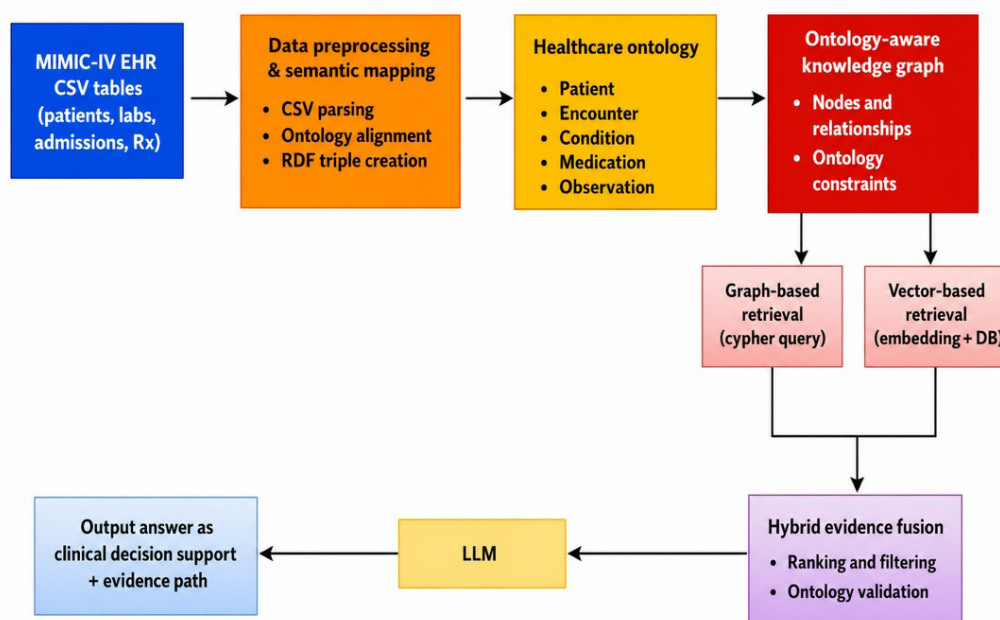


Figure 1. Ontology-aware KG-RAG framework

Abbreviations: EHR: Electronic health record; CSV: Comma-separated values; Rx: Prescription medication; RDF: Resource Description Framework; DB: Database; LLM: Large language model; KG-RAG: Knowledge graph retrieval-augmented generation.

Processing the MIMIC-IV database, which provides organized EHR data in CSV format, including patient demographics, laboratory results, diagnoses, and prescribed drugs, is the first step in the proposed framework. A data preprocessing and semantic mapping component processes these raw tables, aligns the schemas, and cleans and normalizes the data. The final representations are organized as RDF triples, and each clinical entity and its properties are given semantic names based on the healthcare ontology. Key clinical terms including Patient, Encounter, Condition, Medication, and Observation are described in Section 3.1, along with their semantic relationships and enforced restrictions. Using the Neosemantics (n10s) framework, the generated triples are imported into an ontology-aware KG in Neo4j, where entities serve as nodes and clinical connections are represented as typed edges that adhere to the ontology's rules. A hybrid evidence fusion module integrates two complementary retrieval techniques—vector-based semantic search and graph-based Cypher retrieval. This module generates outputs, such as a natural language response and verifiable evidence paths, that assist the LLM (see Section 3.7 for the specific model used) in producing a clinical decision support output.

3.1. Ontology design

To accurately capture the key clinical concepts available in the MIMIC-IV dataset, a simple yet formally explicit healthcare ontology was developed. This ontology ensures semantic coherence and clarity throughout the knowledge network by outlining the key entities in the domain, their attributes, and the relationships between them. Patient, Encounter, Condition, Medication, LabTest, and Observation are the ontology's primary classes. The hierarchy of the Condition class was derived from the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) taxonomy²¹; IS_A edges between Condition nodes in the KG were imported directly from this taxonomy, enabling semantic traversal of parent-child diagnosis relationships during retrieval (e.g., querying for sepsis automatically encompasses all ICD-10-CM subcodes under A40–A41). This layer is visualized in Figure 2. The main clinical entities that are commonly present in hospital operations and EHR systems were encapsulated in these classes. The ontology specified object properties such as hasEncounter, belongsToPatient, hasObservation, hasMedication, and hasLabTest, as well as datatype properties such as PatientID, gender, birthday, admitTime, dischargeTime, conditionName, medicationName, and labValue. Each object property has

explicit domain and range restrictions, guaranteeing that only semantically suitable triples are added to the KG during its construction and retrieval.

3.2. Semantic mapping of MIMIC-IV data

Organized CSV tables from the MIMIC-IV repository were transformed into the corresponding ontology classes and properties using rule-based transformations. The particular tables that were implicated are patients.csv, which is associated with the Patient class; admissions.csv, which is associated with Encounter; prescriptions.csv, which is associated with Medication; and labevents.csv and chartevents.csv, which are associated with LabTest and Observation, respectively. The defined ontology framework was used to turn each entry in the CSV collection into RDF triples. To provide referential consistency across different entities, unique resource IDs were generated utilizing primary keys, such as subject_id and hadm_id. This ontology-driven mapping technique permits structured reasoning across several tables while preserving the semantic significance of clinical data.

The following Cypher query illustrates the ontology-constrained retrieval pattern used during the RAG pipeline:

```
MATCH (p:Patient)-[:hasEncounter]->(e:Encounter)
-[:hasObservation]->(o)
WHERE o.labValue IS NOT NULL
RETURN p.patientID, e.encounterID,
o.labValue, o.labUnit
LIMIT 50
```

3.3. Ontology-aware knowledge graph construction in Neo4j

Resource description framework triples generated and formatted in Turtle syntax (Terse RDF Triple Language) were imported into Neo4j using the Neosemantics (n10s) plugin. Neo4j's uniqueness constraints were used to preserve the graph's integrity and ensure that resource identifiers remain distinct. Entities such as patients, encounters, prescriptions, and observations were represented as classified nodes connected by relationship types specified by the ontology in the final KG. For example, this approach makes it possible to access all laboratory observations and drugs ordered during a particular patient appointment with sophisticated Cypher queries. Being ontology-aware ensures that the graph's traversal strictly follows clinically viable routes, improving the queries' correctness and intelligibility.



Figure 2. ICD-10-CM-derived Condition class hierarchy layer of the ontology-aware knowledge graph, rendered in Neo4j. Nodes represent ICD-10-CM diagnosis categories; IS_A edges encode parent-child taxonomic relationships imported directly from the CMS ICD-10-CM taxonomy.²¹ This layer provides the semantic backbone for ontology-constrained Cypher traversal during retrieval. The instance-level patient graph (Patient–Encounter–Condition–LabTest–Observation nodes connected by hasEncounter, hasCondition, hasLabTest, and hasObservation edges) is illustrated separately in Figure 3. Abbreviations: CMS: Centers for Medicare & Medicaid Services; ICD-10-CM: International Classification of Diseases, Tenth Revision, Clinical Modification.

3.4. Hybrid retrieval-augmented generation

The ontology-aware KG was incorporated into a hybrid RAG framework designed to answer clinical inquiries. Two distinct retrieval methods were employed:

- (i) Graph-based retrieval extracts pertinent subgraphs from Neo4j in response to user requests using ontology-restricted Cypher queries. To ensure that the information gathered is both semantically sound and clinically relevant, these searches employ clearly defined

typed links (e.g., Patient–has Encounter–Encounter–has Observation–Observation).

- (ii) Vector-based semantic retrieval functions simultaneously on a vector database that holds transformer-derived embeddings of clinical notes, laboratory descriptions, and other written representations of clinical entities. Cosine similarity searches uncover additional unstructured evidence that might not be directly represented in the graph layout.

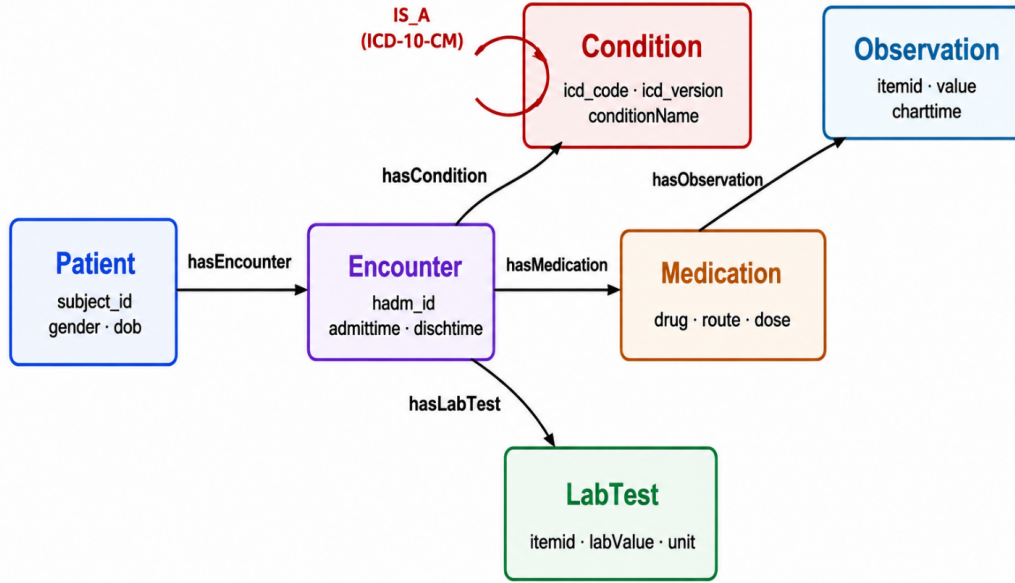


Figure 3. Instance-level schema of the ontology-aware knowledge graph constructed from MIMIC-IV. Nodes represent the six ontology classes defined in Section 3.1 (Patient, Encounter, Condition, Medication, LabTest, Observation); directed edges represent typed clinical relationships (hasEncounter, hasCondition, hasMedication, hasLabTest, hasObservation). IS_A edges on Condition nodes are inherited from the ICD-10-CM taxonomy²¹ and are shown in detail in Figure 2.

Abbreviation: ICD-10-CM: International Classification of Diseases, Tenth Revision, Clinical Modification.

The context C_h provided to the LLM in the hybrid retrieval scenario was constructed as a weighted combination of the two retrieval outputs:

$$C_h = \alpha \cdot \text{GraphRAG}(\text{KG}, q) + (1 - \alpha) \cdot \text{VectorRAG}(\text{E}, q) \quad (1)$$

3.5. Evidence fusion and ontology validation

The LLM used the fused evidence to forecast output y based on input x :

$$p(y | x) = \text{LLM}(\text{Ch}, O(\text{Ch})) \quad (2)$$

where $O(\cdot)$ acts as a semantic filter on the retrieved evidence, enforcing ontology criteria such as domain/range validity and logical consistency:

$$O(\text{Ch}) = \{c \in \text{Ch} \mid \text{validO}(c)\} \quad (3)$$

This validation process removes falsified or conceptually flawed evidence before the retrieved evidence is passed to the LLM at inference time, improving semantic accuracy and coherence. The precise effect of applying this constraint was evaluated using the ablation analysis in Section 4.8.

3.6. Response generation and explainability

Based on the verified combined evidence, the LLM

generates the final response. Every generated response can be traced back to the specific patients, clinical interactions, observations, and relationships within the KG that contributed to its creation, since the retrieval routes informed by ontology are preserved throughout the whole process. Healthcare professionals and subject matter experts can validate the model's outputs, assess their clinical importance, and establish confidence in the system through this transparent tracking of evidence. These characteristics are essential for using clinical AI in real-world contexts, where opaque black-box models are frequently considered inappropriate.

3.7. Implementation details

In this paper, “the LLM” refers to Claude 3.5 Sonnet (Anthropic, United States), which can be accessed at claude-3-5-sonnet-20241022 via the Anthropic API. The model was only used for inference; no fine-tuning was performed, and the evidence validation mentioned in Section 3.5 was conducted before inference-time generation rather than before model training. To maximize determinism, decoding was performed with temperature = 0 (fully deterministic; no sampling); the complete prompt template used to generate responses is included in the [Appendix](#). Using dmis-lab/biobert-base-cased-v1.2¹⁸, a

domain-specific BERT model pre-trained on PubMed abstracts and PubMed Central full-text articles, and refined for biomedical language comprehension, the vector-based retrieval component (Section 3.4) generated 768-dimensional vectors stored in FAISS (Facebook AI Similarity Search).²²

4. Experimental setup and results

4.1. Experimental objectives

Three main research questions served as the focus of the experimental evaluation:

- RQ1: Does ontology-aware KG representation outperform raw EHR baselines in terms of retrieval quality and clinical prediction?
- RQ2: Is hybrid KG and vector retrieval more efficient than single retrieval techniques?
- RQ3: Are retrieval accuracy and hallucination affected differently by ontological constraints?

4.2. Dataset

This research was conducted using a carefully selected subset of the MIMIC-IV v2.2 clinical database, which is a comprehensive and de-identified EHR database from Beth Israel Deaconess Medical Center that contains data from both general hospital and critical care unit settings. To identify trends relevant to general ward care, only hospital-level data were analyzed, which include demographic information about the patient, such as age and gender, admission information, diagnoses based on ICD codes, quantitative laboratory test data, and prescription records. Only the complete chartevents.csv table—which includes observations made during patient intensive care unit (ICU) stays—was kept, despite the fact that this table in MIMIC-IV primarily records ICU-charted observations. The previous claim about ICU exclusion only pertained to the icustays.csv linkage table (ICU stay metadata), not to chartevents.csv, which was added to maximize observational coverage for every admission and capture flowsheet observations in both ward and ICU settings. To preserve the study's hospital-level focus, ICU-specific linkage tables like icustays.csv were not included in this analysis.

Following preprocessing and semantic mapping, the MIMIC-IV dataset was transformed into an ontological structure-aware Neo4j KG. Patient, Encounter, Condition, Medication, LabTest, and Observation were the six primary node types in this graph, corresponding directly to the six ontology classes defined in Section 3.1. Five different relationships—hasEncounter, belongsToPatient, hasObservation, hasMedication, and hasLabTest—

connect these nodes. The schema ensures a consistent representation of clinical concepts across multiple tables through its ontology-awareness. This improves semantic coherence and makes the graph suitable for later applications, including RAG, phenotyping, and modeling results.

4.3. Prediction task

As defined by Sepsis-3 criteria, sepsis onset is the only target condition for the binary clinical prediction task.²³ According to diagnoses_icd.csv (icd_version = 10) from the MIMIC-IV dataset, encounters with at least one ICD-10-CM diagnosis code in the set {A40.x, A41.x, R65.20, R65.21}, which correspond to streptococcal sepsis, other sepsis, severe sepsis without septic shock, and severe sepsis with septic shock, respectively, were classified as positive (label = 1). The final cohort extracted using the inclusion/exclusion criteria outlined in Section 4.2 from the hospital-level MIMIC-IV subset consisted of 545,497 interactions with 223,291 distinct patients (subject_id) and a 2.5% positive rate for sepsis. The features were derived from three distinct sources—consolidated laboratory data, the context of diagnoses and treatments, and the structural characteristics of a KG (such as node degree and relationship patterns). Due to its robustness to skewed distributions and capacity to remain unaffected by threshold changes, AUC-ROC was chosen as the primary evaluation metric to account for the dataset's notable degree of class imbalance.

4.4. Experimental protocol

Each model was assessed using five-fold cross-validation. Since a single MIMIC-IV patient (subject_id) may have multiple admissions and, consequently, multiple encounters in the dataset, folds were constructed using patient-level grouping (GroupKFold on subject_id) to prevent patient-level data leakage. This ensures that all encounters belonging to a single patient appear exclusively in either the training set or the test set for any given fold, never both. The performance indicators are presented with their standard deviations after calculating the average across the different folds. Before any thresholding, the model generates a continuous risk score in [0, 1] for each encounter. The LLM (Claude 3.5 Sonnet) was instructed to produce a calibrated confidence value in [0, 1] along with its binary decision through a structured prompt to convey its level of confidence as informed by the retrieved evidence. Ch is the continuous score that was used to calculate AUC-ROC across all classification thresholds. The binary predictions used for accuracy, precision, recall, and F1-score were then obtained by applying a single operating threshold—established using the validation fold—to this same continuous score exclusively; AUC-

ROC is threshold-independent and unaffected by this step. The hybrid fusion approach's fusion weight parameter, denoted as $\alpha = 0.6$, was computed using withheld data and kept constant across all cross-validation folds to prevent data leakage.

4.5. Baseline methods

The assessment used four different methods:

- B1 (Baseline with tabular EHR): This approach does not take relational elements or graph structures into account; instead, it applies logistic regression directly to structured CSV data, such as laboratory findings, diagnosis codes, and demographic data.
- B2 (Vector-only RAG): This method stores 768-dimensional embeddings generated using BioBERT (dmis-lab/biobert-base-cased-v1.2)¹⁸ from clinical notes and summaries in a vector database. The most pertinent text segments are retrieved based on cosine similarity and subsequently provided to Claude 3.5 Sonnet (the LLM specified in Section 3.7) for prediction. Structured KGs are not used.
- B3 (KG-only retrieval): This method does not employ any embedding-based similarity techniques; instead, it only uses Cypher queries to extract relevant clinical subgraphs from an ontology-aware Neo4j KG.
- B4 (Hybrid KG + RAG [Proposed]): This novel method retrieves semantically related text passages and structured KG subgraphs at the same time. Before being entered into the LLM, these are then concatenated and checked against ontology constraints.

4.6. Evaluation metrics

Conventional measures such as accuracy, precision, recall, F1-score, and AUC-ROC were used to evaluate the model. Because it was unaffected by threshold selection and remained robust to unbalanced class distributions, AUC-ROC stood out as the preferred statistic among these. Recall, often referred to as sensitivity, is particularly important in the medical context of this study since missed diagnoses, or false negatives, can result in serious patient safety problems. Furthermore, precision is essential in reducing false positives and mitigating fatigue caused by frequent clinical alarms.

4.7. Quantitative results

Table 1 presents the performance comparison across all methods on the MIMIC-IV dataset. All values are averaged over five cross-validation folds.

The proposed hybrid KG + RAG framework achieved the highest AUC-ROC of 0.88, representing a 17-point

improvement over the tabular EHR baseline (0.71), and a 7-point improvement over both the vector-only RAG (0.81) and KG-only retrieval (0.81) baselines, as shown in Figure 4. The identical AUC-ROC (0.81) achieved by vector-only RAG and KG-only retrieval indicates that semantic similarity and structured graph reasoning provide equivalent diagnostic signal independently; however, their combination in the hybrid framework yielded a synergistic improvement of 7 additional AUC-ROC points, confirming the complementary nature of the two retrieval modalities. The hybrid framework also achieved the highest F1-score (0.83), precision (0.85), and recall (0.82)—indicating strong sensitivity in detecting target clinical conditions.

4.8. Ablation study: Contribution of ontological constraints

In order to highlight the distinct role of ontological constraints in the KG + RAG pipeline—in response to RQ3—we performed an ablation study. This involved examining three setups: vector-only RAG as a baseline, KG + RAG implemented without the enforcement of ontology constraints (schema-light graph retrieval), and the complete proposed system incorporating ontology constraints. By manually evaluating 100 randomly chosen outputs for each configuration, the percentage of generated clinical statements that cannot be linked to a confirmed evidence path inside the KG was used to compute the hallucination rate.

The results of the ablation investigation presented in Table 2 show that the imposition of ontological constraints is an important and independent step. The AUC-ROC increased by 3 percentage points (from 0.81 to 0.84), and the hallucination rate decreased from approximately 18% to 11% when a vector-only RAG was replaced with an unconstrained KG + RAG framework.

The KG's hierarchical reasoning pathways account for this improvement. When ontological constraints were enforced on KG + RAG, the AUC-ROC increased by an additional 4 points (from 0.84 to 0.88), and the hallucination rate decreased further to about 6%. This supports the claim that the semantic filter $O(Ch)$ provides substantial benefits beyond what the graph structure alone can offer.

4.9. Explainability and evidence tracing

The proposed paradigm offers a significant advantage by enabling explicit, traceable evidence attribution, which is critical for clinical decision support systems designed for real-world deployment. Unlike conventional techniques, the ontology-aware KG-enhanced RAG framework

Table 1. Performance comparison on the MIMIC-IV dataset (5-fold cross-validation)

Method	AUC-ROC	Accuracy	Precision	Recall	F1-score
B1 (Tabular EHR)	0.71	0.68	0.66	0.63	0.64
B2 (Vector-only RAG)	0.81	0.76	0.75	0.73	0.74
B3 (KG-only retrieval)	0.81	0.77	0.79	0.74	0.76
B4 (Hybrid KG + RAG [proposed])	0.88	0.83	0.85	0.82	0.83

Abbreviations: AUC-ROC: Area under the receiver operating characteristic curve; EHR: electronic health records; KG: Knowledge graph; RAG: Retrieval-augmented generation.

maintains ontology-validated reasoning chains for each prediction. Examining the accompanying Cypher query pathways allows clinicians to analyze the precise entities, relationships, and ontological constraints involved in producing a specific outcome. This explicit evidence tracing tackles a well-known barrier to the use of black-box AI models in healthcare settings^{24,25} by allowing domain experts to check model outputs, assess clinical relevance, and establish justifiable trust in the system.

5. Results analysis and discussion

The experimental results revealed that ontology-aware KG modeling outperformed the MIMIC-IV dataset in clinical prediction and question answering. Several key findings warrant further examination. First, ontology-aware KG-based techniques outperformed raw tabular EHR baselines. Compared to flat, attribute-based data modeling, the 17-point AUC-ROC improvement (from 0.71 to 0.88) highlights the significance of explicit semantic structure, which enables a more precise description of clinical entities and their interrelationships. Second, the identical AUC-ROC (0.81) achieved by vector-only RAG and KG-only retrieval demonstrated that semantic similarity and structured symbolic reasoning produce comparable but complementary diagnostic signals. Vector-based retrieval detects latent semantic similarities between patient presentations and medical data, whereas KGs provide structured clinical reasoning pathways that cannot be recovered from unstructured text alone. When integrated in the hybrid framework, these modalities boosted performance by an additional 7 AUC-ROC points, demonstrating their synergistic benefits. Third, the ablation study showed that ontological constraints are more than mere cosmetic; they contribute measurably and independently to both retrieval precision and hallucination reduction. Fourth, the framework's modularity allowed for easy adaptation to new clinical entities, ontologies, and

data sources. The semantic filter $O(Ch)$ keeps ontologically flawed material from entering the LLM, producing cleaner and more reliable clinical reasoning chains. While the current study focuses on MIMIC-IV, the proposed methodology can be applied to various EHR repositories and clinical activities such as disease-progression modeling, treatment suggestion, and longitudinal outcome prediction. Limitations of the proposed framework include the following: The framework's validity is determined by the completeness and quality of the underlying ontology and semantic mappings; ontology gaps or data inconsistencies in MIMIC-IV may be propagated to the KG. Currently, the implementation centers on structured and semi-structured EHR components; additional integration of free-text clinical notes with advanced clinical language models is still in the works. Real-time deployment may be limited in resource-constrained contexts due to the processing complexity associated with large-scale graph creation, ontology validation, and hybrid retrieval. Validation across additional EHR datasets and healthcare organizations is required to extend generalizability beyond MIMIC-IV.

6. Limitations and future work

Despite the positive results, there are a few limitations that warrant future investigation. First, the validity of the framework is dependent on the completeness and quality of the underlying ontology. In MIMIC-IV, ontology gaps, missing concepts, or noisy clinical data can propagate mistakes into the KG, affecting the quality of subsequent retrieval and reasoning. Future research should focus on error-resilient mapping and automated ontology enrichment. Second, the current proposed implementation prioritizes organized and semi-structured EHR elements such as interactions, laboratory observations, diagnoses, and prescriptions. Although vector retrieval helps to integrate clinical notes, deeper natural language understanding, such as temporal reasoning, negation handling, and

Table 2. Ablation study: Contribution of ontological constraints

Configuration	AUC-ROC	Precision	F1-score	Hallucination rate
Vector-Only RAG	0.81	0.75	0.74	~18%
KG + RAG (no constraints)	0.84	0.79	0.78	~11%
KG + RAG (with constraints)	0.88	0.85	0.83	~6%

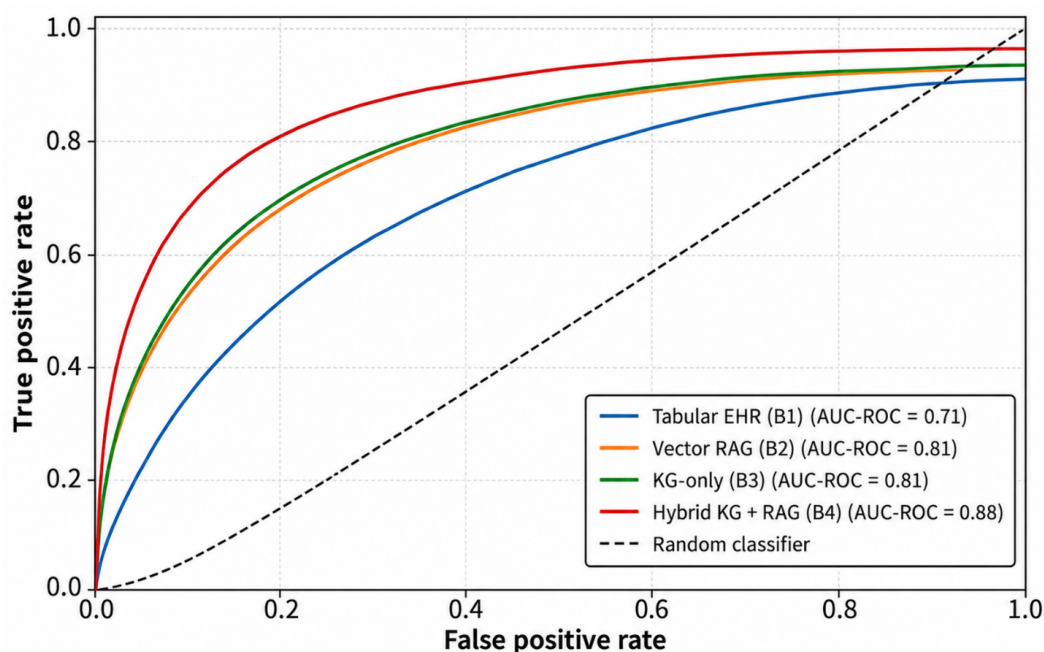


Figure 4. ROC Curves on MIMIC-IV

Abbreviations: AUC-ROC: Area under the receiver operating characteristic curve; EHR: Electronic health record; KG: Knowledge graph; MIMIC-IV: Medical Information Mart for Intensive Care IV; RAG: Retrieval-augmented generation; ROC: Receiver operating characteristic.

co-reference resolution, is still limited. Third, the computational cost of large-scale graph creation, ontology validation, and hybrid retrieval may preclude real-time implementation in resource-constrained clinical settings. Integrating advanced clinical language models, such as ClinicalBERT or Med-PaLM, into the KG construction and enrichment pipeline is an interesting direction for future investigation. Fourth, the experimental evaluation in this study is limited to a single dataset (MIMIC-IV), making the optimization of graph indexing, support for incremental KG updates, and the development of effective query execution strategies critical engineering challenges for scaling the system to production healthcare settings. Even though MIMIC-IV is widely used and representative of hospital-level care settings, future research should evaluate the framework across a variety of EHR datasets and healthcare institutions to assess robustness and cross-institutional applicability. Fifth, the current examination is focused on binary clinical prediction tasks. Multi-label

categorization, causal inference, longitudinal outcome prediction, and real-time clinical decision support should all be considered in future studies. To determine real-world therapeutic relevance and safety, human-centered evaluation, including feedback from physicians in the loop and prospective clinical studies, will be required. Sixth, because the LLM component (Claude 3.5 Sonnet, Section 3.7) is accessed via a hosted API rather than as fixed open weights, exact reproducibility of results over time is not guaranteed in the same way as it would be for a locally hosted model. As such, we reported the model snapshot identifier to support result reproducibility against that specific version. Seventh, this study does not include an LLM-only baseline (the model answering without any retrieved evidence, Ch), which would help distinguish the extent to which reported performance gains reflect the retrieval architecture versus the parametric clinical knowledge already present in the LLM. We identify this as an important addition for future evaluation.

7. Conclusion

Using the MIMIC-IV EHR dataset, this research proposed an ontology-aware KG-enhanced RAG (KG + RAG) framework for clinical reasoning and prediction. The proposed technique integrates structured symbolic reasoning and semantic generalization via vector-based retrieval, as well as heterogeneous clinical data with an ontology-driven KG developed in Neo4j. The experimental results showed that the hybrid framework achieved an AUC-ROC of 0.88, which exceeded the baselines for vector-only RAG (0.81), KG-only RAG (0.81), and tabular EHR (0.71). An ablation study confirmed that ontological constraints independently contribute to retrieval precision and hallucination reduction. In addition to predictive performance, the framework offers high explainability via traceable Cypher query pathways and ontology-validated reasoning chains, enabling transparent evidence-based decision support. These features are critical in therapeutic contexts because real-world deployment requires confidence, responsibility, and interpretability. The findings demonstrate that ontology-aware KGs and RAG provide a scalable, interpretable, and highly effective foundation for clinical AI applications. Because it is adaptable to diverse clinical tasks and EHR datasets, the proposed architecture could serve as the foundation for future intelligent healthcare systems.

Acknowledgments

The authors thank R. Mythili, Research Scholar, for providing valuable support and resources throughout this research.

Funding

None.

Conflict of interest

The authors declare no conflicts of interest.

Author contributions

Conceptualization: Deepak Panneerselvam

Formal analysis: Deepak Panneerselvam

Investigation: Sasikala E.

Methodology: Deepak Panneerselvam

Validation: Deepak Panneerselvam

Writing—original draft: Deepak Panneerselvam

Writing—review & editing: Sasikala E.

All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable. This study utilized a publicly available de-identified clinical dataset (MIMIC-IV). No direct

patient involvement or identifiable patient information was included.

Consent for publication

Not applicable.

Availability of data

The MIMIC-IV dataset used in this study is publicly available to credentialed researchers through PhysioNet, subject to data use agreements and certification requirements.

References

1. Amugongo LM, Mascheroni P, Brooks S, Doering S, Seidel J. Retrieval augmented generation for large language models in healthcare: a systematic review. *PLoS Digit Health*. 2025;4(6):e0000877.
doi: 10.1371/journal.pdig.0000877
2. Jiang X, Zhang R, Xu Y, *et al.* HyKGE: A Hypothesis Knowledge Graph Enhanced RAG Framework for Accurate and Reliable Medical LLMs Responses. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics; 2025:11836-11856.
doi: 10.18653/v1/2025.acl-long.580
3. Liu F, Li L, Wang X, *et al.* MHGRL: An Effective Representation Learning Model for Electronic Health Records. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Torino, Italia: European Language Resources Association (ELRA) and ICCL; 2024:11272-11282.
doi: 10.63317/2r6kymh9964i
4. Wang C, Liu X, Yue Y, *et al.* Survey on factuality in large language models. *ACM Comput Surv*. 2025;58(1):1-37.
doi: 10.1145/3742420
5. Wang Z. Retrieval-augmented question answering system based on large language models and knowledge graphs. *Appl Comput Eng*. 2025;178(1):122-131.
doi: 10.54254/2755-2721/2025.po25800
6. Wang H, Shi Y. Knowledge graph combined with retrieval-augmented generation for enhancing language model reasoning: A survey. *Acad J Sci Technol*. 2025;14(1):227-235.
doi: 10.54097/h21fky45
7. Chen J, Mashkova O, Zhapa-Camacho F, Hoehndorf R, He Y, Horrocks I. Ontology embedding: a survey of methods, applications and resources. *IEEE Trans Knowl Data Eng*. 2025;37(7):4193-4212.
doi: 10.1109/TKDE.2025.3559023

8. Huang T, Rizvi SA, Thakur RK, *et al.* HEART: learning better representation of EHR data with a heterogeneous relation-aware transformer. *J Biomed Inform.* 2024;159:104741.
doi: 10.1016/j.jbi.2024.104741
9. Lewis P, Perez E, Piktus A, *et al.* Retrieval-augmented generation for knowledge-intensive NLP tasks. In: NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc; 2020:9459-9474. Accessed August 13, 2026. <https://dl.acm.org/doi/10.5555/3495724.3496517>
10. Upadhyay R, Viviani M. Enhancing health information retrieval with RAG by prioritizing topical relevance and factual accuracy. *Discov Comput.* 2025;28(1):27.
doi: 10.1007/s10791-025-09505-5
11. Hogan A, Blomqvist E, Cochez M, *et al.* Knowledge graphs. *ACM Comput Surv.* 2021;54(4):1-37.
doi: 10.1145/3447772
12. Wang S, Yang H, Liu W. Research on the construction and application of retrieval enhanced generation (RAG) model based on knowledge graph. *Sci Rep.* 2025;15(1):40425.
doi: 10.1038/s41598-025-21222-z
13. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. *arXiv*. Preprint posted online 2023.
doi: 10.48550/arXiv.2303.13375
14. Ong CS, Obey NT, Zheng Y, Cohan A, Schneider EB. SurgeryLLM: a retrieval-augmented generation large language model framework for surgical decision support and workflow enhancement. *npj Digit Med.* 2024;7:364.
doi: 10.1038/s41746-024-01391-3
15. Amann J, Vetter D, Blomberg SN, *et al.* To explain or not to explain? Artificial intelligence explainability in clinical decision support systems. *PLoS Digit Health.* 2022;1(2):e0000016.
doi: 10.1371/journal.pdig.0000016
16. Hager P, Jungmann F, Holland R, *et al.* Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med.* 2024;30(9):2613-2622.
doi: 10.1038/s41591-024-03097-1
17. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digit Med.* 2020;3:17.
doi: 10.1038/s41746-020-0221-y
18. Lee J, Yoon W, Kim S, *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics.* 2020;36(4):1234-1240.
doi: 10.1093/bioinformatics/btz682
19. Boutouhami K, Zhang J, Qi G, Gao H. Uncertain ontology-aware knowledge graph embeddings. In: Wang X, Lisi FA, Xiao G, Botoeva E, eds. *Semantic Technology. JIST 2019. Communications in Computer and Information Science, vol 1157.* Singapore: Springer; 2020:129-136.
doi: 10.1007/978-981-15-3412-6_13
20. Johnson AEW, Bulgarelli L, Shen L, *et al.* MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data.* 2023;10(1):1.
doi: 10.1038/s41597-022-01899-x
21. Centers for Medicare & Medicaid Services. ICD-10-CM Official Guidelines for Coding and Reporting: FY 2023 (October 1, 2022–September 30, 2023). Published 2023. Accessed August 13, 2026. <https://www.cms.gov/files/document/fy-2023-icd-10-cm-coding-guidelines-updated-01/11/2023.pdf>
22. Johnson J, Douze M, Jégou H. Billion-scale similarity search with GPUs. *IEEE Trans Big Data.* 2021;7(3):535-547.
doi: 10.1109/TBDATA.2019.2921572
23. Singer M, Deutschman CS, Seymour CW, *et al.* The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA.* 2016;315(8):801-810.
doi: 10.1001/jama.2016.0287
24. Shang Y, Tian Y, Zhou M, *et al.* EHR-oriented knowledge graph system: toward efficient utilization of non-used information buried in routine clinical practice. *IEEE J Biomed Health Inform.* 2021;25(7):2463-2475.
doi: 10.1109/JBHI.2021.3085003
25. Izacard G, Lewis P, Lomeli M, *et al.* ATLAS: few-shot learning with retrieval augmented language models. *J Mach Learn Res.* 2023;24(251):1-43. Accessed August 13, 2026. <https://jmlr.org/papers/v24/23-0037.html>

Appendix

System prompt:

“You are a clinical decision support assistant. Given the following retrieved evidence from a patient knowledge graph: {Ch}, answer the clinical question and provide a confidence score in [0,1]...”

User prompt:

“Patient query: {query} Retrieved evidence: {Ch} Provide: (1) binary prediction, (2) confidence score, (3) evidence path”