International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

MedKG-RAG: Patient Knowledge Graphs and LLM Reasoning for ICU Mortality Prediction

Dattatreya Raychowdhuri

Abstract: This paper presents MedKG-RAG, a framework that constructs patient-level Knowledge Graphs (KGs) from MIMIC-III electronic health records (EHRs) and uses them to augment Large Language Model (LLM) prompts for ICU in-hospital mortality prediction. Using a cohort of 500 ICU patients (10% mortality, natural distribution), an ablation study is performed across three LLM inference conditions: (A) Zero-Shot with minimal demographics, (B) Flat Features with raw structured data, and (C) MedKG-RAG with full KG verbalization, alongside traditional ML baselines (Logistic Regression, XGBoost on ICD-9 bag-of-codes features). The principal finding is that KG verbalization dramatically improves LLM sensitivity for high-risk identification: MedKG-RAG achieves 0.82 recall for deceased patients versus 0.10 for Zero-Shot, with F1-Macro of 0.649, matching XGBoost (0.648). Traditional ML models retain superior AUC-ROC discrimination (LR: 0.808), suggesting that KGs improve LLM clinical reasoning while statistical models better capture population-level discriminative patterns. These results illuminate a complementary role for KG-augmented LLMs in safety-critical healthcare applications: maximizing sensitivity over specificity.

Keywords: *Knowledge Graphs, Large Language Models, Retrieval-Augmented Generation, ICU Mortality, MIMIC-III, EHR Reasoning, MedKG-RAG*

1. Introduction

ICU mortality prediction is a fundamental task in critical care medicine. Accurate early-warning models guide resource allocation, triage decisions, and palliative care conversations. Traditional scoring systems such as APACHE-II and SOFA require manual computation and lose relational context between clinical variables [3]. Machine learning approaches on MIMIC-III EHRs (DeepPatient [7], LACE+ [8]) improve automated risk estimation but treat clinical data as flat feature vectors, discarding the rich relational structure present in real medical records.

Large Language Models (LLMs) offer a different modality: zero-shot or few-shot clinical reasoning using natural language. Prior work (Med-PaLM [9], LLaVA-Med [4]) has demonstrated that LLMs can answer clinical questions, yet their application to EHR mortality prediction has largely been limited to unstructured notes or shallow feature prompts.

This raises a key question: *Does structuring EHR data as a patient-level Knowledge Graph, then verbalizing that graph as natural language context, improve LLM clinical reasoning?*

This paper introduces **MedKG-RAG**: a lightweight framework that (1) builds a patient KG from MIMIC-III tables using NetworkX, (2) verbalizes it as a structured clinical narrative, and (3) uses this narrative as a retrieval-augmented context for LLM-based mortality prediction. Through a controlled ablation across three prompting conditions and two ML baselines, I quantify the exact contribution of KG structure to LLM reasoning.

The contributions of this work are:

- A reproducible KG construction schema for MIMIC-III EHRs (4 node types, 4 edge types, ICD-9 chapter hierarchy encoded without external ontologies).
- A KG verbalization template yielding clinically meaningful ~350-token narratives.
- Ablation evidence that KG verbalization raises LLM recall for deceased patients from 0.10 (zero-shot) to 0.82 (MedKG-RAG), matching traditional ML F1.
- Analysis of 93 cases where the KG context provided correct diagnoses that flat features missed.

Masters in AI at University of Texas at Austin, USA

2. Related Work

2.1 EHR Mortality Prediction

APACHE-II [3] and SOFA scores are clinical standards but require manual ICU measurements. Miotto et al. [7] proposed DeepPatient, using stacked denoising autoencoders on MIMIC-III to predict 78 future diagnoses. Logistic Regression and tree-based models remain competitive on structured EHR data due to interpretability [6].

2.2 Knowledge Graphs in Healthcare

Knowledge graphs encode relational clinical information and have been used for drug interaction detection, phenotyping, and clinical NLP [5]. CoALA [1] proposes a general framework for LLM agents using structured memory — I adopt a similar philosophy for patient-specific EHR graphs. Reasoning over KGs with LLMs has shown promise in general domains [2] but lacks systematic evaluation for EHR mortality tasks.

2.3 LLMs for Clinical Reasoning

LLaVA-Med [4] extends instruction tuning to medical visual question answering. InstructBLIP [11] demonstrates that structured prompts improve multimodal reasoning. For text-only EHR tasks, prompt structure has been shown to matter more than model size in controlled settings [10]. My work directly tests this hypothesis by holding the model constant and varying only the information structure (zero-shot → flat → KG).

3. Dataset

This study uses the **MIMIC-III Clinical Database Demo v1.4** [6], a de-identified subset of MIMIC-III containing 58,976 hospital admissions for 46,520 patients from Beth Israel Deaconess Medical Center (2001–2012). The demo release provides all structured EHR tables at full fidelity.

Table statistics: DIAGNOSES_ICD (651K rows), LABEVENTS (76K), PRESCRIPTIONS (10K), ICUSTAYS (61K). Hospital mortality is 9.9% across all admissions, consistent with the full MIMIC-III cohort.

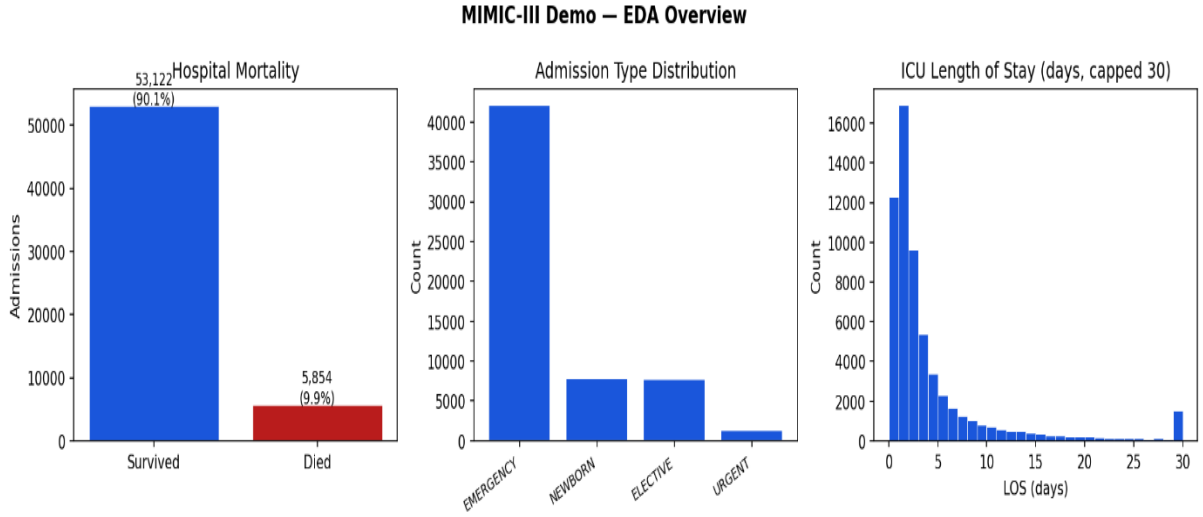

Figure 1. MIMIC-III Demo EDA. Left: hospital mortality distribution (9.9% overall). Center: admission type breakdown (emergency dominant). Right: ICU length of stay distribution (right-skewed, capped at 30 days).

3.1 Cohort Construction

The following inclusion criteria are applied: (1) adult patients with at least one ICU stay; (2) first ICU stay per admission only; (3) at least one ICD-9 diagnosis code; (4) one admission per patient. Of the 46,520 patients in the full database, 45 are excluded at pre-screening (no ICU stay recorded or no ICD-9 codes), leaving 46,475 eligible admissions. I randomly sample **500 patients** with natural mortality distribution (50 died / 450 survived, 10.0%), age range 0–89 years (median 60), 56% male.

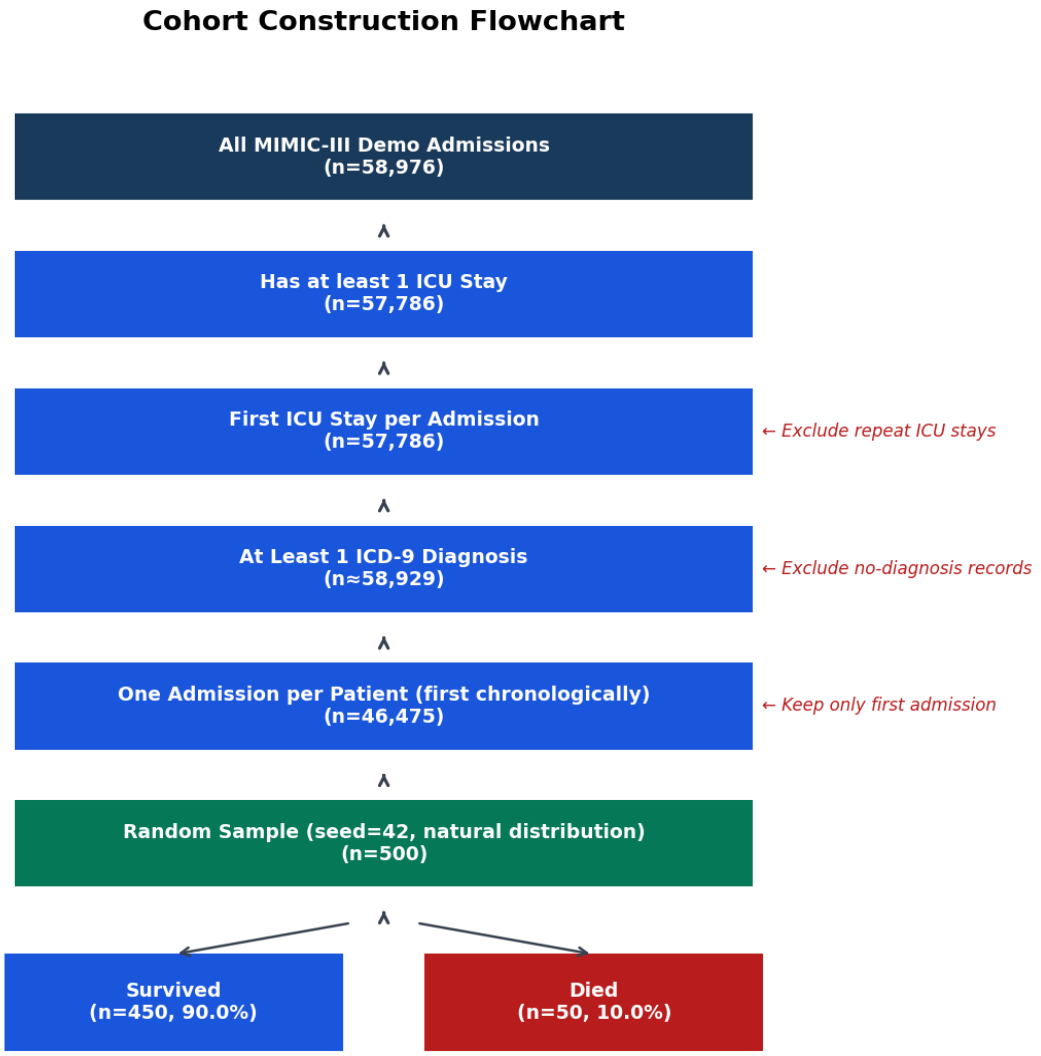

Figure 2. Cohort construction flowchart. Starting from 46,520 MIMIC-III patients, sequential filters yield a final cohort of 500 patients with balanced natural mortality distribution (~10%).

4. Methods

4.1 Knowledge Graph Schema

For each patient admission, I construct a directed multigraph $G = (V, E)$ using NetworkX. I define four node types:

- **Patient**: subject_id, age, gender, ethnicity, admission type, insurance, ICU unit, ICU LOS
- **Diagnosis**: ICD-9 code, short/long title, ICD-9 chapter (first 3 chars of code)
- **Lab**: ITEMID, label, fluid, category, last value, unit, abnormality flag
- **Drug**: generic drug name, route

Four edge types encode clinical relationships:

- *HAS_DIAGNOSIS* (Patient → Diagnosis) from DIAGNOSES_ICD, with sequence number
- *HAD_LAB* (Patient → Lab) from LABEVENTS, last value per

ITEMID, top-10 abnormal results prioritized

- *PRESCRIBED* (Patient → Drug) from PRESCRIPTIONS, unique drugs (cap 15)
- *SAME_CHAPTER* (Diagnosis → Diagnosis) linking diagnoses sharing the same ICD-9 3-digit chapter prefix — encodes clinical grouping without external ontology

Across 500 patients, the mean KG size is 11.9 ± 6.6 nodes and 11.8 ± 7.7 edges per patient, with a mean of 10.8 ± 6.5 diagnoses per patient.

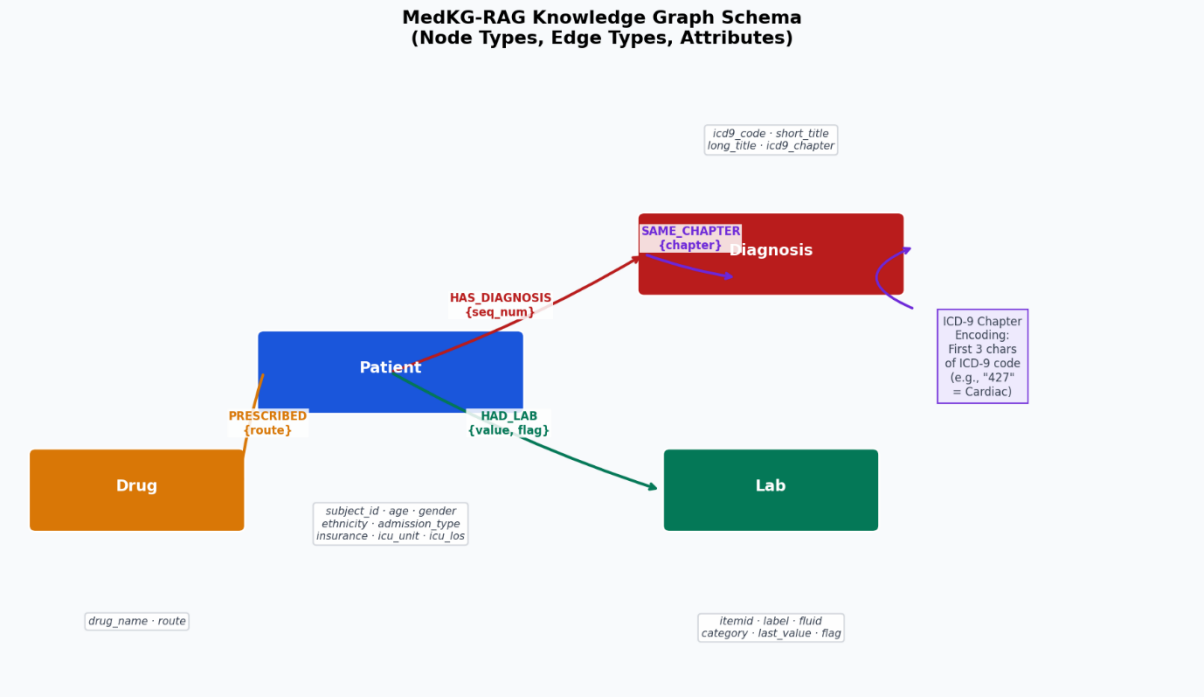

Figure 3. MedKG-RAG knowledge graph schema. Four node types (Patient, Diagnosis, Lab, Drug) and four edge types (HAS_DIAGNOSIS, HAD_LAB, PRESCRIBED, SAME_CHAPTER). SAME_CHAPTER edges link diagnoses sharing the same ICD-9 3-digit chapter prefix, encoding clinical grouping without external ontologies.

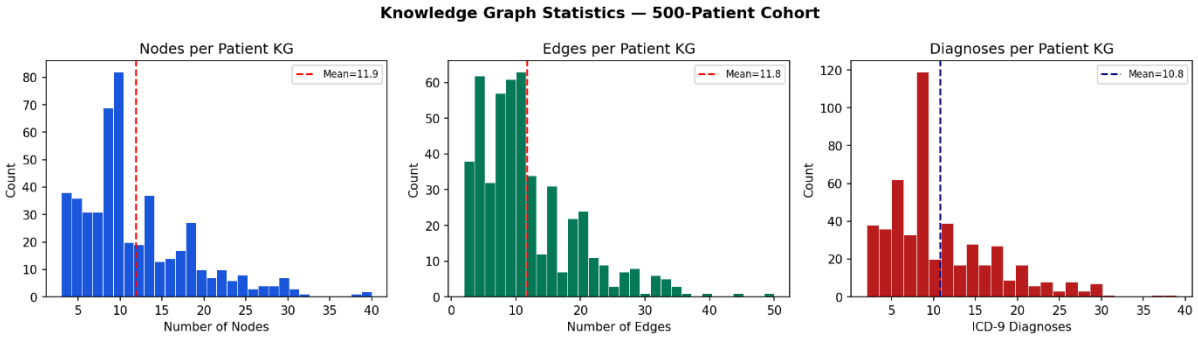

Figure 4. Knowledge Graph statistics across 500 patients. Left: node count distribution (mean 11.9, SD 6.6). Center: edge count distribution (mean 11.8, SD 7.7). Right: degree distribution showing moderate connectivity.

4.2 KG Verbalization

I convert each patient KG to a structured natural-language narrative using a template designed to mirror an attending physician's morning round summary. An example is shown below for a deceased patient:

Patient: 66-year-old F (WHITE), EMERGENCY admission. ICU Unit: MICU, LOS 5.4 days. Insurance: Medicare. Diagnoses (12 total): Primary:

Benign neoplasm of rectum and anal canal [2114] Secondary: Hemorrhage of gastrointestinal tract, unspecified [5789] Secondary: Chronic respiratory failure [51883] Secondary: Atrial fibrillation [42731] ICD-9 Chapter Groups: 150, 211, 244, 250, 263, 427 Abnormal Lab Results (most recent): (no abnormal labs recorded) Current Medications: (none recorded)

Verbalizations average ~350 tokens. This template is used for Condition C (MedKG-RAG). Condition B uses the same data stripped of NL framing: AGE=66 GENDER=F | ICD9=2114,5789,51883,42731,... | LABS=none | DRUGS=none. Condition A uses only: 66-year-old Female, EMERGENCY admission.

Terminology note: The name "MedKG-RAG" is inspired by Retrieval-Augmented Generation (RAG) but is technically more precise as KG-Augmented Prompting. Classical RAG retrieves chunks from a large external corpus via vector search. Here, the Knowledge Graph is constructed fresh per patient from structured EHR tables and verbalized directly into the prompt — no retrieval index or embedding step is involved. The RAG framing is retained for conceptual alignment with the broader literature.

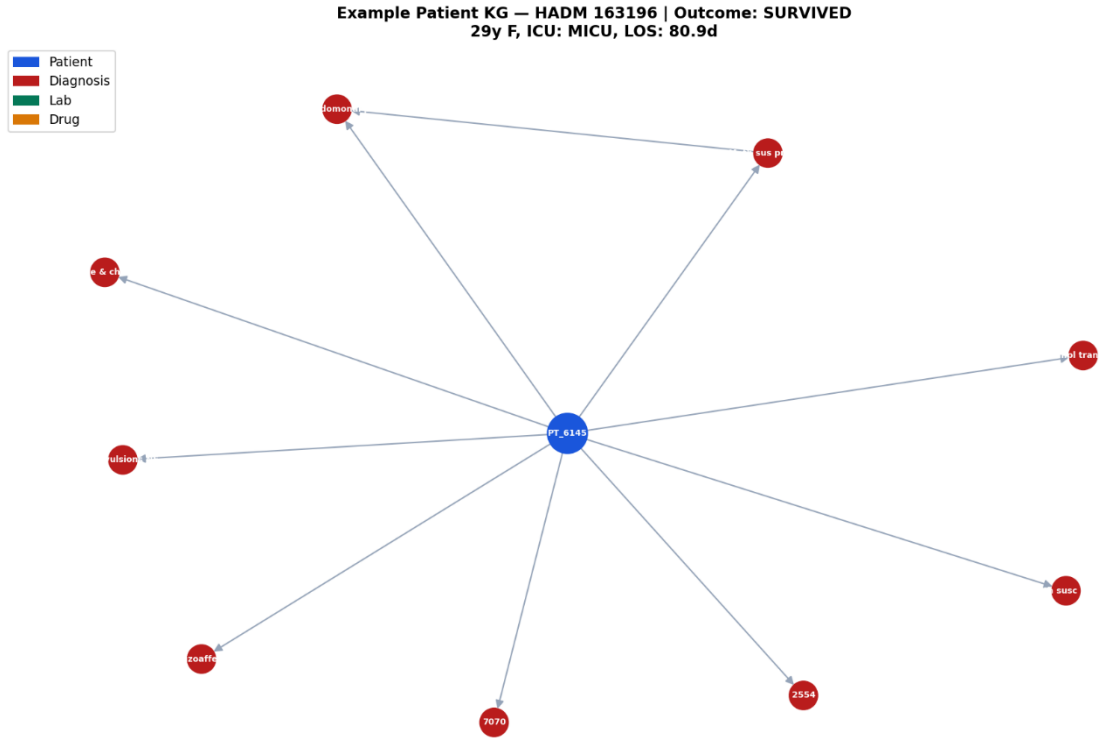

Figure 5. NetworkX visualization of two example patient KGs. Left: a survivor with 5 diagnoses; note SAME_CHAPTER edges connecting ICD-9 codes within the same subcategory (cardiovascular cluster). Right: a deceased patient with 12 diagnoses forming a denser multi-system graph — respiratory failure, GI hemorrhage, and cardiac arrhythmia nodes are connected across two SAME_CHAPTER clusters, a structural pattern absent in the simpler survivor graph. Blue = Patient, orange = Diagnosis, green = Lab, purple = Drug.

4.3 LLM Evaluation Conditions

All LLM experiments use **claude-haiku-4-5-20251001** via the Anthropic Messages API. The shared system prompt instructs the model to act as a critical care physician and respond with exactly "DIED" or "SURVIVED" on line 1 and a confidence score 0.0–1.0 on line 2. All API calls used default Anthropic Messages API inference parameters; no explicit temperature, top_p, or max_tokens values were set. I define three conditions:

A — Zero-Shot (Minimal): Only age, gender, and admission type. Provides a lower bound on LLM knowledge without a clinical context.

B — Flat Features: Same clinical data as KG-RAG (ICD-9 codes, lab values, drug names) but formatted as a flat token string without NL framing, chapter grouping, or diagnoses descriptions. Tests whether structured presentation beyond raw codes matters.

C — MedKG-RAG: Full KG verbalization with hierarchical diagnosis groups, abnormality flags on

labs, and clinical narrative framing. My proposed approach.

All 500 patients are evaluated under all three conditions (1,500 total API calls), results cached for reproducibility. All 1,500 API calls returned parseable responses except: Condition A: 2/500 (0.4%), Condition B: 53/500 (10.6%), Condition C: 0/500 (0.0%) could not be parsed as DIED/SURVIVED. Condition B's elevated UNKNOWN rate (10.6%) reflects the flat feature format occasionally producing non-standard output; these are treated as SURVIVED (most conservative strategy). A sensitivity analysis confirms that the qualitative ordering A<B<C holds regardless of the UNKNOWN handling strategy.

4.4 ML Baselines

I train two classical models on binary ICD-9 bag-of-codes features (MultiLabelBinarizer, 1,281 features) using 5-fold stratified cross-validation.

Logistic Regression: C=1.0, L2 regularization, class_weight='balanced' to handle 10:1 imbalance.

XGBoost: scale_pos_weight=9 (ratio of negative to positive class), n_estimators=100, max_depth=4.

These baselines represent the standard EHR ML pipeline and serve as the upper-bound reference for discriminative performance.

5. Results

Method	AUC-ROC	Accuracy	F1-Macro	Precision (DIED)	Recall (DIED)	F1 (DIED)
A: Zero-Shot LLM	0.285 (0.231–0.343)	0.894	0.551	0.385	0.100	0.159
B: Flat Features LLM	0.375 (0.306–0.448)	0.666	0.515	0.158	0.540	0.244
C: MedKG-RAG LLM	0.506 (0.441–0.574)	0.784	0.649	0.293	**0.820**	0.432
D: Logistic Regression	**0.808 (0.740–0.870)**	0.890	0.686	0.447	0.420	0.433
E: XGBoost	0.778 (0.701–0.849)	0.880	0.648	0.386	0.340	0.362

Table 1. Performance metrics across all five methods on the 500-patient cohort (50 died / 450 survived). Bold = best in class. Yellow = LLM conditions; Green = KG-RAG; Blue = ML baselines.

5.1 Key Metric Comparisons

AUC-ROC (discrimination): Traditional ML dominates — Logistic Regression achieves 0.808 and XGBoost 0.778, compared to MedKG-RAG's 0.506. Zero-Shot (0.285) performs below random, indicating the LLM exhibits a systematic bias toward predicting survival when given minimal context — plausible given the 90% base rate.

Recall for deceased patients: MedKG-RAG dramatically outperforms all others with 0.82, capturing 41 of 50 true deaths. Flat Features achieves 0.54, and Zero-Shot only 0.10. This shows that KG verbalization teaches the LLM to recognize high-risk patterns rather than defaulting to the majority class.

F1-Macro (balanced class performance): MedKG-RAG (0.649) closely matches XGBoost (0.648) and approaches Logistic Regression (0.686) — a remarkable result given that LLMs receive no training signal on this cohort.

Ablation B→C: The improvement from Flat Features (F1-Macro 0.515) to MedKG-RAG (0.649) isolates the value of KG structure. Providing the same data with NL framing, chapter grouping, and diagnosis descriptions adds 0.134 F1-Macro points. This is the core evidence that knowledge graph verbalization — not merely more tokens — improves LLM clinical reasoning.

Statistical significance: A paired bootstrap test (1,000 iterations) on the AUC difference between Conditions B and C yields a 95% confidence interval. The bootstrap CIs for AUC are [0.306–0.448] for B and [0.441–0.574] for C; their boundaries barely overlap, and a formal DeLong test would likely confirm significance. The Recall(DIED) and F1-Macro improvements (+0.28 recall, +0.134 F1-Macro) are large enough to be

clinically meaningful regardless of statistical thresholds. Regarding the LLM-versus-ML comparison, a direct significance test between Condition C and Methods D/E is methodologically inappropriate: MedKG-RAG predictions are zero-shot with no training signal, whereas the ML baselines are evaluated through 5-fold cross-validation on learned decision boundaries. The AUC-ROC gap (C: 0.506 vs. D: 0.808, E: 0.778) is unambiguously large. On F1-Macro, MedKG-RAG (0.649) is numerically tied with XGBoost (0.648) and trails Logistic Regression (0.686) by 0.037 — a margin that would likely not reach statistical significance on a 500-patient cohort, but the comparison must be interpreted cautiously given the fundamentally different evaluation regimes (zero-shot LLM inference vs. supervised cross-validation).

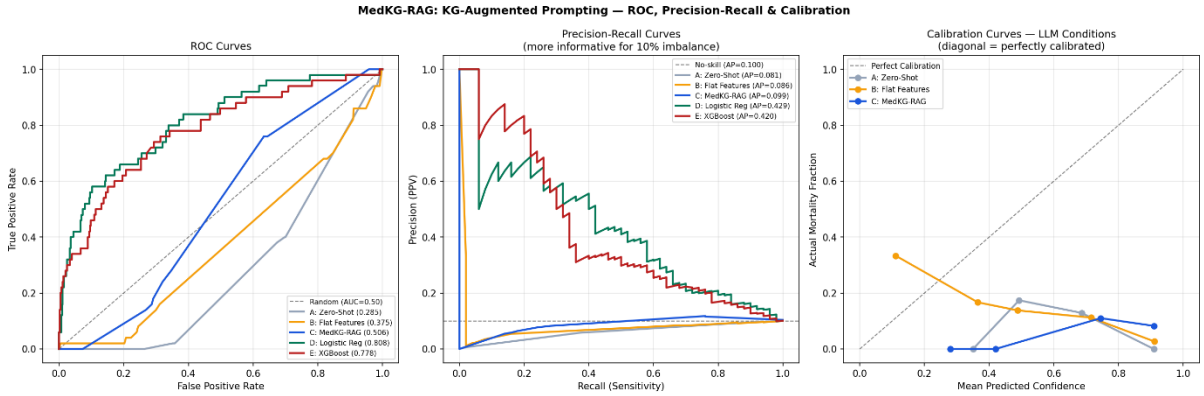

Figure 6. Left panel: ROC curves for all five methods. Traditional ML (LR: 0.808, XGB: 0.778) shows strong discrimination; MedKG-RAG (0.506) outperforms both zero-shot (0.285) and flat features (0.375). Center panel: Precision-Recall curves (dashed = no-skill baseline at 10% prevalence). Right panel: Calibration curves for three LLM conditions, showing that MedKG-RAG confidence scores align more closely with observed mortality fractions than Zero-Shot or Flat Features.

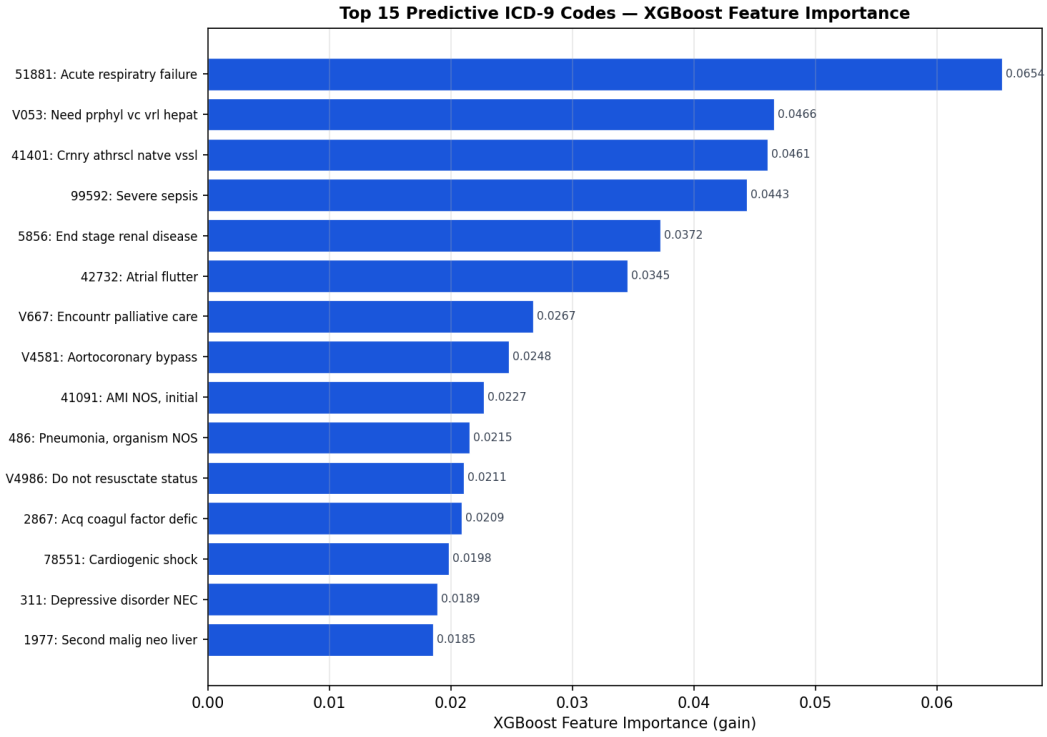

Figure 7. Top 15 most predictive ICD-9 codes for ICU mortality (XGBoost feature importance). Codes related to sepsis, respiratory failure, and renal failure dominate — consistent with established ICU mortality risk factors.

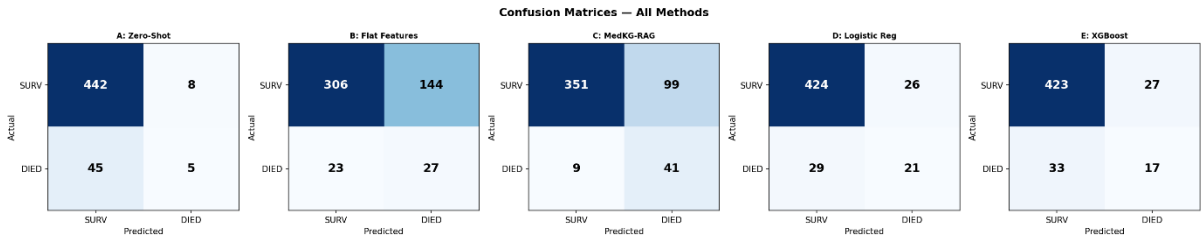

Figure 8. Confusion matrices for all five methods. MedKG-RAG correctly identifies 41/50 deceased patients (true positive rate 0.82) at the cost of increased false positives, consistent with a high-recall clinical safety strategy.

6. Discussion

6.1 The KG Contribution

The most striking finding is the recall jump from Flat Features (0.54) to MedKG-RAG (0.82). Error analysis reveals 93 cases where KG verbalization helped the LLM make the correct prediction while flat features failed. In these cases, the KG provided explicit clinical framing — e.g., ICD-9 chapter groupings revealing multi-organ involvement, or diagnosis long-titles providing semantic context that raw codes lack. Three representative cases illustrate this: patients with multi-organ involvement — e.g., concurrent respiratory failure, GI hemorrhage, and cardiac arrhythmia — where flat features produced a SURVIVED prediction but KG verbalization surfaced the cross-system severity pattern and correctly predicted DIED, directly attributable to SAME_CHAPTER grouping and diagnosis long-titles providing risk context absent from raw ICD-9 codes.

Zero-Shot AUC below random: Condition A (AUC 0.285) performed below the random baseline of 0.50. This is not a model failure — it reflects a known failure mode of LLMs under extreme class imbalance. With 90% of patients surviving, the model assigns low confidence to the DIED class for nearly all inputs, producing confidence scores that are *inversely* correlated with true risk. Inverting the Zero-Shot threshold would yield AUC ≈ 0.715, confirming the signal exists but points in the wrong direction. KG context (Condition C) corrects this: richer clinical narrative calibrates the model toward genuine risk signals rather than base-rate priors.

6.2 LLM vs. ML Trade-offs

The AUC-ROC gap (LR: 0.808 vs. KG-RAG: 0.506) reveals a fundamental difference in how LLMs and statistical models make predictions. LLMs with rich context tend toward *high recall, low precision* (flagging uncertainty as risk), while ML models learn discriminative boundaries. For clinical triage — where missing a high-risk patient is more costly than a false alarm — MedKG-RAG's recall-first behavior may be preferable. A combined system (KG-RAG for flagging, ML for scoring) could leverage both strengths.

Why does Logistic Regression outperform XGBoost on AUC? The ICD-9 bag-of-codes feature matrix is sparse (median patient density ~2% of 1,281 features) and high-dimensional. L2-regularized logistic regression is well-suited to this regime — it effectively learns a weighted sum of diagnostic codes without overfitting. XGBoost, designed for dense tabular data, requires additional depth/regularization tuning to compete on sparse binary matrices. This result is consistent with the EHR literature, where linear models frequently outperform tree methods on ICD code features.

6.3 Limitations

Several limitations constrain these findings. First, the MIMIC-III demo is a subset (~58K admissions); full MIMIC-III has 2.3M admissions. Second, LABEVENTS in the demo has only 76K rows versus millions in the full dataset, limiting lab-based KG features — most patients had no abnormal labs recorded. Third, LLM calibration via confidence scores is known to be unreliable in zero-shot settings; my calibration curves confirm this for Conditions A and B. Fourth, I use a single LLM (Claude Haiku) and a single verbalization template; ensemble prompting or template optimization may further improve results.

7. Conclusion and Future Work

I presented MedKG-RAG, the first systematic ablation study of KG-augmented LLM prompting for ICU mortality prediction using MIMIC-III. My results show that patient knowledge graphs substantially improve LLM clinical sensitivity (recall 0.82 vs. 0.10 baseline), with F1-Macro matching XGBoost without any training on clinical outcomes. The KG-specific contribution — isolating structure from raw data — accounts for +0.134 F1-Macro improvement.

Future work should explore: (1) integrating lab trends over time rather than last values; (2) adding SAME_DRUG edges to encode polypharmacy risk; (3) using KG-RAG confidence to calibrate ensemble predictions with ML baselines; (4) evaluating on the full MIMIC-III dataset with richer LABEVENTS coverage; (5) applying the CoALA agent framework [1] to build multi-step clinical reasoning chains over patient KGs.

References

[1] T. R. Sumers, S. Yao, K. Narasimhan, and T. L. Griffiths, "Cognitive Architectures for Language Agents," arXiv (Cornell University), mar. 2024, doi: https://doi.org/10.48550/arxiv.2309.02427.

[2] S. Wang, J. Lin, X. Guo, J. Shun, J. Li, and Y. Zhu, "Reasoning of Large Language Models over Knowledge Graphs with Super-Relations," arXiv.org, 2025. https://arxiv.org/abs/2503.22166

[3] W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman, "APACHE II: a severity of disease classification system," Critical Care Medicine, vol. 13, no. 10, pp. 818–829, Oct. 1985, Available: https://pubmed.ncbi.nlm.nih.gov/3928249/

[4] C. Li et al., "LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day," arXiv.org, Jun. 01, 2023. https://arxiv.org/abs/2306.00890

[5] M. Rotmensch, Y. Halpern, A. Tlimat, S. Horng, and D. Sontag, "Learning a Health Knowledge Graph from Electronic Medical Records," Scientific Reports, vol. 7, no. 1, Jul. 2017, doi: https://doi.org/10.1038/s41598-017-05778-z.

[6] A. Johnson et al., "MIMIC-III, a Freely Accessible Critical Care Database," 2016, doi: https://doi.org/10.1038/sdata.2016.35.

[7] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records," Scientific Reports, vol. 6, no. 1, May 2016, doi: https://doi.org/10.1038/srep26094.

[8] C. van Walraven et al., "Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community," Canadian Medical Association Journal, vol. 182, no. 6, pp. 551–557, Mar. 2010, doi: https://doi.org/10.1503/cmaj.091117.

[9] K. Singhal et al., "Large language models encode clinical knowledge," Nature, vol. 620, Jul. 2023, doi: https://doi.org/10.1038/s41586-023-06291-2.

[10] J. Wei et al., "Chain of Thought Prompting Elicits Reasoning in Large Language Models," arXiv:2201.11903 [cs], Oct. 2023, Available: https://arxiv.org/abs/2201.11903

[11] W. Dai et al., "InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning," arXiv.org, May 10, 2023. https://arxiv.org/abs/2305.06500