Med-KG: A General-Purpose Knowledge Graph for Understanding Clinical Interactions

Purvi Sehgal*, Sahithi Ankireddy*

Department of Computing + Mathematical Sciences
California Institute of Technology, Pasadena, USA
{psehgal, sankired}@caltech.edu

*These authors contributed equally to this work.

***Abstract*—This paper introduces Med-KG, a general-purpose knowledge graph that models clinical interactions such as patient similarity using large language models (LLMs). Unlike prior knowledge graphs that captured only one aspect of the patient experience, Med-KG models the entire patient journey with all the associated interactions and intermediate steps. Its general purpose nature addresses a key inefficiency in current studies - the need to repeatedly create new, specialized knowledge graphs for each task. The Med-KG pipeline consists of three main stages. First, in the feature extraction stage, relevant clinical information from the MIMIC-IV database — diagnoses, symptoms, medications, procedures, and demographics — was processed and concatenated into standardized representations. Next, LLMs were queried to convert these structured inputs into a clinical knowledge graph. Finally, Med-KG was applied to a downstream task of patient similarity assessment. Controlled and direct evaluation frameworks compared predicted similarity against diagnosis-based ground truth. Additionally, a benchmark dataset was created for the patient similarity task. This framework not only demonstrates the effectiveness of using Med-KG for patient similarity, but also provides a foundation that can be extended to other downstream clinical applications.**

***Index Terms*—Healthcare AI, Knowledge Graphs, Large Language Models, MIMIC-IV, Clinical Interactions, Patient Similarity, Multimodal Data Integration**

I. Introduction

The medical field generates a wide range of data distributed across various modalities and sources. Clinical information may take the form of unstructured physician notes, structured electronic health records (EHRs), medical imaging (e.g., X-rays, CT scans), laboratory test results, and genomic data. These modalities span a different format —text, images, time-series, or structured tabular records. They each capture distinct aspects of a patient's health status, and together offer a more holistic view of the clinical journey.

However, effectively integrating this multimodal data remains a key challenge in healthcare. The data ranging from free-text to high-dimensional omics data makes it difficult to create unified representations. Bridging these sources in a way that preserves semantic meaning and clinical utility is a key step toward enabling more effective care. This challenge has led to an active area of research focused on multimodal machine learning, representation learning, and the development of frameworks that can reason across diverse data types in a meaningful way [1] [2] [3] [4].

Knowledge graphs (KGs) have emerged as a promising approach for organizing this fragmented information. By representing clinical concepts—such as diagnoses, symptoms, medications, procedures, and providers—as nodes, and their interactions as edges, KGs unify diverse data sources into a structured representation. Biomedical KGs have been successfully applied to various downstream tasks, including drug repurposing, clinical decision support, and adverse event prediction [5] [6] [7]. However, most existing KGs are either domain-specific or built around a narrow set of entities, limiting their generality. Moreover, the construction of new knowledge graphs for different use cases reduces efficiency, is redundant, and can be simplified with a more general knowledge graph. However, building a comprehensive, general-purpose KG that captures the full scope of a patient's clinical journey remains an open challenge.

Med-KG is a large-scale, general-purpose knowledge graph that captures information across the entire clinical journey — not just patient- or diagnosis-centric data. It includes details about providers, pharmacists, procedures, medications, and more, enabling a comprehensive view of healthcare interactions. Unlike existing graphs that are narrowly focused, Med-KG is designed to be broadly applicable.

To extract insights from Med-KG, LLMs were leveraged. Traditional approaches for analyzing knowledge graphs rely on graph algorithms and various neural network encoders, which learn latent representations from node features and the underlying graph structure and oftentimes have difficulty generalizing [8][9]. More recently, LLMs have shown strong performance in processing natural language representations of structured data. By converting relational triples into text, LLMs were used to reason over Med-KG which helped support a wide range of downstream applications, including patient similarity analysis.

The key contributions of this project are as follows,

- The construction of a large, general-purpose clinical knowledge graph that models multiple healthcare entities.

- The application of LLMs to reason over this graph for more use cases.
- A focus on patient similarity as one representative application, showing how Med-KG supports various clinical analyzes.

II. Literature Review

Knowledge graphs have been used in the biomedical field for years. These graphs have traditionally been very specialized and only designed for a small range of specific downstream applications. In a review paper describing all the cutting-edge knowledge graphs in the biomedical space, most had under 20 node types (implying a small number of features) and all were built for their specific downstream use cases [10].

In research conducted by Halamka and Cerrato (2023), knowledge graphs were utilized to extract diagnosis and procedure codes for more accurate coding, but they only contained enough information for that task. Another knowledge graph was built only from report summaries to provide information on drug interactions [11]. Collectively, such graphs are currently being used to solve a wide range of tasks within biomedicine, underscoring their importance across various domains within the field. However, because they are highly specialized for their own tasks, each graph cannot be applied to many other problems besides their own.

Next, the literature review turns to patient-centric knowledge graphs built from Electronic Health Records (EHRs), as this is the emphasis of this research. In general, these knowledge graphs were also found to be specific to the downstream application. According to Zheng et al. [12], PharmKG only identified relationships between genes, drugs, and diseases, and according to Zhang and Che [13], the DRKF knowledge graph only linked drugs and Parkinson's disease. Another research paper used EHRs to create a patient similarity graph using diagnoses, procedures, and medications information [7]. This graph only had a limited number of features (diagnoses, procedures, and medications) and was tailored only for the downstream patient similarity task but none others. The general purpose knowledge graph developed in this research was also applied to this patient similarity task to show its effectiveness in solving problems previously solved by specialized graphs.

Additionally, because there were a large number of features in this Med-KG graph, LLMs were queried to infer relations between sets of features. This approach is backed by recent advances in research, in which LLMs accomplished state-of-the-art performance in tasks such as relation prediction [14]. While this research was conducted for specialized knowledge graphs, the findings were applied to inform methodology for this general-purpose knowledge graph.

In summary, knowledge graphs utilized for biomedical and patient-centric healthcare tasks were previously specialized and independent of each other, which added a layer of redundancy. To resolve these issues, a general purpose knowledge graph was designed to follow the patient's journey including symptoms, procedures, provider and pharmacy orders, diagnoses, and subsequent admissions, etc. This comprehensive view of the patient journey from start to end provides a unified framework which can help researchers solve a multitude of downstream tasks and understand clinical interactions more holistically.

III. Med-KG Generation

A. Overview

The methodology consists of three steps. First, in the feature extraction step, relevant data from the MIMIC-IV database was processed and aggregated. Second, this data was then used to generate relational information and a knowledge graph. Third, the resulting graph was used for a patient similarity downstream application to demonstrate its effectiveness in solving a real-world problem. The details are explained in the following sections.

B. Dataset

The MIMIC-IV database was used for this project, which includes tabular, textual, and image data for more than 300,000 patients and all their hospital admissions. This includes information such as diagnoses, procedures, etc. The MIMIC-IV-Note database contains clinical notes from hospital admissions, and GPT-4o was queried to retrieve symptom information from the notes. All this information was spread over 26 files and was decentralized. The relevant information was then concatenated, extracted, and organized into two tables as explained below in the Feature Extraction section.

C. Feature Extraction

For each patient, relevant information was aggregated. The first table included patient information such as demographics and other commonalities across hospital admissions. The second table included patient admissions information where each row in the table contained the patient's admission information such as symptoms, prescribed medications, and provider name for that admission. In addition to extracting information relevant to this Med-KG graph from the 26 files mentioned above, some other necessary associations were created between features. For example, two provider orders were connected when one was discontinued by another and linked medications listed in a different file to provider orders. All codes were cross-referenced with a lookup table to retrieve their descriptions, which were then added as features to the table. For image data, detailed captions present in the MIMIC-IV database were incorporated into the knowledge graph.

All these features were organized in a dictionary by patient and admission numbers.

D. Knowledge Graph Creation

Once these features were extracted, a knowledge graph was generated. For each patient and admission, GPT-4o was queried with the feature dictionary and feature descriptions. This produced a list of relational triples which linked one entity to another entity. An example of this is the patient being

linked to the admission number with the label, "has admission" or the admission number being linked to a medicine with the label, "is prescribed."

These relational triples were concatenated and then iterated through to create a digraph using the Graphviz library. This graph captures the complex relationships between patients, providers, and other healthcare professionals, tracking each patient's full clinical journey—including orders placed, medications prescribed or discontinued, and other intermediate steps. Unlike other knowledge graphs that focus on a single aspect, such as diagnoses, this approach models the entire clinical care journey.

While this knowledge graph includes information for all patients and all admissions, Figure 1 below displays only a segment of the knowledge graph for one patient and one admission for clarity.

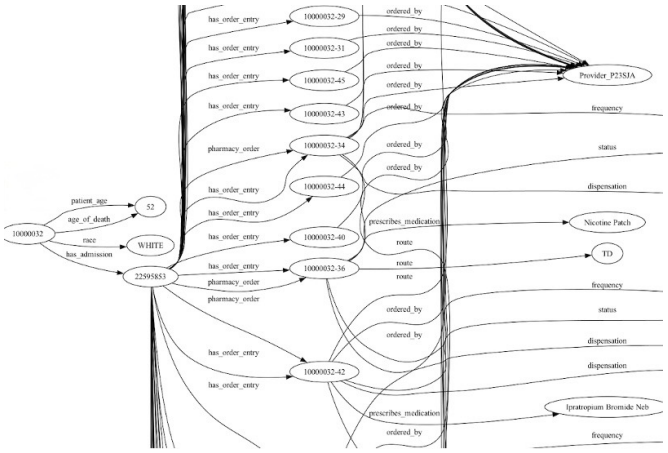

Fig. 1. Segment of the knowledge graph for a single patient and admission.

In Figure 1, the left-most node with the number 10000032 refers to the patient number and all of the information shown corresponds to that of the patient. The first level child nodes of the patient contains basic information and the admission number (22595853). All child nodes of that admission number have information from that patient's specific admission, including features like pharmacy orders, provider orders, and the associated providers. The features depicted in this figure were just a small subset of the features included in the final knowledge graph.

IV. LLM Application: Patient Similarity

Once constructed, the knowledge graph can be applied to a wide range of downstream tasks. Its large scale captures information on patients, providers, medications, procedures, and symptoms, making it a powerful tool for clinical decision support.

A. Patient Similarity

A primary application and focus of this research was patient similarity analysis, where the goal was to identify patients with similar medical profiles. This is valuable for providers aiming to better understand complex cases, as well as for junior medical professionals who can more easily learn by referencing similar past patient cases. By identifying patients most similar to the current patient, clinicians can gain insights into likely disease trajectories, effective treatment responses, and potential risks. This, in turn, enables more personalized care and improves clinical decision-making and outcomes.

To implement this, knowledge graph triples for all patients were retrieved and saved to a pickle file. These relation triples were then reformatted into natural language strings in a two step process. First, a triple like (99384712, is_diagnosed, asthma) was converted to the natural language sentence: "99384712 is diagnosed with asthma." These strings were then further organized by category (e.g., diagnoses, prescriptions, procedures) to introduce more structure and facilitate analysis.

This transformation was applied to all triples for each patient to create a consolidated textual representation per patient. This master string was then sent to an LLM, where it extracted similarity scores for each feature (e.g., diagnoses, prescribed medications, etc.) between every pairwise combination of patients. Each similarity score was calculated on a scale from 0 to 1 based on the number of overlapping items. A final overall similarity score was computed by averaging the sub-scores across all features. The resulting information was structured into a JSON format for each pairwise set of patients, enabling the final similarity scores to be easily sorted in order to identify the top k most similar patients for any given individual. Refer to Table 1 below for a sample similarity report between two patients.

TABLE I
SAMPLE SIMILARITY REPORT BETWEEN TWO PATIENTS, SHOWING COMPONENT-WISE SIMILARITIES AND SHARED CLINICAL FEATURES CONTRIBUTING TO THE OVERALL SIMILARITY SCORE.

Patient 1	99384712
Patient 2	88246109
Diagnosis Similarity	0.00
Prescription Similarity	0.71
Procedure Similarity	0.00
Symptoms Similarity	0.50
Overall Similarity	0.3025
Key Contributors	Shared prescriptions: Metoprolol, Sertraline, Atorvastatin, Albuterol Inhaler, Furosemide Shared symptoms: Fatigue, Shortness of Breath, Swelling in Legs, Difficulty Concentrating

B. Experimental Set-up

To evaluate the quality of the patient similarity method, two primary strategies were employed, along with a long-term evaluation method of creating a benchmark.

The first approach involved injecting known similar patients into the dataset and creating controlled evaluation cases. The second approach was more direct and evaluated performance on the data itself. As a long-term effort, a benchmark dataset is being developed to directly compare model-predicted

similarity scores—computed using all available clinical attributes—against expert-reviewed ground truth similarity annotations.

1) Controlled Evaluation: For the first evaluation method, for every real patient in the knowledge graph, three patient profiles were added. The first patient profile's diagnosis was the same as that of the real patient ($\sim$100% similarity). The second patient's diagnosis was somewhat similar ($\sim$50% similarity), and the third patient's diagnosis was completely unrelated to that of the real patient ($\sim$0% similarity), allowing for patients with varying similarity levels to be injected into the data. These known diagnoses similarity percentages were used as ground truth similarity scores of $\sim$1.0, $\sim$0.50, and $\sim$0.0.

The authors of this paper had prior experience with the ICD-10-CM coding system, having been trained by professional medical coders. Thus, these ground truths and patient profiles were carefully constructed using the ICD-10-CM coding system, which assigns diagnosis codes to various medical conditions. The hierarchical structure of this system was utilized to relate similar medical conditions to each other. Thus, this tool enabled the creation of patient profiles close to 100%, 50%, and 0% similarity. These ground truth similarity scores were cross-validated with LLM results. Patient profiles were then appended to the KG data.

Once the ground truths were determined, an LLM was used to predict three similarity scores for each real patient - one for every corresponding patient profile. Multiple LLM queries demonstrated consistent scoring. This process was repeated for all real patients.

The predicted and ground truth similarity scores were evaluated by calculating the Spearman correlation coefficient and Pearson correlation coefficients. Spearman correlation assessed whether the ranking order of predicted similarities matched the ranking order of ground truth similarities. Pearson correlation determined the degree of the linear relationship between predicted and ground truth values. Accounting for different predicted and ground truth similarity scales in this evaluation framework, these correlation coefficients together provided insight into whether the pipeline could accurately identify the relative ranking and degree of patient similarity.

2) Direct Evaluation: Unlike the first approach, this strategy did not modify the dataset and instead evaluates model performance on the patient pairs directly. Similar to the first approach, a ground truth similarity score is computed based solely on the overlap in diagnoses between each pair of patients. Then, the model determines a predicted patient similarity score based on all other available clinical features such as symptoms, medications, and procedures, while explicitly excluding diagnosis-related information.

To implement this evaluation framework, a subset of patients was randomly selected. For each patient in the subset, all associated knowledge graph triples were retrieved and converted into natural language strings. These strings were then filtered to separate diagnosis-related information from the rest. Given there was no pre-existing ground truth information for this data, diagnosis information served as a proxy under the assumption that similar patients have similar diagnoses.

For each pair of patients, two similarity reports were then generated using an LLM: one using only the diagnosis-related strings (ground truth), and the other using all remaining information excluding diagnoses (predicted score). This included prescription, procedure, symptom, and other similarities, which were averaged into a final predicted similarity score.

The key metric in the evaluation framework was the overall similarity score, which provided a single numerical value representing patient similarity. This shared metric allowed for direct comparison between the ground truth and predicted scores, enabling evaluation of how closely the predicted similarity (based on non-diagnostic information) aligned with the diagnosis-based ground truth.

The difference between the diagnosis-based ground truth score and the predicted similarity score was calculated for each pair. This delta served as a proxy for evaluating how well the method captures clinically meaningful similarities without relying on diagnosis information explicitly.

This setup was used only during the evaluation phase to assess how well the model could infer patient similarity without access to diagnoses. In practice, during inference mode, the model utilizes all available information, including diagnoses, to compute a comprehensive similarity score between patients. This distinction ensures that the evaluation reflects the model's ability to learn from indirect clinical signals, while inference uses the full knowledge graph for more accurate downstream usage.

3) Benchmark Creation: As a longer-term objective, in parallel, a benchmark dataset was also constructed, containing 20 patient records and their pairwise similarity scores. Clinical data for each of the 20 patients, including diagnoses, symptoms, prescriptions, and other relevant information, were generated. The ground truth pairwise similarity scores were manually curated using LLM assistance and will undergo review by clinical professionals affiliated with Caltech. This benchmark provides an additional evaluation, offering expert-validated ground truth against which predicted similarity scores can be directly compared without having to use diagnosis as a proxy for ground truth.

All together, these two evaluation strategies along with a benchmark will enable a robust assessment of the model's ability to infer patient similarity.

C. Performance Metrics

1) Controlled Evaluation: Spearman correlation and Pearson correlation was determined between predicted and ground truth similarity scores.

A Spearman correlation coefficient of 0.7728 reflects a strong relationship between the rank of predicted similarity scores and the rank of ground truth similarity scores. Thus, this pipeline is effective at ranking patient pairs' similarity.

Pearson correlation of 0.7608 indicates a strong relationship between the magnitude of predicted and ground truth similarity

TABLE II
PATIENT SIMILARITY EVALUATION: CORRELATION COEFFICIENTS

Spearman Correlation	0.7728
P-value	0.0001698
Pearson Correlation	0.7608
P-value	0.0002461

scores. Thus, this pipeline is effectively able to determine the magnitude of similarity scores.

Low p-values demonstrate that both of these correlations are statistically significant and that these results were not likely a result of randomness. Together, both indicate a strong positive correlation between predicted and ground truth similarity scores, highlighting that this pipeline can be utilized to effectively rank and estimate patient similarity. Thus, this knowledge graph tool can be applied to a variety of clinical patient similarity tasks from clinical trial matching to ICU mortality estimation.

2) Direct Evaluation: In the direct evaluation method run on the given patient data itself, once both the diagnosis-based ground truth and the predicted patient similarity scores were generated, each patient pair was categorized as either a full match, partial match, or no match—based on the absolute difference between the two scores. In particular,

- A difference of **0.1 or less** is considered a **full match**, indicating strong alignment between the predicted and actual similarities.
- A difference of **0.4 or less** is considered a **partial match**, suggesting moderate agreement.
- A difference **greater than 0.4** is labeled as a **no match**, reflecting a discrepancy.

In the initial test run, the outcome across all patient pairs was a full to partial match, demonstrating that the model captures meaningful similarities but still leaves room for improvement in aligning with diagnosis-based ground truth.

3) Benchmark: While the benchmark dataset is still being refined, initial comparisons indicate moderate accuracy between the model's predicted similarity scores (using all features) with the benchmark ground truth.

TABLE III
PATIENT SIMILARITY EVALUATION: PREDICTED VS. GROUND TRUTH SCORES

Patient Pair	Patient [99384712] vs [88246109]
Predicted Similarity	0.321
Ground Truth Similarity	0.100
Match	Full

CONCLUSION

This work introduces Med-KG, a general-purpose clinical knowledge graph constructed from MIMIC-IV data to represent the complete clinical journey—including diagnoses, symptoms, procedures, medications, and provider interactions. Unlike prior task-specific graphs, this broad design supports diverse downstream applications. Its utility is demonstrated through patient similarity analysis, where LLMs reason over relation triples to generate similarity scores between patient pairs. To evaluate performance, diagnosis data was removed to isolate the contribution of non-diagnosis features, and predictions were compared to diagnosis-based ground truth in both controlled and direct evaluation settings. An expert-annotated benchmark is also being developed.

Overall, this work shows the promise of combining large-scale clinical knowledge graphs with LLMs for generalizable healthcare tasks.

REFERENCES

[1] Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. Multimodal biomedical AI. *Nat. Med.*, 28(9):1773–1784, September 2022.

[2] Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1), November 2022.

[3] Yaara Artsi, Vera Sorin, Benjamin S. Glicksberg, Girish N. Nadkarni, and Eyal Klang. Advancing clinical practice: The potential of multimodal technology in modern medicine. *Journal of Clinical Medicine*, 13(20):6246, October 2024.

[4] Qiong Cai, Hao Wang, Zhenmin Li, and Xiao Liu. A survey on multimodal data-driven smart healthcare systems: Approaches and applications. *IEEE Access*, 7:133583–133599, 2019.

[5] Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1), February 2023.

[6] Xuehong Wu, Junwen Duan, Yi Pan, and Min Li. Medical knowledge graph: Data sources, construction, reasoning, and applications. *Big Data Mining and Analytics*, 6(2):201–217, June 2023.

[7] Rajat Mishra and S. Shridevi. Knowledge graph driven medicine recommendation system using graph neural networks on longitudinal medical records. *Scientific Reports*, 14(1), October 2024.

[8] Jiaxin Bai, Tianshi Zheng, and Yangqiu Song. Sequential query encoding for complex query answering on knowledge graphs, 2023.

[9] Jiacheng Xu, Kan Chen, Xipeng Qiu, and Xuanjing Huang. Knowledge graph representation with jointly structural and textual encoding, 2016.

[10] Hejie Cui, Jiaying Lu, Ran Xu, Shiyu Wang, Wenjing Ma, Yue Yu, Shaojun Yu, Xuan Kan, Chen Ling, Liang Zhao, Zhaohui S. Qin, Joyce C. Ho, Tianfan Fu, Jing Ma, Mengdi Huai, Fei Wang, and Carl Yang. A review on knowledge graphs for healthcare: Resources, applications, and promises, 2023.

[11] John Halamka and Paul Cerrato. Knowledge graphs can move healthcare into the future, December 2023.

[12] Shuangjia Zheng, Jiahua Rao, Ying Song, Jixian Zhang, Xianglu Xiao, Evandro Fei Fang, Yuedong Yang, and Zhangming Niu. PharmKG: a dedicated knowledge graph benchmark for bomedical data mining. *Brief. Bioinform.*, 22(4), July 2021.

[13] Xiaolin Zhang and Chao Che. Drug repurposing for parkinson's disease by integrating knowledge graph completion model and knowledge fusion of medical literature. *Future Internet*, 13(1):14, January 2021.

[14] Liang Yao, Jiazhen Peng, Chengsheng Mao, and Yuan Luo. Exploring large language models for knowledge graph completion. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025.