Bridging Stepwise Lab-Informed Pretraining and Knowledge-Guided Learning for Diagnostic Reasoning

Pengfei Hu, *Member, IEEE*, Chang Lu, Fei Wang, *Senior Member, IEEE*, and Yue Ning , *Member, IEEE*

***Abstract*— Despite the growing use of Electronic Health Records (EHR) for AI-assisted diagnosis prediction, most data-driven models struggle to incorporate clinically meaningful medical knowledge. They often rely on limited ontologies, lacking structured reasoning capabilities and comprehensive coverage. This raises an important research question: Will medical knowledge improve predictive models to support stepwise clinical reasoning as performed by human doctors? To address this problem, we propose DuaLK, a dual-expertise framework that combines two complementary sources of information. For external knowledge, we construct a Diagnosis Knowledge Graph (KG) that encodes both hierarchical and semantic relations enriched by large language models (LLM). To align with patient data, we further introduce a lab-informed proxy task that guides the model to follow a clinically consistent, stepwise reasoning process based on lab test signals. Experimental results on two public EHR datasets demonstrate that DuaLK consistently outperforms existing baselines across four clinical prediction tasks. These findings highlight the potential of combining structured medical knowledge with individual-level clinical signals to achieve more accurate and interpretable diagnostic predictions. The source code is publicly available on `https://github.com/humphreyhuu/DuaLK`.**

***Index Terms*— Diagnosis Prediction, Diagnosis Knowledge Graph, LLM, Clinical Reasoning, Stepwise, Lab Tests**

I. Introduction

The digitization of patient information such as EHR has transformed healthcare in terms of data storage, information retrieval, and computational pattern recognition. EHR-based prediction models, including those for diagnosis [1], heart failure [2], and readmission prediction tasks [3], leverage EHR data to forecast patient outcomes. To reduce bias risks of predictions based on data patterns rather than clinical understanding, it is essential to integrate medical domain knowledge from both individual health records and public medical knowledge into the predictive process. (1) Individual knowledge can help model capture progression patterns (e.g. disease complications) of personal health status on EHR datasets [4]–[6]; (2) Public knowledge bases and clinical websites have been used to enhance the understanding of medical concepts [7]–[9], which reveals that even simple hierarchical structures of diseases can improve prediction performance.

Despite recent progress, performing diagnostic decision-making in a manner similar to physicians remains a challenging goal, with several challenges ahead:

1) **Underutilized public medical knowledge with limited quality.** While external knowledge beyond the training data has been widely utilized by previous approaches, these models often rely on simple ontologies (e.g., the hierarchy in ICD-9-CM [10]), which fail to capture the comprehensive medical knowledge required for clinical prediction [4], [7]. Textual and other unstructured features [5], [11], [12] have also been explored to improve predictive performance, but semantic ambiguity and irrelevant context remain challenges.
2) **Lack of diagnostic reasoning when using EHR data.** Some studies [4], [13], [14] take advantage of customized pretraining tasks to fully exploit EHR data, overlooking the causal and procedural nature of clinical decision-making. There are also some models using co-occurrence graphs to learn medical interactions (e.g., complications) [1], [5], [6], but the relations they capture are simple associations, failing to reflect the multi-step diagnostic reasoning that physicians typically follow.
3) **Inefficiency of integrating medical knowledge in EHR.** Most models rely on either public knowledge or individual health data independently, as combining both may incur high computational costs during deployment [15]. However, models that rely on a single source of data fail to capture generalized patterns, leading to biased predictions and degraded robustness.

To overcome these limitations, a stepwise and knowledge-informed framework is required to better align with real-world clinical decision-making [11]. Clinicians typically begin by ordering lab tests informed by patients' symptoms and medical history, interpreting the results, and making a diagnostic decision [16]. Inspired by this process, we introduce a lab-informed pretraining strategy, in which lab tests are

Pengfei Hu is with the Department of Computer Science, Stevens Institute of Technology, Hoboken, NJ 07030 USA (e-mail: phu9@stevens.edu).

Chang Lu was with the Department of Computer Science, Stevens Institute of Technology, Hoboken, NJ 07030 USA. He is now the independent researcher (e-mail: luchang.cs@gmail.com).

Fei Wang is with the Weill Cornell Medicine, Cornell University, Boulder, NY 10065 USA (e-mail: few2001@med.cornell.edu).

Yue Ning is now with the Department of Computer Science, Stevens Institute of Technology, Hoboken, NJ 07030 USA (e-mail: yue.ning@stevens.edu).

treated as intermediate clinical signals. It enables the model to perform diagnostic reasoning in a stepwise manner rather than all at once. To enhance the quality of medical knowledge, a diagnosis-specific KG is constructed by selecting highly relevant clinical entities (e.g., diseases, drugs, phenotypes) and incorporating structured triples transformed from unstructured textual information. To mitigate the high computational cost during deployment, we decouple the KG from the model's prediction phase. This design allows the use of large-scale knowledge graphs without introducing additional inference-time overhead, thereby improving scalability and efficiency.

Therefore, we propose DuaLK, a **Dual**-Expertise Synergy framework, that enhances representation by both **L**ab-informed pretraining and **K**nowledge-guided learning for diagnostic reasoning. We construct a unified KG by merging ontologies and using large language models (LLMs) to capture both hierarchical and semantic relationships. Next, a lab-informed learning is designed to utilize clinical records through graph neural networks and two pretraining proxy tasks based on lab test data to rectify patient-level embeddings. These two expertise are collaborated within an encoder-decoder architecture, enabling mutual refinement and accurate clinical prediction. The main contributions of this paper are as follows:

1) We construct a diagnosis-specific KG upon multiple knowledge sources to support predictive modeling. Experimental results show that our KG enhances diagnostic prediction performance (e.g. F1-score and recall), while maintaining compatibility with various baseline models through a flexible and adaptable embedding strategy.
2) To the best of our knowledge, this is the first work to leverage lab data for pretraining. Different from previous work [2], [17], we model stepwise diagnostic reasoning through lab test assignment and abnormality detection. Our experiments also reveal that incorporating such lab-informed reasoning improves predictive performance.
3) We propose a diagnostic prediction framework explicitly designed to fuse both public and individual medical knowledge. Our approach consistently outperforms current state-of-the-art baselines, demonstrating strong generalization and adaptability—even as data volume or diagnostic code complexity increases.

II. Related Work

A. Clinical Prediction Models

Deep learning models have been widely applied in predictive healthcare to provide guidance for preventive care. Both RNN-based [18]–[22] and CNN-based models [23], [24] have been developed to capture temporal and spatial dependencies for clinical risk predictions. With the rise of graph neural networks (GNNs), many works incorporate ontologies over medical concepts (e.g. diseases) to enrich representations, such as CGL [1], SeqCare [11], and GraphCare [9]. Meanwhile, Transformer-based models like G-BERT [4] and its extensions leverage proxy tasks to extract information from structured EHR data, particularly for patients with limited visit histories [6], [13], [14]. Most of these methods rely on graphs with simple, single-type relations and demonstrate superiority across various EHR prediction tasks, which highlights the potential of enhancing prediction by more complex, multi-relational medical knowledge.

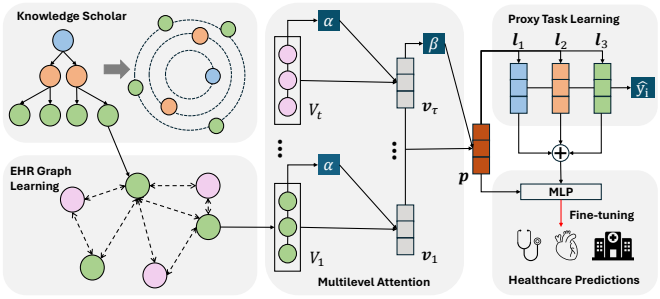


Fig. 1: **Overview of the DuaLK Framework.** We construct a unified KG $\tilde{\mathcal{G}}_H$ and learn disease embeddings $\mathbf{E}_{\mathcal{D}}$, which are then incorporated into the disease–lab EHR graph $\mathcal{G}$ to support graph learning with multil-evel attention, yielding visit-level representations $\{\mathbf{v}_1, \ldots, \mathbf{v}_\tau\}$ and the patient-level embedding $\mathbf{p}$. Next, we apply a stepwise proxy-task learning module on $\mathbf{p}$ to obtain lab-informed representations $\{\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3\}$, which together facilitate downstream clinical predictions $\hat{\mathbf{y}}$.

B. Knowledge-Guided Clinical Prediction

Predictive models that leverage graph and transformer architectures have achieved state-of-the-art performance in various clinical prediction tasks, highlighting their potential to advance predictive healthcare through the use of medical knowledge. Such knowledge can be categorized into individual knowledge and public knowledge bases, as outlined below:

- **Individual Knowledge:** Prior studies such as CGL [1] and Sherbet [5] construct EHR graphs based on interactions (e.g., disease complications) among medical codes in admissions, capturing hidden associations for clinical prediction. Transformer-based models like G-BERT [4] also adopt customized pretraining tasks to exploit extra supervision signals from EHR data, thereby improving convergence and enhancing clinical prediction [6], [25]. However, existing methods either rely on oversimplified associations among medical concepts or employ proxy tasks disconnected from clinical reasoning, limiting their ability to model diagnostic decision processes.
- **Public Knowledge:** External knowledge is derived from public data sources beyond the EHR dataset. Researchers have explored simple ontologies such as the ICD-9 hierarchy [10], retrieval-augmented generation (RAG) methods with unstructured medical descriptions like RAM-EHR [12], and large-scale ontologies for prediction, such as in GraphCare [9]. However, they often emphasize either hierarchical or semantic knowledge in isolation, which may introduce noisy or misaligned information and impair prediction performance.

While public and individual knowledge each provide complementary insights, most existing studies rely on only one of them [15]. Effectively combining both hierarchical and semantic knowledge sources is essential for improving the robustness and generalizability of predictive healthcare models.

C. Clinical Prediction with Multi-aspect Features

In addition to leveraging medical-domain knowledge, combining both structured and unstructured features also enhance predictions. MiME [26] and GCT [2] use graph structures with lab results to augment representations, while CGL [1] and MedGTX [27] leverage unstructured clinical notes to highlight the value of additional information. However, most approaches rely heavily on these multi-aspect features, limiting their applicability to data without those features [17]. Our proposed framework integrates laboratory features as auxiliary data in graph construction and proxy tasks, enhancing predictions even when no desired feature is available.

D. Reasoning in Clinical Prediction

Many studies treat graph learning as a form of reasoning based on neighboring entities [5], [9], [11], which often reduces reasoning to static associations. Reinforcement learning-based methods attempt to emulate clinical inquiry processes through multi-step interactions [28]; however, their reliance on iterative decision-making leads to high training complexity and substantial data requirements, making them difficult to scale and less applicable in real-world clinical settings [29]. More recently, LLMs have been directly used as predictors to simulate clinician-like reasoning [30], [31], a strategy which is prone to hallucinations in medical contexts and involves high computational costs. Overall, most of them fall short in explicitly modeling the stepwise and causally structured nature of clinical reasoning, underscoring the need for more faithful and efficient reasoning frameworks.

III. Methodology

In this section, we propose DuaLK, a dual-expertise framework designed for predictive healthcare that aims to provide comprehensive diagnoses based on public and individual knowledge. An overview of DuaLK is shown in Figure 1.

A. General Notation

An EHR dataset $\mathcal{S}$ is a collection of temporal admission records of N patients $\{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_N\}$, and $\mathbf{p}_u$ denotes patient u. Such patient can also be represented as a sequence of T_u admission records $(V_1^u, V_2^u, .., V_{T_u}^u) \in \mathbf{p}_u$ in chronological order. We omit the patient index u in the following sections and explain our framework using a single patient as an example for clarity. We denote the entire set of medical concepts as $\mathcal{C} = \{c_1, c_2, ..., c_{|\mathcal{C}|}\}$ where $|\mathcal{C}|$ is the vocabulary size of medical concepts. Each admission, like t-th admission $V_t \in (V_1, V_2, .., V_T)$, contains a subset of $\mathcal{C}$. We consider diseases and lab results as medical concepts, denoted by the vocabularies $\mathcal{D}$ and $\mathcal{L}$ with $\mathcal{D}, \mathcal{L} \subset \mathcal{C}$, where diseases are $\{d_1, \ldots, d_{|\mathcal{D}|}\}$ and lab tests are $\{l_1, \ldots, l_{|\mathcal{D}|}\}$.

B. Knowledge Graph Construction

We establish knowledge scholar by converting both existing medical KGs and verified textual knowledge into the structured triplet set via the in-context-learning power of LLMs. Figure 2 illustrates how to augment KG with textual information. In general, we aim to integrate and standardize a uniform KG based on both medical ontologies and textual knowledge.

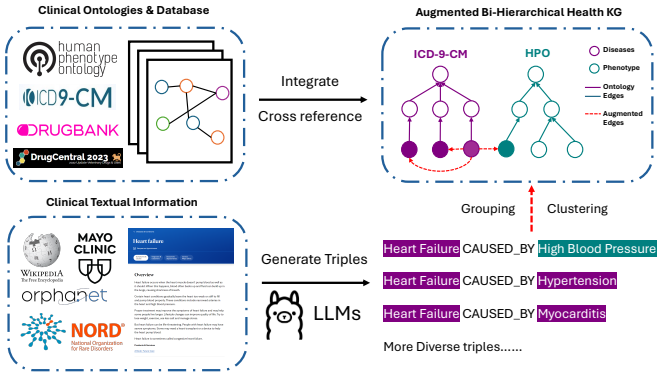


Fig. 2: Overview of Augmented Bi-hierarchical KG

Ontology KG Construction: The objective of this ontology KG is to link medical concepts by merging ontologies from multiple datasets, providing a more complex structure than all-in-one ontologies like UMLS [32] to support precise clinical prediction. Therefore, five standardized ontologies and experimentally derived data are adopted for KG construction, ensuring broad coverage across key medical entities:

1) *DrugBank* [33]: The comprehensive resource containing detailed pharmaceutical knowledge. We retrieved the latest version (5.1.12), published on March 14, 2024. Our focus is on synergistic drug interactions, which represent the bidirectional connections between two drugs.
2) *DrugCentral* [34]: The verified database that provides drug-disease relations, including indications, contraindications, and off-label uses. We utilized the updated SQL Database, released on November 1, 2023, for our study.
3) *HPO* [35]: The Human Phenotype Ontology offers details on phenotype abnormalities associated with diseases. We used the most recent updates, from April 19, 2024, focusing on disease-phenotype and phenotype-phenotype relationships, and extracted verified associations with expertly curated annotations.
4) *ICD-9-CM* [10]: The encoding system is to classify medical conditions for billing purposes, and it underlies the diagnostic information in most EHR data. For our purposes, we extracted the parent-child relationships among codes to describe disease interactions.
5) *SIDER* [36]: The Side Effect Resource provides side effects caused by various drugs. We retrieved this information from the data release dated October 21, 2015.

The HPO and the ICD-9-CM hierarchies serve as the foundation for the resulting KG, with bi-hierarchical structure enhancing its robustness and connectivity. We denote ontology KG as $\mathcal{G}_M$, containing M tuples of head, relation, and tail.

LLM KG Construction: We further extract relational triples from text information by leveraging the in-context learning capabilities of LLMs. Text information is parsed from the following four clinical websites/databases:

1) *Mayo Clinic* [37] maintains a knowledge base with clinical information on over 2000 diseases. We collected Mayo Clinic web knowledge and mainly focus on both "Symptoms & Causes" (Overview, Symptoms, When to

see a doctor, Causes, Risk Factors, Complications, and Prevention) and "Diagnosis & Treatment" parts.

2) *Orphanet* [38] is a clinical database dedicated to rare diseases. We collected data focusing on information about definitions, prevalence, treatment, and clinical descriptions for rare diseases.
3) *Rare Disease Database* [39] contains supplementary knowledge source for uncovered rare diseases. We collected data to retrieve information on disease overviews, symptoms, causes, treatments, and clinical trials.
4) *Wikipedia* is the most well-known online encyclopedia. It designs a specific webpage for ICD-9-CM codes[1]. We retrieved related information for each disease on the separate webpage. For example, the page of "Typhoid Fever" provides detailed information on symptoms, causes, diagnosis, prevention, treatment, epidemiology, history, terminology, and societal impact, offering more comprehensive and diverse knowledge compared to other textual sources.

Incorporating unstructured textual data into structured KG is achieved by leveraging the in-context learning capabilities of LLMs to extract relational triples from curated descriptions. As all inputs originate from verified or expert-curated sources, the risk of hallucination is minimal. A prompting-based approach is adopted instead of retrieval-augmented generation (RAG), showing better empirical performance due to reduced semantic loss. The LLM-generated triples in the format `[ENTITY 1, RELATION, ENTITY 2]` are aggregated to form a knowledge graph $\mathcal{G}_N$ containing N triples. To stabilize inference output, we adopt two strategies: (1) **Iterated Prompting**, where the full input is re-prompted to ensure completeness; and (2) **Re-read Prompting**, where the same prompt is passed repeatedly for refinement. Final triples are manually reviewed before generating the LLM KG. Note that, such is performed using `Llama-3.1-8b` [40], an open-source language model selected for its accessibility and compatibility with local deployment. However, other open-source LLMs are also compatible with the proposed workflow. A detailed prompt example is provided in Appendix-D.

Final KG Integration & Normalization: To get the final KG $\mathcal{G}_H$, we integrate the ontology KG $\mathcal{G}_M$ and the LLM KG $\mathcal{G}_N$:

$$\mathcal{G}_M = \bigcup\nolimits_M (\mathbf{h}_m^{(M)}, \mathbf{r}_m^{(M)}, \mathbf{t}_m^{(M)}) \tag{1}$$

$$\mathcal{G}_N = \bigcup\nolimits_{x,y} \bigcup\nolimits_N (\mathbf{h}_n^{(N)}, \mathbf{r}_n^{(N)}, \mathbf{t}_n^{(N)}) \tag{2}$$

$$\mathcal{G}_H = \mathcal{G}_M \cup \mathcal{G}_N,\ \tilde{\mathcal{G}_H} = \textsc{Normalize}(\mathcal{G}_H), \tag{3}$$

where x denotes the number of iteration for iterative running and and y denotes the number of times the generated output is re-evaluated (re-reading) to refine the quality of triples.

Our approach for normalizing $\mathcal{G}_H$ involves two key steps: **(1) Uniform encoding system for entities:** In addition to phenotype nodes sourced from HPO, we align disease and drug entities by using ICD-9-CM [10] and ATC [41] identifiers. A cross-referencing method is employed to convert $\mathcal{G}_M$ and $\mathcal{G}_N$ into a shared hierarchy. Given that drugs are not the primary focus, we use ATC-4 to simplify the numerous synergistic interactions in DrugBank. **(2) Clustering duplicated or unmatched triples:** The prompting process is run multiple times, which can generate duplicated nodes and edges. Moreover, cross-referencing files may not cover all concepts, and unmatched entities leave noise in KG. To address these problems, we calculate cosine similarity for nodes and edges based on their embeddings. We then group entities within categories (e.g. "disease" and "drug") with a predefined threshold θ, and the hierarchical clustering is applied to reorganize duplicated edges.

[1]The Wikipedia ICD-9-CM webpage: `https://en.wikipedia.org/wiki/List_of_ICD-9_codes`

TABLE I: Statistical Comparison among KGs. (* means results might fluctuate by different parameter settings.)

KG	# Nodes	# Edge Types	# Triples
GraphCare: [Disease, Procedure, Drug, Other]			
- Ontology*	10,805	54	81,073
- GPT-4 output*	4,599	752	31,325
- Final KG*	12,284	785	104,460
PrimeKG: [Disease, Drug, Protein ... 10 in total]			
- Public KG	**129,375**	30	**4,050,249**
- Text Description	-	-	-
DuaLK KG: [Disease, Drug, Phenotype, Other]			
- Ontology KG	32,941	6	477,992
- LLM KG*	42,436	**3,642**	82,191
- Final KG*	52,604	**3,645**	560,183

Advancement Analysis: To demonstrate the advancements of our proposed KG, we compare it with two recent health-related KGs: GraphCare [9] and PrimeKG [42]. A statistical summary of these KGs is presented in Table I. The analysis highlights the following advantages of our KG:

1) Compared to GraphCare [9], our KG is larger in scale upon ICD-9-CM and HPO systems, offering both hierarchical and diverse knowledge. In contrast, GraphCare relies on zero-shot generation from LLMs without fine-tuning, which can result in incorrect triples due to hallucination [12], along with unrelated entities (e.g. social impact) irrelevant to the clinical task. Hence, our KG provides more diverse and precise information in downstream prediction tasks.
2) Compared to PrimeKG [42], our KG includes more edge types, primarily augmented by LLM-generated triples. The LLM reasoning process not only enhances connectivity within the KG but also converts unstructured text into structured triples, making information retrieval more efficient from single-modal data.

Although PrimeKG covers a broader range of medical concepts, it is more prone to data quality issues that can introduce bias into EHR predictions. For example, genes and proteins included in PrimeKG are seldom considered in physicians' decision-making process. In contrast, our KG contains numerous normalized triples with diverse edges, making it highly reproducible and well-suited for various diagnosis-related healthcare predictions. Such analysis will then be supported by the results of the ablation study.

C. Knowledge Graph Embedding

Traditional KG embedding methods often emphasize either contextual or hierarchical information, limiting their ability to represent both simultaneously. Our constructed KG encodes both types of signals, requiring an embedding method capable of capturing their interplay. To meet this need, we adopt polar-space embedding [43]. A comparison with other KG embedding methods [44]–[47] is provided in Appendix-A.

We leverage two properties of the polar coordinate: (1) **Radial coordinate:** map entities across different levels of the hierarchy; (2) **Angular coordinate:** map entities at the same level based on contextual information. We denote entities, including head $\mathbf{h}$ and tail $\mathbf{t}$, and relation embeddings as $\mathbf{r}$. Each entry of $\mathbf{h}_{r,a}$ or $\mathbf{t}_{r,a}$ corresponds to a radial or angular value, while each entry of $\mathbf{r}_{r,a}$ represents a scaling transformation between the corresponding head and tail. We then formulate radial and angular coordinates

$$\mathbf{h}_r \circ \mathbf{r}_r = \mathbf{t}_r, \quad \mathbf{h}_r, \mathbf{t}_r, \mathbf{r}_r \in \mathbb{R}^k \tag{4}$$

$$(\mathbf{h}_a + \mathbf{r}_a) \bmod 2\pi = \mathbf{t}_a, \quad \mathbf{h}_a, \mathbf{t}_a, \mathbf{r}_a \in [0, 2\pi)^k \tag{5}$$

According to the coordinates for triples, we can further calculate radial and angular distances as follows:

$$d_r(\mathbf{h}_r, \mathbf{t}_r) = \|\mathbf{h}_r \circ \mathbf{r}_r - \mathbf{t}_r\|_2, \tag{6}$$

$$d_a(\mathbf{h}_a, \mathbf{t}_a) = \|\sin((\mathbf{h}_a + \mathbf{r}_a - \mathbf{t}_a)/2)\|_1, \tag{7}$$

where $\|\cdot\|_1$, $\|\cdot\|_2$ denote ℓ_1, ℓ_2 norms, and $d(\mathbf{h}, \mathbf{t})$ is the weighted sum of both radial and angular distances. We can get $\mathbf{h}$ and $\mathbf{t}$ by concatenating both radial and angular vectors. Such embedding model can then be pretrained individually through link prediction tasks, which helps us understand priors from KG into trainable embeddings. The regular negative sampling loss function L is adopted for optimization:

$$L = -\log \sigma(\gamma - d(\mathbf{h}, \mathbf{t})) - \sum_{i=1}^{n} \log \sigma(d(\mathbf{h}', \mathbf{t}') - \gamma), \tag{8}$$

where γ and σ are marginal value and sigmoid function respectively, and (h', r, t') is a negative triple. Therefore, we can get diagnosis embedding matrix $\mathbf{E}_{\mathcal{D}}$ for condition codes. Note that, we also experiment with a simpler objective function (i.e., bringing connected nodes closer while pushing unconnected nodes apart) as a replacement for the link prediction task. However, due to the complexity of the KG, the model fails to fully learn the positions of nodes (in Section V-B).

D. Graph Learning

The continual enhancement process unfolds in two steps: graph learning and proxy-task learning, which together refine the priors by fully leveraging EHR data. On the first stage, we construct a disease-complication graph $\mathcal{G}$ including both condition and laboratory codes. Compared to regular complication graphs involving only diseases, we have broader scope to explain complication in terms of shared abnormal lab tests in the same admission. To represent the complications between two diseases via patient admissions, we add bidirectional edges (c_i, c_j), (c_i, c_m), and (c_j, c_m) in graph $\mathcal{G}$, where $c_i, c_j \in \mathcal{D}$ are diagnostic codes and $c_m \in \mathcal{L}$ is a laboratory code. Therefore,

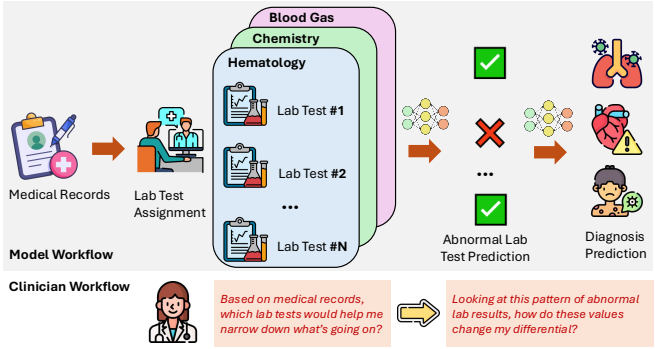


Fig. 3: Illustration of the Lab-Informed Reasoning Process

we then compute co-occurrence matrix $\mathbf{B} \in \mathbb{N}^{|\mathcal{C}|\times|\mathcal{C}|}$ and adjacent matrix $\mathbf{A} = (1-\varphi)\mathbf{B} + \varphi\mathbf{I} \in \mathbb{R}^{|\mathcal{C}|\times|\mathcal{C}|}$ considering self loop based on EHR graph $\mathcal{G}$, where $\mathbf{I}$ is an identity matrix. A GNN with L layers is adopted by EHR graph $\mathcal{G}$ and node features $\mathbf{E}$, where we initialize diseases by priors $\mathbf{E}_{\mathcal{D}}$ from KG. Then, the hidden representation $\mathbf{H}^{(l)}$ can be calculated by input $\mathbf{H}^{(l-1)}$ through the l-th Convolution layer: After the last GNN layer, we obtain the hidden representation of diseases $\mathbf{X} = \text{GNN}(\mathbf{A}, \mathbf{E}) = \mathbf{H}^{(L)}$, which can be further refined through the bi-attention mechanism across codes and admissions for patient embedding $\mathbf{p}$:

$$\mathbf{v}_\tau = \sum_{i=1}^{n} \alpha_\tau^i \mathbf{x}_i, \quad \mathbf{p} = \sum_{\tau=1}^{T} \beta_\tau \tilde{\mathbf{v}}_\tau, \tag{9}$$

where α_τ^i and β_τ are attention scores computed as follows:

$$\mathbf{z}_i = \tanh(\mathbf{W}_c \mathbf{x}_i), \quad \mathbf{r}_\tau = \tanh(\mathbf{W}_v \sigma(\mathbf{W}_u \mathbf{v}_\tau)) \tag{10}$$

$$\alpha_\tau^i = \frac{\exp(\mathbf{z}_i)}{\sum_{j=1}^{n} \exp(\mathbf{z}_j)}, \quad \beta_\tau = \frac{\exp(\mathbf{r}_\tau)}{\sum_{\tau=1}^{T} \exp(\mathbf{r}_\tau)} \tag{11}$$

$\mathbf{W}_c$, $\mathbf{W}_u$, and $\mathbf{W}_v$ are weighted matrices. The attention score α_τ measures the distribution of medical codes of admission, and β measures the distribution across admissions. Note that, we project each admission $\mathbf{v}_\tau$ to $\tilde{\mathbf{v}}_\tau$ for fitting the patient dimension through a single layer with $\mathbf{W}_u$.

The proposed encoder module ensures that both the visit-level diagnostic information and the longitudinal history of a patient's health are effectively captured in the final patient embedding $\mathbf{p}$, laying a robust foundation for the subsequent self-supervised learning phase.

E. Lab-Informed Pretraining

Most existing methods [4]–[6] adapt pretraining tasks from language models to exploit the whole EHR dataset, but fewer of them consider whether definitions are clinically meaningful. To bridge this gap, we harness laboratory insights to help predictive models better mimic the clinical reasoning process of a human doctor. The proposed proxy-task learning combines two consecutive steps, (1) **lab test assignment** and (2) **abnormality detection**, as shown in Figure 3.

The EHR dataset for the proxy task learning is denoted as $\mathcal{S}'$ with N' patients in total. Since those single-admission

patients and final admission records of multi-admission patient lack labels regarding their next admission, we assume that the learned representations $\{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{N'}\}$ will be able to reflect their medical histories (i.e. historical lab test results). Considering diverse patterns across different lab tests within the vocabulary $\mathcal{L} = \{l_1, l_2, \ldots, l_{|\mathcal{L}|}\}$, we further categorize lab tests into several distinct clusters. Note that, the categorical information of each test is documented in most EHR datasets. Here we take lab tests in MIMIC datasets as an example:

1) **Hematology** (# items: 417) Analyze blood components like red and white blood cells to monitor conditions.
2) **Chemistry** (# items: 286) Evaluate chemical factors in blood and provide information about organ's function.
3) **Blood Gas** (# items: 35) Measure the levels of gas in blood to assess respiratory and metabolic functions.

However, it still remains challenging to enable models to grasp the diagnostic reasoning process. Typically, a doctor first forms preliminary diagnostic hypotheses based on symptoms and medical history, orders relevant lab tests accordingly, and then confirms the diagnosis based on abnormal test results. This sequential reasoning underscores how to quantify conditional probabilities, which assign lab tests before observing abnormal outcomes, for our proxy-task learning module.

For better demonstration, we then separate the vocabulary $\mathcal{L}$ into $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ in order and take single patient as an example. Since we aim to decode embedding $\mathbf{p}$ into the probability distribution $\hat{\mathbf{y}}$: $\hat{\mathbf{y}}_i = P(l_i \in \mathcal{L}_m|\mathbf{p})$ for each category m, it is intuitive that the equation can be parameterized by a multi-layer perceptron (MLP) for estimation $\hat{\mathbf{y}} = \sigma(\text{MLP}(\mathbf{p}))$. However, such naïve method has limitation in mimicking the conditional probability for both assignment or abnormality step. Assuming assigned and abnormal lab results are denoted as $\hat{\mathbf{y}}^{a,m}$ and $\hat{\mathbf{y}}^{b,m}$ respectively for certain category m, the abnormal lab results $\hat{\mathbf{y}}^{b,m}$ is equivalent to a joint probability $P\left(p^a_{l_i} \in \mathcal{L}_m, p^b_{l_i} \in \mathcal{L}_m|\mathbf{p}\right)$ regarding assignment as priors, which can be represented as follow with given $\mathbf{p}$:

$$\begin{aligned}\hat{\mathbf{y}}^{b,m} &= P\left(p^a_{l_i} \in \mathcal{L}_m, p^b_{l_i} \in \mathcal{L}_m|\mathbf{p}\right) \\ &= P\left(p^b_{l_i} \in \mathcal{L}_m|p^a_{l_i} \in \mathcal{L}_m, \mathbf{p}\right) \times P\left(p^a_{l_i} \in \mathcal{L}_m|\mathbf{p}\right) \quad (12)\end{aligned}$$

where i represent the index within the vocabulary $\mathcal{L}$, so that $p^a_{l_i}$ and $p^b_{l_i}$ are probabilities of the i-th assigned and abnormal lab results. Even though we separate our target $\hat{\mathbf{y}}^b$ by two consecutive conditional components, it is still difficult to parameterize $\text{P}\left(p^b_{l_i} \in \mathcal{L}_m|p^a_{l_i} \in \mathcal{L}_m, \mathbf{p}\right)$ for estimation. Hence, we adopt the constraint relaxation method and parameterize $\text{P}\left(p^b_{l_i} \in \mathcal{L}_m|\mathbf{p}\right)$ as the lower bound approximation:

$$P\left(p^b_{l_i} \in \mathcal{L}_m|p^a_{l_i} \in \mathcal{L}_m, \mathbf{p}\right) \geq P\left(p^b_{l_i} \in \mathcal{L}_m|\mathbf{p}\right) \quad (13)$$

$$P\left(p^a_{l_i} \in \mathcal{L}_m|\mathbf{p}\right) = \sigma(\mathbf{w}_a\mathbf{p})_{l_i} \quad (14)$$

$$P\left(p^b_{l_i} \in \mathcal{L}_m|\mathbf{p}\right) = \sigma(\mathbf{w}_b\mathbf{p})_{l_i} \quad (15)$$

$$\hat{\mathbf{y}}^{b,m}_i = [\sigma(\mathbf{w}^m_a\mathbf{p}) \otimes \sigma(\mathbf{w}^m_b\mathbf{p})]_{l_i} \quad (16)$$

$$\begin{aligned}L^m = \omega \times \text{BCE}(\hat{\mathbf{y}}^{a,m}, \mathbf{y}^{a,m}) \\ + (1-\omega) \times \text{BCE}(\hat{\mathbf{y}}^{b,m}, \mathbf{y}^{b,m})\end{aligned} \quad (17)$$

where $\mathbf{w}^m_a, \mathbf{w}^m_b \in \mathbb{R}^{|\mathcal{L}_m| \times p}$ decodes patient embedding $\mathbf{p}$ to the probability of the laboratory codes for assignment and abnormality respectively. In L^m, $\text{BCE}(\cdot)$ denotes the binary cross-entropy loss and ω is the hyperparameter scaling two losses. $\mathbf{y}^{a,m}$ and $\mathbf{y}^{b,m}$ are ground truth for prediction. Note that, as a self-supervised learning problem, such design help model learn the conditional relationship between both steps, which further augment the robustness of learned parameters. Moreover, to ensure the lab embeddings are well-aligned with each category, the proxy task will be trained in two phrases:

- **Joint Training.** We get preliminary trained parameters in encoder Θ_e and optimize prediction through the average loss across three categories, where we can adjust patient embedding $\tilde{\mathbf{p}}$ through updated parameters.
- **Individual Training.** We individually update parameters in decoder Θ_d after maintaining Θ_e unchanged given from the former phase, where we get refined lab embeddings $\tilde{\mathbf{l}}_1$, $\tilde{\mathbf{l}}_2$, and $\tilde{\mathbf{l}}_3$ through updated decoder.

After two training phases, we get pretrained parameter $\Theta = [\Theta_d; \Theta_e]$ for the predictive model. Similar to the real scenario, doctors always provide diagnoses relying on their historical conditions and abnormal lab results. By training model to understand such underlying laboratory pattern, it provides a pathway to deeply integrate additional laboratory knowledge into the model, ensuring that the predictions are not only accurate but also clinically meaningful.

F. Fine-Tuning and Inference

The proposed model can be trained with EHR data either with or without lab tests, making it adaptable to data missing desired features. If training with only diagnostic information, we can directly compute the estimated output $\hat{\mathbf{y}}'$ and the downstream loss $\mathcal{L}'$ for optimization:

$$\hat{\mathbf{y}}'_{\text{direct}} = \sigma(\mathbf{W}\mathbf{p}) \in \mathbb{R}^o \quad (18)$$

$$L'_{\text{direct}} = \text{Loss}(\mathbf{y}', \hat{\mathbf{y}}') \quad (19)$$

If lab features are included, the model leverages pretrained parameters Θ to get both adjusted patient embedding $\tilde{\mathbf{p}}$ and lab embeddings $\tilde{\mathbf{l}}_1$, $\tilde{\mathbf{l}}_2$, and $\tilde{\mathbf{l}}_3$. The downstream outputs $\hat{\mathbf{y}}'$ and fine-tuning loss $\mathcal{L}'$ are then calculated as follows:

$$\hat{\mathbf{y}}'_{\text{finetune}} = \sigma(\mathbf{W}[\tilde{\mathbf{l}}_1 : \tilde{\mathbf{l}}_2 : \tilde{\mathbf{l}}_3]|\tilde{\mathbf{p}}) \in \mathbb{R}^o \quad (20)$$

$$L'_{\text{finetune}} = \text{Loss}(\mathbf{y}', \hat{\mathbf{y}}'|\Theta) \quad (21)$$

Here, $\mathbf{W} \in \mathbb{R}^{o \times p}$ is the weight matrix and [:] means concatenation. Note that $\mathbf{y}'$, σ, o, and $\mathcal{L}'$ depend on the specific tasks. During optimization with backpropagation, the embedding matrix and parameters remain learnable.

IV. Experimental Setup

A. Datasets

To evaluate our proposed model, we utilize two public and widely-used EHR datasets: MIMIC-III [48] and MIMIC-IV [49]. MIMIC-III includes 7,493 patients with multiple visits ($T \geq 2$) recorded between 2001 and 2012, while MIMIC-IV contains 85,155 such patients from 2008 to 2019. To ensure a comparable sample size and avoid temporal

TABLE II: Statistics of MIMIC-III and MIMIC-IV datasets

Dataset	MIMIC-III	MIMIC-IV
# patients	7,493	10,000
Max. # visit	42	55
Avg. # visit	2.66	3.66
# codes	4,880	6,102
Max. # codes per visit	39	50
Avg. # codes per visit	13.06	13.38

overlap with MIMIC-III, we randomly sample 10,000 patients from MIMIC-IV who were admitted between 2013 and 2019. Table II summarizes the basic statistics of the MIMIC-III and MIMIC-IV datasets. We divide both datasets into training, validation, and test sets by patient: 6,000/493/1,000 patients for MIMIC-III and 8,000/1,000/1,000 for MIMIC-IV, respectively. For each patient, the most recent visit is used as the prediction target, with all previous visits serving as input features. All diagnoses are encoded using the ICD-9-CM coding system. There are 36.70% (15.70%) positive samples and 63.30% (85.30%) negative samples in the test set of MIMIC-III (MIMIC-IV).

B. Baselines

To evaluate the performance of our proposed model, we selected the following 10 EHR models as comparison methods: (1) **RNN/CNN-based models**: RETAIN [18] and Timeline [21]. (2) **Graph/KG-based models**: GCT [2], CGL [1], SeqCare [11], and RAM-EHR [12]. (3) **Transformer-based models**: G-BERT [4], HiTANet [50], Med-BERT [25], and GT-BEHRT [6]. Note that although GraphCare [9] is another KG-based method, it is not included as a baseline in experiments because its defined KGs cannot be directly transferred from CCS (only about 330 diseases) to ICD-9-CM codes. However, we feed our constructed KG into GraphCare framework for comparison as shown in Table V.

C. Tasks & Evaluation Metrics.

We conduct our experiments on two tasks as downstream examples for clinical prediction, following the similar settings as in previous studies [1], [7]:

1) **Diagnosis Prediction.** This task involves multi-label prediction, where the goal is to predict all condition codes for the next admission based on admission history.
2) **Heart Failure Prediction (HF).** This is a binary classification task that predicts whether patients will be diagnosed with heart failure in their next admission.

Given the label imbalance in EHR data, we use the weighted F_1 score (w-F_1) and recall at k ($R@k$) as evaluation metrics for diagnosis prediction and use the area under the ROC curve (AUC) and the F_1 score for HF prediction.

D. Implementation Details

During the transformation of textual information into triples via LLMs, each inference undergoes iterative prompting and re-read prompting strategies with 2 and 1 rounds, respectively. Polar-space KG embeddings are used as node features in the graph learning module, with vector size $2k = 2000$ ($k = 1000$ for both hierarchical and semantic components). The disease-lab graph is modeled using 2 GAT convolution layers with hidden dimensions of 256, and default settings elsewhere. Attention modules adopt dimensions $a = 256$ and $b = 256$, with a dropout rate of 0.2 applied to each. Lab test embeddings are generated using three decoders, each implemented as a three-layer MLP with hidden sizes of 256 and 128, and a uniform dropout rate of 0.4 across all layers for both pretraining and direct training. A two-layer classifier with 256 hidden units and a dropout rate of 0.5 is used to transform concatenated features into logits for downstream tasks. KG embeddings are trained for 180,000 steps to capture hierarchical and semantic relations. Proxy-task learning is performed with 10 epochs for both joint and individual training. We adopt 500 and 50 training epochs for diagnosis and HF prediction, respectively. The Adam optimizer is employed with an initial learning rate of 0.001, decaying by a factor of 0.5 every 50 (5) epochs for diagnosis (HF) prediction. Hyperparameters for DuaLK are selected on the validation set, and baseline are defined by following their recommended settings. All experiments are conducted using Python 3.10 and PyTorch 2.3.1 with CUDA 12.4 on a system equipped with two AMD EPYC 9254 processors, 528 GB RAM, and four NVIDIA L40S GPUs.

V. Experimental Results

A. Main Results

Table III demonstrates that DuaLK consistently outperforms existing baselines across both diagnosis and heart failure prediction tasks on both MIMIC-III and MIMIC-IV datasets. For diagnosis prediction, we can observe that DuaLK boosts the weighted F_1 score by about 6% over the best baseline on MIMIC-III. Similar trends are observed in the R@10 and R@20 metrics, where DuaLK surpasses the strongest baselines by around 4–5%. The improvements on MIMIC-IV are even more pronounced, since DuaLK reaches a w-F_1 of 29.87%, an improvement of nearly 11%. Regarding HF prediction, the gains are somewhat smaller but remain consistent. Note that all reported improvements are statistically significant, with significance verified using two tests as detailed in Appendix-B. Moreover, we further extend our model to two modified tasks on the MIMIC-III dataset, as shown in Table IV.

1) Real-time (RT) Diagnosis: To fairly compare performance of all baselines, we feed lab tests information for all baselines and proposed model, which not only augments representation for diagnosis prediction but also mimics real-time diagnosis of physician within the beginning of an admission. We can observe that the proposed model still outperforms all baselines, which has even a larger improvement on w-F_1 over regular diagnosis. We suppose that DuaLK, which originally relies on lab tests for auxiliary diagnosis, will be more suitable for lab features input than other baselines.

2) Non-chronic (NC) Diagnosis: Since MIMIC-III collects ICU data, we remove all chronic diseases and focus on predicting acute diseases. Concretely, there are 142 codes removed by

TABLE III: Prediction Results on MIMIC-III and MIMIC-IV for Diagnosis (Task 1) and HF (Task 2) Prediction. We report the average performance (%) and standard deviation (in bracket) of each model over 10 runs. The best results are highlighted.

Model	Task 1: Diagnosis Prediction						Task 2: Heart Failure Prediction			
	MIMIC-III			MIMIC-IV			MIMIC-III		MIMIC-IV	
	w-F1	R@10	R@20	w-F1	R@10	R@20	AUC	F1	AUC	F1
RETAIN	$18.37_{(0.8)}$	$32.12_{(0.8)}$	$32.54_{(0.6)}$	$23.11_{(0.8)}$	$37.32_{(0.8)}$	$40.15_{(0.6)}$	$83.21_{(0.3)}$	$71.32_{(0.2)}$	$84.14_{(0.3)}$	$71.23_{(0.2)}$
Timeline	$20.46_{(0.2)}$	$30.73_{(0.1)}$	$34.83_{(0.1)}$	$23.76_{(0.2)}$	$37.89_{(0.1)}$	$40.87_{(0.1)}$	$82.34_{(0.3)}$	$71.03_{(0.2)}$	$83.45_{(0.3)}$	$72.30_{(0.2)}$
GCT	$20.66_{(0.2)}$	$32.73_{(0.2)}$	$35.44_{(0.2)}$	$24.16_{(0.2)}$	$38.24_{(0.2)}$	$41.65_{(0.2)}$	$82.08_{(0.3)}$	$70.35_{(0.2)}$	$84.80_{(0.3)}$	$69.52_{(0.2)}$
CGL	$22.63_{(0.2)}$	$33.64_{(0.3)}$	$37.87_{(0.2)}$	$25.74_{(0.2)}$	$39.23_{(0.3)}$	$42.67_{(0.2)}$	$84.19_{(0.2)}$	$71.77_{(0.1)}$	$87.91_{(0.2)}$	$70.71_{(0.3)}$
SeqCare	$21.78_{(0.1)}$	$34.17_{(0.2)}$	$36.46_{(0.3)}$	$24.39_{(0.1)}$	$38.42_{(0.2)}$	$41.62_{(0.3)}$	$81.55_{(0.2)}$	$68.78_{(0.1)}$	$85.55_{(0.2)}$	$69.82_{(0.4)}$
RAM-EHR	$23.71_{(0.1)}$	$37.97_{(0.2)}$	$40.18_{(0.2)}$	$27.01_{(0.1)}$	$42.86_{(0.2)}$	$46.92_{(0.2)}$	$82.88_{(0.1)}$	$72.03_{(0.1)}$	$84.80_{(0.2)}$	$72.34_{(0.5)}$
G-BERT	$22.28_{(0.3)}$	$35.62_{(0.2)}$	$36.46_{(0.2)}$	$25.12_{(0.3)}$	$39.91_{(0.2)}$	$43.25_{(0.2)}$	$81.50_{(0.2)}$	$71.18_{(0.1)}$	$85.76_{(0.2)}$	$72.88_{(0.1)}$
HiTANet	$23.15_{(0.2)}$	$34.68_{(0.3)}$	$35.97_{(0.1)}$	$24.53_{(0.2)}$	$38.42_{(0.3)}$	$41.89_{(0.1)}$	$85.13_{(0.1)}$	$73.15_{(0.2)}$	$86.34_{(0.4)}$	$71.35_{(0.2)}$
Med-BERT	$21.68_{(0.1)}$	$33.47_{(0.1)}$	$36.53_{(0.1)}$	$23.58_{(0.1)}$	$36.79_{(0.1)}$	$40.45_{(0.1)}$	$81.36_{(0.1)}$	$69.54_{(0.1)}$	$83.61_{(0.5)}$	$70.46_{(0.1)}$
GT-BEHRT	$23.92_{(0.1)}$	$38.64_{(0.2)}$	$39.97_{(0.1)}$	$26.97_{(0.3)}$	$43.07_{(0.4)}$	$47.19_{(0.2)}$	$83.24_{(0.1)}$	$74.12_{(0.1)}$	$87.43_{(0.3)}$	$72.26_{(0.2)}$
DuaLK	$\mathbf{25.37}_{(0.3)}$	$\mathbf{40.52}_{(0.2)}$	$\mathbf{41.86}_{(0.3)}$	$\mathbf{29.87}_{(0.3)}$	$\mathbf{45.66}_{(0.2)}$	$\mathbf{51.73}_{(0.3)}$	$\mathbf{86.53}_{(0.1)}$	$\mathbf{75.35}_{(0.1)}$	$\mathbf{90.32}_{(0.2)}$	$\mathbf{73.54}_{(0.1)}$

TABLE IV: Supplementary Prediction Results on MIMIC-III for Real-time and Non-chronic Diagnosis Prediction.

Models	RT Diagnosis			NC Diagnosis	
	w-F_1	R@10	R@20	w-F_1	R@10
RETAIN	$19.56_{(0.2)}$	$32.18_{(0.3)}$	$33.67_{(0.3)}$	$18.22_{(0.5)}$	$30.15_{(0.4)}$
Timeline	$21.12_{(0.2)}$	$33.05_{(0.3)}$	$34.23_{(0.3)}$	$18.04_{(0.3)}$	$29.58_{(0.4)}$
GCT	$22.54_{(0.2)}$	$35.67_{(0.3)}$	$37.10_{(0.4)}$	$18.95_{(0.3)}$	$30.08_{(0.3)}$
CGL	$24.13_{(0.1)}$	$38.21_{(0.3)}$	$39.64_{(0.4)}$	$19.45_{(0.3)}$	$30.52_{(0.5)}$
SeqCare	$23.21_{(0.1)}$	$36.51_{(0.1)}$	$37.75_{(0.3)}$	$19.13_{(0.2)}$	$30.34_{(0.4)}$
RAM-EHR	$24.19_{(0.1)}$	$38.74_{(0.2)}$	$41.81_{(0.2)}$	$19.56_{(0.3)}$	$31.27_{(0.4)}$
G-BERT	$24.35_{(0.3)}$	$37.12_{(0.5)}$	$39.35_{(0.4)}$	$19.26_{(0.2)}$	$31.96_{(0.3)}$
HiTANet	$25.67_{(0.3)}$	$39.72_{(0.4)}$	$41.85_{(0.4)}$	$18.81_{(0.3)}$	$31.08_{(0.4)}$
Med-BERT	$22.74_{(0.3)}$	$35.27_{(0.4)}$	$36.54_{(0.4)}$	$17.94_{(0.3)}$	$30.84_{(0.4)}$
GT-BEHRT	$24.53_{(0.4)}$	$40.31_{(0.2)}$	$41.45_{(0.3)}$	$19.93_{(0.3)}$	$33.53_{(0.4)}$
DuaLK	$\mathbf{26.89}_{(\mathbf{0.3})}$	$\mathbf{42.86}_{(\mathbf{0.1})}$	$\mathbf{43.15}_{(\mathbf{0.2})}$	$\mathbf{21.69}_{(\mathbf{0.2})}$	$\mathbf{34.35}_{(\mathbf{0.4})}$

TABLE V: Ablation Results of the Variants of DuaLK.

Model	Diagnosis Prediction			HF Prediction	
	w-F1	R@10	R@20	AUC	F1
DuaLK_a	$21.21_{(0.2)}$	$35.25_{(0.3)}$	$36.46_{(0.3)}$	$80.75_{(0.4)}$	$69.81_{(0.3)}$
DuaLK_b	$22.32_{(0.3)}$	$36.28_{(0.3)}$	$36.95_{(0.4)}$	$81.66_{(0.3)}$	$70.01_{(0.3)}$
DuaLK_c	$23.15_{(0.3)}$	$37.78_{(0.4)}$	$38.79_{(0.4)}$	$83.16_{(0.3)}$	$72.73_{(0.3)}$
DuaLK_d	$22.49_{(0.3)}$	$34.86_{(0.3)}$	$35.79_{(0.4)}$	$80.90_{(0.3)}$	$69.24_{(0.3)}$
DuaLK_e	$24.58_{(0.3)}$	$39.78_{(0.4)}$	$40.46_{(0.4)}$	$83.92_{(0.3)}$	$73.62_{(0.3)}$
DuaLK_f	$24.96_{(0.3)}$	$40.16_{(0.3)}$	$40.87_{(0.4)}$	$85.81_{(0.3)}$	$73.96_{(0.3)}$
PrimeKG_g	$23.74_{(0.3)}$	$38.19_{(0.3)}$	$39.52_{(0.4)}$	$84.73_{(0.3)}$	$72.85_{(0.3)}$
GraphCare_h	$24.14_{(-)}$	$39.89_{(-)}$	$40.14_{(-)}$	$83.73_{(-)}$	$\mathbf{76.28}_{(-)}$
DuaLK	$\mathbf{25.37}_{(\mathbf{0.3})}$	$\mathbf{40.52}_{(\mathbf{0.3})}$	$\mathbf{41.86}_{(\mathbf{0.3})}$	$\mathbf{86.53}_{(\mathbf{0.3})}$	$75.35_{(\mathbf{0.3})}$
- *w/o lab*	$24.21_{(0.3)}$	$39.47_{(0.3)}$	$40.35_{(0.3)}$	$85.73_{(0.3)}$	$73.92_{(0.3)}$

the chronic condition indicator [2] for ICD-9-CM, and most of them frequently appears in admission records. Therefore, all models suffer performance degradations compared to the regular setting, and DuaLK still achieves superior performance. We conjecture that the removal of chronic diseases in admissions skews the distribution, making rare diseases more prominent and, consequently, making the task more challenging.

Overall, these results suggest that the integration of public knowledge from KGs and individual patient data is crucial. DuaLK not only bridges the gap between structured medical information and dynamic EHR data but also outperforms models that rely on either temporal learning, medical ontology, or Transformer-based architectures alone.

B. Ablation Study

To evaluate the effectiveness of involved modules in DuaLK, we compare performance of variants with some modules removed or replaced. We use both future diagnosis and HF prediction tasks on MIMIC-III as examples:

1) **DuaLK_a**: Replacing the embedding of the Diagnosis KG with `GloVe`, as used in GRAM [7], to only focus on hierarchical information for diseases.
2) **DuaLK_b**: Replacing the embedding of the Diagnosis KG with `text-embedding-3-large` [51] to only retrieve semantic information for diseases.
3) **DuaLK_c**: Replacing the link prediction tasks with minimizing distances among connected codes and maximizing distances among unconnected ones.
4) **DuaLK_d**: Removing the Graph Learning module and directly feeding KG embeddings into the Attention layer.
5) **DuaLK_e**: Replacing EHR graph by a simplified graph with disease complications, excluding lab test nodes.
6) **DuaLK_f**: Replacing Attention by GRU modules.
7) **PrimeKG_g**: Using PrimeKG as the KG for initialization of disease embeddings by mapping identifier from MONDO to ICD-9-CM.
8) **$\text{GraphCare}_\text{h}$**: Feeding learned embeddings from the Diagnosis KG into the predictor used by Graphcare aligning to the same encoding system.

The results of ablation variants are shown in Table V. The comparison with $\text{DuaLK}_{\text{a-b}}$ shows that both hierarchical and semantic information in the KG contribute to performance. Results from $\text{DuaLK}_{\text{d-f}}$ demonstrate the importance of graph learning and attention for representation refinement, with lab tests further enhancing performance. The gain over PrimeKG_g highlights the value of a task-specific KG with relevant entities. Although $\text{GraphCare}_\text{h}$ achieves higher F1, its higher com-

[2]https://hcup-us.ahrq.gov/toolssoftware/chronic/chronic.jsp

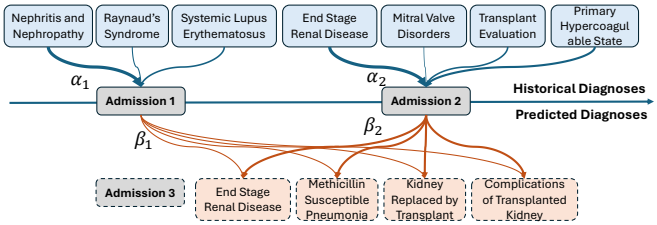


Fig. 4: Contribution of diagnoses to each admission, and admissions to predicted diagnoses. In this case, the patient has two historical admissions. We use two admissions to predict diagnoses in the third admission.

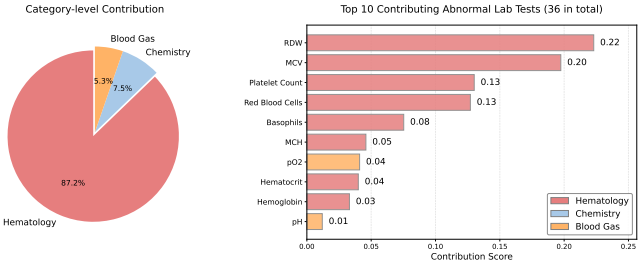


Fig. 5: Contribution of abnormal lab test results to the complications of transplanted kidney (the 4th disease of admission 3 in Fig. 4). Normalized important scores (gradients) are shown.

putational cost underscores the efficiency of leveraging multi-aspect information beyond fine-grained graphs. Moreover, the model *w/o lab* using modified GNN and pretraining tasks in Sherbet [5] still achieves superiority over baselines.

C. Model Interpretability

We interpret DuaLK by demonstrating the contributions of both historical diagnoses and abnormal lab test results introduced in Section III-D and III-E, respectively.

1) Quantification of Multilevel Attention: To measure the contributions both at the code level and at the admission level, we visualize the code-level attention distribution α in code-level attention, and visit-level coefficient β for a given patient on the diagnosis prediction task. Figure 4 demonstrates the historical diagnoses, the admissions, and the predicted diagnoses for a patient in rectangles. The contributions α and β are represented by blue and orange lines, and thicker and darker lines correspond to larger values. Note that, we only select important diagnoses with high attention scores.

In the first visit, DuaLK focus more on systemic lupus erythematosus (SLE) and nephritis. These disorders are highly related: SLE commonly causes immune–mediated nephritis, which can progress to chronic kidney damage. By the second visit, the patient's renal pathology has progressed to end-stage renal disease (ESRD), which receives the highest disease-level attention. DuaLK also places attention on a primary hypercoagulable state, which contributes to respiratory complications. Overall, predicted diagnoses, including complications of a transplanted kidney, recurrent ESRD-related conditions, and respiratory infections, closely match the patient's risk profile.

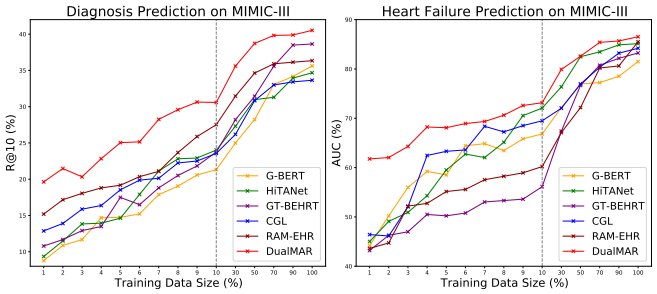


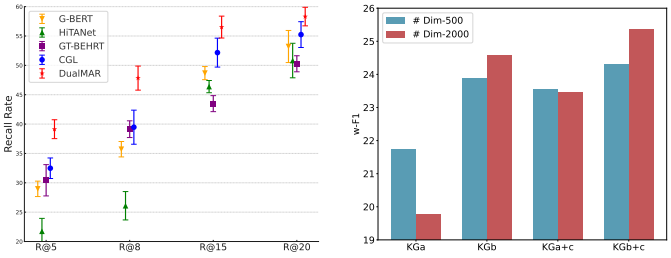
Fig. 6: Performance by EHR Training Data Sizes. Values on the x-axis indicate % of the entire training data. The dotted lines separate two ranges: [1, 10] and [10, 100] (%).

2) Interpretation of Abnormal Lab Tests: We get contributions of abnormal lab tests to diagnoses by two steps. Integrated gradients are adopted to measure the attribution from lab embeddings to diagnosis outputs, then project these attributions back to individual lab test space, enabling interpretation of which lab abnormality drives each disease. We select the complications of a transplanted kidney as the target condition. RDW, MCV, and platelet count contributed most prominently, indicating systemic inflammation, disrupted red cell production, and abnormalities in platelet activity. These patterns commonly occur in patients with chronic graft dysfunction, medication related marrow suppression, or early transplant rejection. The dominance of hematology in the attribution is therefore clinically reasonable and consistent with the pathophysiology of renal transplant complications.

D. Case Study

1) Effect of EHR Data Size: By incorporating the diagnosis KG as medical prior knowledge before the training phase, DuaLK is expected to exhibit greater robustness to variations in training data size. Hence, it is crucial to evaluate how well models can provide accurate diagnoses with limited labeled data, both with and without prior knowledge. Figure 6 presents the results for R@10 in Diagnosis Prediction and AUC in heart failure prediction. As shown, all models experience performance degradation as the amount of labeled data decreases, especially when less than 10% of the training set is available. However, DuaLK consistently outperforms other baselines, demonstrating that knowledge injection from the KG effectively enhances information utilization, even in scenarios with scarce labeled data.

2) Diagnosis for Low-Prevalence Diseases: We also design experiments focusing on predicting less frequent condition codes in MIMIC-III, similar to low-prevalence medical coding tasks. We select and predict 100 ICD codes that occur fewer than 10 times in the dataset. We follow the previous settings which excludes records without low-prevalence diseases to stabilize training. We use recall at 5/8/15/20 for evaluation. Figure 7a compares DuaLK with baselines using ontology or transformer architectures. DuaLK outperforms other selected methods on R@5, R@8, R@15, and R@20. Notably, CGL and G-BERT show significant improvements, indicating that



(a) R@k in Less-frequent Diagnosis Prediction.

(b) **w-**F_1 of Diagnosis Prediction with different KGs.

Fig. 7: Results for the 1^{st} and 2^{nd} case studies, full results are shown in Appendix -C.

TABLE VI: Accessibility Analysis with Adjusted GraphCare.

Models	GPU Memory Usage	Runnable
DuaLK - *batch sizes: 256, 32, and 32*		
- KG Embedding	11.73 GB	✓
- Pretrain & Finetune	9.78 GB	✓
- Direct Training	9.62 GB	✓
GraphCare + Diagnosis KG		
- Batch Size: 2	33.95 GB	×
- Batch Size: 4	42.57 GB	×
- Batch Size: 8	98.23 GB	×
- Batch Size: 16	172.04 GB	×

external knowledge bases can aid in diagnosing less-frequent conditions. This underscores the importance of incorporating medical ontologies or complex KGs to enhance model resilience against EHR data limitations.

3) Advanced Analysis of KG Embeddings: To assess the importance of each component in our Diagnosis KG, we divided it into three parts: ICD9-Hierarchy $\texttt{KG}_\text{a}$, Ontology-KG $\texttt{KG}_\text{b}$, and LLM-KG $\texttt{KG}_\text{c}$. We then evaluated the predictive performance of the model with these varying complexities. Figure 7b shows how the complexity and embedding dimensions of KGs influence diagnosis prediction. Key observations include: (1) LLM-generated triples ($\texttt{KG}_\text{c}$) enhance comprehensive KG representation, as seen in $\texttt{KG}_\text{a+c}$ and $\texttt{KG}_\text{b+c}$. (2) Ontology-KG ($\texttt{KG}_\text{b}$) from diverse data sources is more reliable for accurate predictions than ICD9-Hierarchy ($\texttt{KG}_\text{a}$). (3) Embedding dimensions significantly impact predictions; higher dimensions offer a deeper understanding but can lead to overfitting, especially with simpler KGs like $\texttt{KG}_\text{a}$.

4) Accessibility: We evaluated the accessibility of DuaLK by comparing with GraphCare [9] on an NVIDIA RTX-4090 GPU. For a fair comparison, we adapted `UMLS-GPT-KG` in GraphCare to our KG, adjusting only the batch size due to its end-to-end structure without pretraining module. Table VI shows average GPU usage for diagnosis prediction. GraphCare struggles with training on the GPU when transitioning from CCS KG to ICD-9 KG, limiting its scalability with large KGs. In contrast, DuaLK efficiently separates learning processes for KGs and EHR data, allowing it to scale to larger KGs and easily deploy across various EHR tasks and datasets.

TABLE VII: Different Embedding Baselines Comparison (%)

MODEL	MRR	HITS@1	HITS@3	HITS@10
TransE	54.77	45.90	58.30	72.40
RotatE	57.82	47.95	62.56	75.28
ModE	63.02	55.46	66.25	77.85
TripleRE	64.02	57.21	68.03	79.32
HyperE	59.03	53.69	64.82	74.93
HAKE	**67.37**	**60.23**	**70.62**	**81.21**

VI. Conclusion

We proposed DuaLK, a dual-expertise framework for enhancing clinical prediction by jointly leveraging structured medical knowledge and lab test information. Experimental results demonstrate that DuaLK consistently outperforms existing predictive baselines across multiple clinical tasks, which highlights the potential of combining structured knowledge and clinical evidence to advance EHR modeling. Overall, DuaLK not only broadens the scope of medical-domain knowledge but also addresses the limitations of traditional EHR models by incorporating more complex and nuanced medical data. Future work could explore expanding DuaLK to large-scale KGs or a broader range of clinical predictions beyond diagnoses. DuaLK also reveals the potentials of lab data in decision making, suggesting that leveraging lab-informed signals in predictive healthcare remains a promising direction.

Appendix

A. KG Embeddings Comparison

Most existing hierarchical KG embedding methods leverage Euclidean or non-Euclidean geometries to encode hierarchy. Euclidean-based approaches typically use coordinate projections to distinguish entity levels and retrieve hierarchical information more efficiently than hyperbolic or spherical methods, which often require additional steps or higher computational cost. To further improve efficiency, we decouple embedding training from graph learning by using link prediction tasks, allowing efficient GPU memory usage even on complex KGs. In our approach, each entity is projected onto polar coordinates to facilitate efficient embedding retrieval. As shown in Table VII, our polar-space embedding method, adapted from HAKE [43], outperforms five widely-used baselines [44]–[47] in capturing both hierarchical and semantic features, evaluated by MRR and HITS@k metrics (k = 1, 3, 10).

B. The Significance of Improvement

To evaluate the statistical significance of performance differences between our proposed DuaLK model and baseline models, we employed both the paired t-test and Wilcoxon signed-rank test. The t-test was chosen under the assumption of normally distributed performance metrics. To account for potential deviations from normality, the non-parametric Wilcoxon signed-rank test was also conducted. These tests allow us to robustly determine whether the observed improvements in DuaLK's performance are statistically significant compared to the baselines. Here we only take the w-F_1 of

TABLE VIII: Wilcoxon Signed-Rank Test and t-Test Results for DuaLK vs. Baseline Models with 10 runs. "p" means p-value for certain evaluation method, and "CI" means Confidence Interval calculated by T-testing method.

Models	Wilcoxon p	T-test p	95% CI
DuaLK vs. RETAIN	1.95e-03	1.62e-14	(7.41, 7.80)
DuaLK vs. Timeline	1.95e-03	2.01e-14	(7.17, 7.53)
DuaLK vs. GCT	1.95e-03	6.27e-10	(6.46, 7.39)
DuaLK vs. CGL	1.95e-03	9.77e-13	(5.93, 6.39)
DuaLK vs. SeqCare	1.95e-03	5.53e-13	(5.31, 5.67)
DuaLK vs. RAM-EHR	1.95e-03	1.23e-12	(4.87, 5.21)
DuaLK vs. G-BERT	1.95e-03	3.46e-11	(4.49, 4.88)
DuaLK vs. HiTANet	1.95e-03	7.65e-13	(5.62, 6.01)
DuaLK vs. Med-BERT	1.95e-03	1.99e-12	(5.89, 6.27)
DuaLK vs. GT-BEHRT	1.95e-03	4.57e-11	(3.98, 4.32)

TABLE IX: Less Frequent Codes Diagnosis Prediction.

	Rare-10			Rare-20		
Model	**R@5**	**R@8**	**R@15**	**R@5**	**R@8**	**R@15**
CGL	32.48	39.47	52.17	12.53	16.03	24.40
G-BERT	28.97	35.72	48.68	**14.95**	17.11	25.10
HiTANet	21.74	26.09	46.38	11.99	15.43	22.49
GT-BEHRT	30.43	39.13	43.48	12.95	17.56	23.79
DuaLK	**39.13**	**47.83**	**56.52**	14.28	**18.00**	**26.30**

Diagnosis Prediction in MIMIC-IV as an example to show the evaluation process on Table VIII. We observe that the Wilcoxon signed-rank test consistently yielded p-values of 1.95×10^{-3}, indicating significant differences between DuaLK and all baseline models. The paired t-tests further confirmed these results, with all p-values well below 1×10^{-10}. These results demonstrate the robustness of DuaLK's improvements in weighted F_1 scores over the baselines.

C. Details in Case Studies

We design experiments focusing on predicting less frequent condition codes in MIMIC-III, similar to rare ICD code prediction tasks. We conduct two subtasks:

- `Rare-20`: Predicting 200 ICD codes that occur fewer than 20 times in MIMIC-III.
- `Rare-10`: Predicting 100 ICD codes that occur fewer than 10 times in MIMIC-III.

Table IX shows the full results of both `Rare-10` and `Rare-20`. It highlights DuaLK's superior performance in predicting less frequent ICD codes across both subtasks, `Rare-10` and `Rare-20`. Specifically, DuaLK consistently outperforms baseline models, achieving the highest Recall at all thresholds (R@5, R@8, R@15). Notably, for `Rare-20`, DuaLK significantly surpasses other models, particularly in R@8 and R@15, reflecting its strong capability in handling rare condition codes. Table X shows the full results of DuaLK with varying KGs and embedding sizes. The results indicate that incorporating KG components ($\texttt{KG}_{\texttt{b+c}}$) consistently yields the highest Recall across R@10 and R@20, particularly with 2000-dimensional embeddings. Notably, $\texttt{KG}_{\texttt{b}}$ also shows strong performance, suggesting the significant contributions of fundamental Ontology-KG in enhancing diagnostic predictions. The comparison highlights the critical role of augmenting KG to achieve superior recall in diagnosis prediction tasks.

TABLE X: Diagnosis Prediction Results by Different KGs.

Models	# Dims	w-F_1	R@10	R@20
DuaLK-$\texttt{KG}_a$	500	$21.75_{(0.3)}$	$33.27_{(0.4)}$	$35.32_{(0.4)}$
DuaLK-$\texttt{KG}_a$	2000	$19.79_{(0.5)}$	$32.12_{(0.4)}$	$33.57_{(0.4)}$
DuaLK-$\texttt{KG}_b$	500	$23.87_{(0.3)}$	$37.73_{(0.4)}$	$39.18_{(0.4)}$
DuaLK-$\texttt{KG}_b$	2000	$24.58_{(0.3)}$	$39.27_{(0.4)}$	$40.32_{(0.4)}$
DuaLK-$\texttt{KG}_{a+c}$	500	$23.56_{(0.2)}$	$37.43_{(0.3)}$	$38.74_{(0.4)}$
DuaLK-$\texttt{KG}_{a+c}$	2000	$23.47_{(0.3)}$	$37.89_{(0.4)}$	$38.54_{(0.4)}$
DuaLK-$\texttt{KG}_{b+c}$	500	$24.32_{(0.3)}$	$39.73_{(0.4)}$	$40.15_{(0.4)}$
DuaLK-$\texttt{KG}_{b+c}$	2000	$\mathbf{25.37}_{\mathbf{(0.3)}}$	$\mathbf{40.52}_{\mathbf{(0.2)}}$	$\mathbf{41.86}_{\mathbf{(0.3)}}$

D. Prompting LLM

The following template in Table XI is used to instruct LLMs to convert textual information into structured triples.

ACKNOWLEDGMENT

This work is supported in part by the US National Science Foundation under grants 2047843 and 2437621. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] C. Lu, C. K. Reddy, P. Chakraborty, S. Kleinberg, and Y. Ning, "Collaborative graph learning with auxiliary text for temporal event prediction in healthcare," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021*, 2021, pp. 3529–3535.

[2] E. Choi, Z. Xu, Y. Li, M. Dusenberry, G. Flores, E. Xue, and A. M. Dai, "Learning the graphical structure of electronic health records with graph convolutional transformer," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, 2020, pp. 606–613.

[3] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis," *IEEE journal of biomedical and health informatics*, vol. 22, no. 5, pp. 1589–1604, 2017.

[4] J. Shang, T. Ma, C. Xiao, and J. Sun, "Pre-training of graph augmented transformers for medication recommendation," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019*, 2019, pp. 5953–5959.

[5] C. Lu, C. K. Reddy, and Y. Ning, "Self-supervised graph learning with hyperbolic embedding for temporal health event prediction," *IEEE Trans. Cybern.*, vol. 53, no. 4, pp. 2124–2136, 2023.

[6] R. Poulain and R. Beheshti, "Graph transformers on EHRs: Better representation improves downstream performance," in *The Twelfth International Conference on Learning Representations*, 2024.

[7] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "GRAM: graph-based attention model for healthcare representation learning," in *Proceedings of the 23rd ACM SIGKDD, 2017*, 2017, pp. 787–795.

[8] F. Ma, Q. You, H. Xiao, R. Chitta, J. Zhou, and J. Gao, "KAME: knowledge-based attention model for diagnosis prediction in healthcare," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018*, 2018, pp. 743–752.

[9] P. Jiang, C. Xiao, A. R. Cross, and J. Sun, "Graphcare: Enhancing healthcare predictions with personalized knowledge graphs," in *The Twelfth International Conference on Learning Representations*, 2023.

[10] W. H. Organization *et al.*, "International classification of diseases—ninth revision (icd-9)," *Weekly Epidemiological Record= Relevé épidémiologique hebdomadaire*, vol. 63, no. 45, pp. 343–344, 1988.

TABLE XI: General Prompting Framework for Triple Generation Tasks of LLMs

Name	**Prompt Template**
Variables	<category>: The prompt is not only suitable for disease but also available for concepts within medical domains such as medication and treatment. In this study, only condition concepts are considered. <term>: The concept name. In this study, it means disease names provided by web-scraped text. <topics>: The topics related to crawled text, and it has been provided when we crawled them from websites. <text>: The content of crawled text.
Skeleton	Given a crawled text about specific topic of certain <category>, please find triples related to the given <category> in terms of crawled text. • Filling triples in updates based on given information and strictly following output style of example updates. • Each update should follow the format of [ENTITY 1, RELATIONSHIP, ENTITY 2] with directed edge. • Both ENTITY 1 and ENTITY 2 should be noun, and one of them must be <term>. • Just output each unique triple once, don't output repeatedly. • It is possible that <category> name not exactly matched in crawled text (abbreviated or partly matched), consider it as the same thing. Example: ## An example demo is shown below... Given a paragraph about specific topic of certain <category>, please find triples related to the given <category> in the text. **Given Information:** <category> **Name**: <term> **Topics**: <topics> **Text**: <text> **Updates:** ## LLM's output is expected to generate here in terms of given information... ## [Head Entity, Relation, Tail Entity]...
Example	**Disease Name:** Heart Failure **Topics:** Overview **Text:** *Heart failure occurs when the heart muscle doesn't pump blood as well as it should. When this happens, blood often backs up and fluid can build up in the lungs, causing shortness of breath. Certain heart conditions gradually leave the heart too weak or stiff to fill and pump blood properly. These conditions include narrowed arteries in the heart and high blood pressure. Proper treatment may improve the symptoms of heart failure and may help some people live longer. Lifestyle changes can improve quality of life. Try to lose weight, exercise, use less salt and manage stress. But heart failure can be life-threatening. People with heart failure may have severe symptoms. Some may need a heart transplant or a device to help the heart pump blood. Heart failure is sometimes called congestive heart failure.* **Updates:** • [Heart Failure, IS_CAUSED_BY, Narrowed Arteries], • [Heart Failure, IS_CAUSED_BY, High Blood Pressure], • [Heart Failure, HAS_SYMPTOMS, Shortness of Breath], • [Heart Failure, HAS_SYMPTOMS, Fluid Build-up in Lungs], • [Heart Failure, NEEDS_TREATMENT, Proper Treatment], • [Heart Failure, NEEDS_TREATMENT, Lifestyle Changes] ## You can use more examples to refine output of LLMs, but only one example is also fine for this task.

[11] Y. Xu, X. Chu, K. Yang, Z. Wang, P. Zou, H. Ding, J. Zhao, Y. Wang, and B. Xie, "Seqcare: Sequential training with external medical knowledge graph for diagnosis prediction in healthcare data," in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 2819–2830.

[12] R. Xu, W. Shi, Y. Yu, Y. Zhuang, B. Jin, M. D. Wang, J. C. Ho, and C. Yang, "Ram-ehr: Retrieval augmentation meets clinical predictions on electronic health records," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

[13] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," *npj Digit. Medicine*, vol. 4, 2021.

[14] P. K. S. Prakash, S. Chilukuri, N. Ranade, and S. Viswanathan, "Rarebert: Transformer architecture for rare disease patient identification using administrative claims," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, 2021, pp. 453–460.

[15] H. O. Boll, A. Amirahmadi, M. M. Ghazani, W. O. de Morais, E. P. de Freitas, A. Soliman, F. Etminani, S. Byttner, and M. Recamonde-Mendoza, "Graph neural networks for clinical risk prediction based on electronic health records: A survey." *J. Biomed. Informatics*, vol. 151, p. 104616, 2024.

[16] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: benefits, risks, and strategies for success," *NPJ digital medicine*, vol. 3, no. 1, p. 17, 2020.

[17] E. Yang, P. Hu, X. Han, and Y. Ning, "Mplite: Multi-aspect pretraining for mining clinical health records," in *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2024, pp. 5096–5102.

[18] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. F. Stewart, "RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, 2016, pp. 3504–3512.

[19] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor AI: predicting clinical events via recurrent neural networks," in *Proceedings of the 1st Machine Learning in Health Care*, vol. 56, 2016, pp. 301–318.

[20] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao, "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, 2017, pp. 1903–1911.

[21] T. Bai, S. Zhang, B. L. Egleston, and S. Vucetic, "Interpretable representation learning for healthcare via capturing disease progression through time," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018*, 2018, pp. 43–51.

[22] L. Ma, C. Zhang, Y. Wang, W. Ruan, J. Wang, W. Tang, X. Ma, X. Gao, and J. Gao, "Concare: Personalized clinical feature embedding via capturing the healthcare context," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, 2020, pp. 833–840.

[23] P. Nguyen, T. Tran, N. Wickramasinghe, and S. Venkatesh, "`Deepr`: A convolutional net for medical records," *IEEE J. Biomed. Health Informatics*, vol. 21, no. 1, pp. 22–30, 2017.

[24] L. Ma, J. Gao, Y. Wang, C. Zhang, J. Wang, W. Ruan, W. Tang, X. Gao, and X. Ma, "Adacare: Explainable clinical health status representation learning via scale-adaptive feature extraction and recalibration," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, 2020, pp. 825–832.

[25] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," *NPJ digital medicine*, vol. 4, p. 86, 2021.

[26] E. Choi, C. Xiao, W. F. Stewart, and J. Sun, "Mime: Multilevel medical embedding of electronic health records for predictive healthcare," in *Neural Information Processing Systems 2018, NeurIPS 2018*, 2018, pp. 4552–4562.

[27] S. Park, S. Bae, J. Kim, T. Kim, and E. Choi, "Graph-text multi-modal pre-training for medical representation learning," in *Conference on Health, Inference, and Learning, CHIL 2022, 7-8 April 2022, Virtual Event*, vol. 174, 2022, pp. 261–281.

[28] X. Zou, W. He, Y. Huang, Y. Ouyang, Z. Zhang, Y. Wu, Y. Wu, L. Feng, S. Wu, M. Yang *et al.*, "Ai-driven diagnostic assistance in medical inquiry: Reinforcement learning algorithm development and validation," *Journal of Medical Internet Research*, vol. 26, p. e54616, 2024.

[29] B. Smith, A. Khojandi, and R. Vasudevan, "Bias in reinforcement learning: A review in healthcare applications," *ACM Computing Surveys*, vol. 56, no. 2, pp. 1–17, 2023.

[30] T. Nguyen, T. Huynh, M. H. Phan, Q. V. H. Nguyen, and P. Le Nguyen, "Carer-clinical reasoning-enhanced representation for temporal health risk prediction," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 10 392–10 407.

[31] P. Jiang, C. Xiao, M. Jiang, P. Bhatia, T. Kass-Hout, J. Sun, and J. Han, "Reasoning-enhanced healthcare predictions with knowledge graph community retrieval," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=8fLgt7PQza

[32] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[33] C. Knox, M. Wilson, C. M. Klinger, M. Franklin, E. Oler, A. Wilson, A. Pon, J. Cox, N. E. Chin, S. A. Strawbridge *et al.*, "Drugbank 6.0: the drugbank knowledgebase for 2024," *Nucleic acids research*, vol. 52, no. D1, pp. D1265–D1275, 2024.

[34] O. Ursu, J. Holmes, J. Knockel, C. G. Bologa, J. J. Yang, S. L. Mathias, S. J. Nelson, and T. I. Oprea, "Drugcentral: online drug compendium," *Nucleic acids research*, p. gkw993, 2016.

[35] M. A. Gargano, N. Matentzoglu, B. Coleman, E. B. Addo-Lartey, A. V. Anagnostopoulos, J. Anderton, P. Avillach, A. M. Bagley, E. Bakštein, J. P. Balhoff *et al.*, "The human phenotype ontology in 2024: phenotypes around the world," *Nucleic acids research*, vol. 52, no. D1, pp. D1333–D1346, 2024.

[36] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, "The sider database of drugs and side effects," *Nucleic acids research*, vol. 44, no. D1, pp. D1075–D1079, 2016.

[37] M. D. Brennan, K. M. Miner, and R. A. Rizza, "The mayo clinic," *The Journal of Clinical Endocrinology & Metabolism*, vol. 83, no. 10, pp. 3427–3434, 1998.

[38] S. S. Weinreich, R. Mangon, J. Sikkens, M. E. Teeuw, and M. Cornel, "Orphanet: a european database for rare diseases," *Nederlands tijdschrift voor geneeskunde*, vol. 152, no. 9, pp. 518–519, 2008.

[39] A. Schieppati, J.-I. Henter, E. Daina, and A. Aperia, "Why rare diseases are an important medical and social issue," *The Lancet*, vol. 371, no. 9629, pp. 2039–2041, 2008.

[40] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[41] G. Nahler and G. Nahler, "Anatomical therapeutic chemical classification system (atc)," *Dictionary of pharmaceutical medicine*, pp. 8–8, 2009.

[42] P. Chandak, K. Huang, and M. Zitnik, "Building a knowledge graph to enable precision medicine," *Scientific Data*, vol. 10, no. 1, p. 67, 2023.

[43] Z. Zhang, J. Cai, Y. Zhang, and J. Wang, "Learning hierarchy-aware knowledge graph embeddings for link prediction," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, 2020, pp. 3065–3072.

[44] A. Bordes, N. Usunier, A. García-Durán, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*, 2013, pp. 2787–2795.

[45] M. Nickel and D. Kiela, "Poincaré embeddings for learning hierarchical representations," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017, pp. 6338–6347.

[46] Z. Sun, Z. Deng, J. Nie, and J. Tang, "Rotate: Knowledge graph embedding by relational rotation in complex space," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

[47] V. Fionda and G. Pirrò, "Learning triple embeddings from knowledge graphs," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*. AAAI Press, 2020, pp. 3874–3881.

[48] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[49] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark, "Mimic-iv," *PhysioNet. Available online at: https://physionet.org/content/mimiciv/1.0/(accessed August 23, 2021)*, pp. 49–55, 2020.

[50] J. Luo, M. Ye, C. Xiao, and F. Ma, "Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records," in *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2020, pp. 647–656.

[51] OpenAI, "Embeddings," 2024. [Online]. Available: https://platform.openai.com/docs/guides/embeddings