MedCPI: A Construct–Personalize–Integrate Framework for KG-enhanced Clinical Prediction

Hang Wang **Hang Dong** **Lu Liu***
Department of Computer Science, University of Exeter, Exeter, UK
{hw876, H.Dong2, L.Liu3}@exeter.ac.uk

Abstract

Electronic health records (EHRs) provide longitudinal evidence for clinical prediction, but EHR data are sparse, incomplete, and heterogeneous, which can limit robustness. Medical knowledge graphs (MKGs) have therefore been incorporated to support KG-enhanced clinical prediction by linking heterogeneous EHR codes to shared medical concepts via structured relations. However, existing KG-enhanced approaches remain limited in two aspects: (i) task-specific knowledge selection when extracting knowledge from a large multi-source MKG; and (ii) patient-level personalization and knowledge integration, where personalization is often weakly controlled and knowledge integration is not sufficiently aligned with longitudinal patient trajectories. To address these issues, we propose MedCPI, a unified Construct–Personalize–Integrate framework. MedCPI first performs task-guided schema induction and KG normalization to build a task-specific Concept MKG as a denoised knowledge pool, then constructs controlled patient-level PKGs via local expansion and short path search, and finally integrates PKG representations with time-aware EHR representations via cross-attention for prediction. Experiments on MIMIC-III and MIMIC-IV across four clinical prediction tasks show consistent improvements over strong EHR-only and KG-enhanced baselines. Ablations and additional analyses further validate the contribution of each stage and illustrate how MedCPI utilizes structured medical knowledge.

1 Introduction

Electronic health records (EHRs) capture longitudinal patient trajectories through diagnoses, procedures, medications, and other clinical events, providing rich, fine-grained evidence for clinical prediction (Rajkomar et al., 2018; Tomašev et al., 2019; Landi et al., 2020; Li et al., 2022; Kraljevic et al., 2024). Building on this foundation, recent deep learning models have leveraged EHR data to learn patient representations and demonstrate promising performance across a variety of predictive tasks (Choi et al., 2016; Ma et al., 2017; Choi et al., 2018, 2020; Ma et al., 2020; Luo et al., 2020). However, EHR data are inherently sparse, incomplete, and heterogeneous, leading to challenges in model robustness and reliability (Xiao et al., 2018; Amirahmadi et al., 2023; Li et al., 2023; Yu et al., 2024). To mitigate these issues, Medical Knowledge Graphs (MKGs) have been incorporated into clinical prediction models. By organizing medical entities and their relations into a structured graph, MKGs link heterogeneous EHR codes to shared medical concepts and provide prior knowledge that complements patient representations and smooths over data sparsity, thereby enabling KG-enhanced clinical prediction (Shang et al., 2019a; Mao et al., 2022; Xu et al., 2023).

Building on MKGs, KG-enhanced clinical prediction has evolved through several stages. Early work exploits hierarchical structures in medical ontologies, using parent-child or ancestor relations to enrich the representations of sparse medical codes (Choi et al., 2017; Ma et al., 2018; Zhang et al., 2020). More recent approaches move beyond pure hierarchies by leveraging multi-relational MKGs that integrate diverse biomedical sources and are encoded with graph neural networks to provide global relational context for clinical prediction tasks (Shang et al., 2019b; Ge et al., 2024). The latest line of research constructs personalized knowledge graphs (PKGs) for each patient by expanding from their EHR codes into the MKG and performing graph reasoning over the resulting PKGs, thereby yielding representations that explicitly capture individual clinical histories and their associated medical context (Yang et al., 2023b,a; Jiang et al., 2024).

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 24266–24282

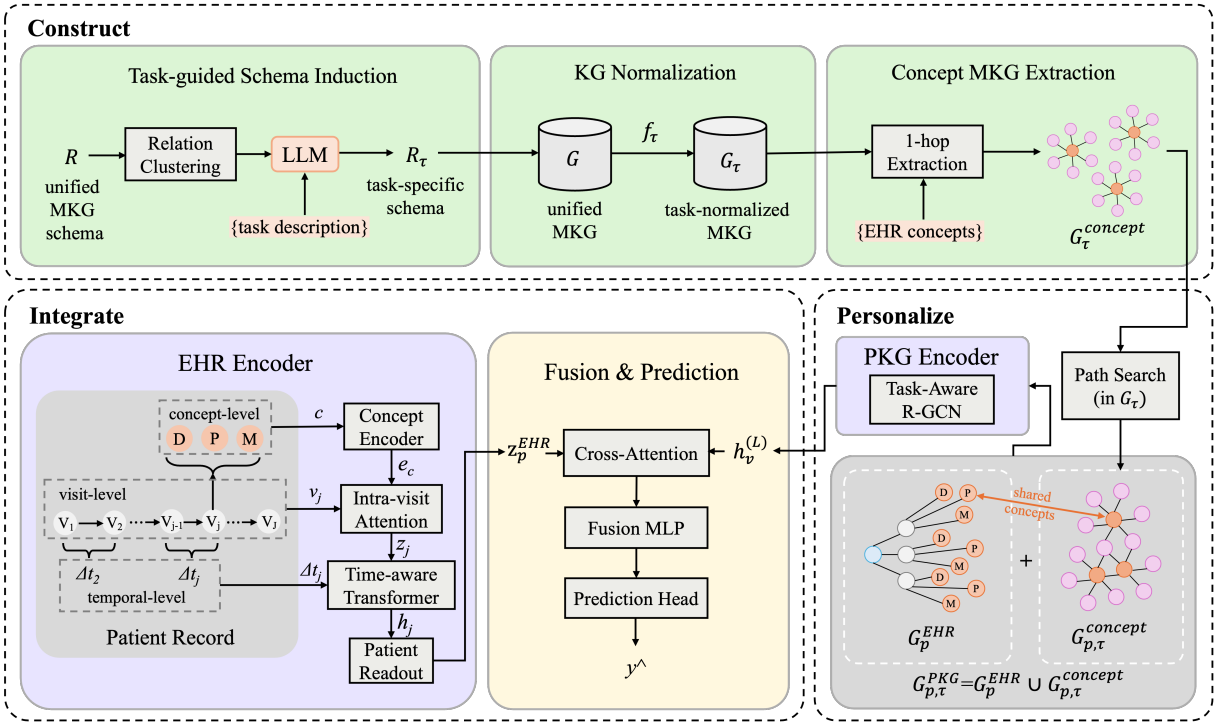


Figure 1: **MedCPI**: a Construct–Personalize–Integrate framework for KG-enhanced clinical prediction. **Construct** (§4.1) performs task-guided schema induction and KG normalization on the unified MKG, followed by global 1-hop extraction from EHR concepts to obtain a task-specific Concept MKG. **Personalize** (§4.2) constructs a patient-level PKG by augmenting the EHR subgraph with task-relevant knowledge and short paths in the task-normalized MKG, and encodes it with a task-aware R-GCN. **Integrate** (§4.3) encodes longitudinal EHRs and fuses them with PKG node representations via cross-attention for final prediction.

Despite this progress, existing KG-enhanced approaches for clinical prediction still face two key limitations in how medical knowledge is constructed and used: **(a) Knowledge extraction from a large, multi-source MKG**. First, the large, multi-source MKG integrates heterogeneous resources with varying relation granularity, so its schema often contains redundancy and semantic overlap, introducing substantial schema-level noise. Second, existing methods typically lack explicit task-specific knowledge selection and thus retain many task-irrelevant relations when extracting knowledge for prediction. Here, a task refers to a concrete clinical prediction problem (e.g., short-term outcomes such as in-hospital mortality and 30-day readmission, or disease onset prediction). For example, for type 2 diabetes mellitus (T2DM) onset prediction, risk-factor and diagnostic relations (e.g., `has_risk_factor`, `diagnosed_by`) are often the most informative. Without task-specific selection, neighborhood expansion can over-retrieve heterogeneous, weakly related nodes and relations, diluting task-relevant evidence and impairing predictive performance. **(b) Patient-level personalization and knowledge integration**. Existing approaches often construct PKGs through unconstrained neighborhood expansion or retrieval-based assembly, so the resulting graph structure depends heavily on the patient's observed concepts and retrieved neighborhoods or paths, which may introduce weakly related nodes and relations. In addition, knowledge integration is often weakly aligned with the patient's longitudinal clinical context, making it difficult to incorporate structured knowledge in a way that reflects the evolving patient trajectory across visits.

To address these limitations, we propose **MedCPI**, a unified Construct–Personalize–Integrate framework for KG-enhanced clinical prediction (Figure 1). In response to limitation (a), *Construct* performs task-guided schema induction and KG normalization to produce a task-specific Concept MKG. In contrast to prior methods that directly extract or encode knowledge from a large multi-source MKG with heterogeneous relation semantics, this stage yields a compact and denoised knowledge pool with explicitly task-relevant relations. In response to limitation (b), *Personalize* constructs patient-level PKGs from the task-specific Concept MKG and the patient's EHR via

constrained local expansion and short path instantiation, then *Integrate* performs cross-attention over PKG nodes using time-aware EHR representations.

The main contributions are summarized as follows:

- We propose MedCPI, a unified Construct–Personalize–Integrate framework that organizes task-guided knowledge construction, patient-level personalization, and EHR–PKG integration within a structured pipeline, thereby providing a systematic paradigm for knowledge construction and utilization in KG-enhanced clinical prediction.
- We introduce a task-guided knowledge selection mechanism that derives a task-specific Concept MKG from large multi-source MKGs through schema induction and KG normalization, which reduces schema noise caused by heterogeneous relations and provides a semantically consistent and task-specific knowledge pool for clinical prediction.
- Experiments on MIMIC-III and MIMIC-IV across multiple clinical prediction tasks demonstrate the effectiveness of MedCPI, while ablations and analyses validate the roles of Construct, Personalize, and Integrate.

2 Related Work

In this section, we introduce related work on clinical prediction and group existing methods into two families: EHR-only models and KG-enhanced prediction models.

2.1 EHR-only models

EHR-only models rely solely on structured EHR data without external medical knowledge. Some adopt general sequence models such as GRU (Chung et al., 2014) and Transformer (Vaswani et al., 2017), while others design architectures tailored to EHR structure. Representative works include RETAIN (Choi et al., 2016), which uses reverse-time and dual-level attention for interpretability; Dipole (Ma et al., 2017), which combines bidirectional RNNs with attention for inter-visit dependencies; MiME (Choi et al., 2018), which models the multi-layer patient-visit-diagnosis-treatment structure to alleviate sparsity; and HiTANet (Luo et al., 2020), which introduces hierarchical time-aware attention for irregular intervals.

2.2 KG-enhanced models

KG-enhanced models incorporate MKGs to complement EHR-based representations, including static KG-enhanced models and PKG reasoning models. **Static KG-enhanced models** integrate external knowledge during EHR encoding without constructing patient-specific PKGs. GRAM (Choi et al., 2017) propagates ontology hierarchies to smooth sparse diagnosis representations, and KAME (Ma et al., 2018) uses ancestor information with knowledge-level attention to support prediction. Moving beyond pure hierarchies, GAMENet (Shang et al., 2019b) leverages a multi-relational MKG to provide richer relational context, and DKEC (Ge et al., 2024) encodes a multi-relational MKG with graph neural networks to model global relations for clinical prediction. **PKG reasoning models** construct a PKG for each patient by expanding from observed EHR concepts into an MKG and perform graph reasoning over the resulting PKGs. KGxDP (Yang et al., 2023b) applies type-specific graph attention to enable interpretable prediction, KerPrint (Yang et al., 2023a) introduces hierarchical attention over local and global graphs for patient representation learning, and GraphCare (Jiang et al., 2024) combines LLM-extracted and KG-retrieved knowledge in a PKG-based reasoning framework to enhance clinical prediction.

Despite these advances, most KG-enhanced models still construct and use knowledge directly from MKGs with heterogeneous relation semantics, and lack explicit task-specific knowledge selection, motivating the unified, task-guided design of MedCPI.

3 Preliminaries

EHR Data and Patient Subgraph. Let $\mathcal{P}$ be the set of patients. Each patient $p \in \mathcal{P}$ has a time-ordered EHR sequence $\mathbf{E}_p = [V_{p,1}, \ldots, V_{p,L_p}]$ of visits, where $V_{p,t} = (\mathcal{D}_{p,t}, \mathcal{P}_{p,t}, \mathcal{M}_{p,t})$ contains diagnosis, procedure, and medication codes recorded at visit t. We represent the EHR of patient p as a patient subgraph $G_p^{\text{ehr}} = (V_p^{\text{ehr}}, R^{\text{ehr}}, \mathcal{E}_p^{\text{ehr}})$, where V_p^{ehr} contains the patient node p, visit nodes, and the medical concept nodes appearing in $\mathbf{E}_p$. We map EHR codes to UMLS concepts using SapBERT (Liu et al., 2021) and treat the

aligned concepts as the medical concept nodes in V_p^{ehr}. $\mathcal{E}_p^{\text{ehr}}$ links p to each visit and links each visit to its aligned concepts via relation types R^{ehr} (e.g., `has_admission`, `has_diagnosis`, `has_procedure`, `has_medication`).

Unified Medical Knowledge Graph. Given a unified MKG $G = (V, R, \mathcal{E})$ integrated from multiple sources, the set of medical concepts is denoted by V, the set of relation types by R, and $\mathcal{E} \subseteq V \times R \times V$ is the set of triples (h, r, t). Due to heterogeneous sources and fine-grained vocabularies, G contains noisy and semantically overlapping relations, motivating the *Construct* stage in Section 4.1.

Problem Setting. Let $\mathcal{T}$ be the set of clinical prediction tasks, and let T_τ denote the description of task τ. For each task $\tau \in \mathcal{T}$, let $\mathcal{I}_\tau$ denote the set of prediction instances, where an instance $i \in \mathcal{I}_\tau$ corresponds to either a patient or a specific visit, depending on the task definition. For each instance i associated with patient p, the goal is to predict an outcome $\hat{y}_i$ from the patient's EHR history $\mathbf{E}_p$, the unified MKG G, and the task description T_τ. Formally, we learn a predictor

$$F : \big(\mathbf{E}_p,\, G,\, T_\tau\big) \rightarrow \hat{y}_i, \tag{1}$$

where MedCPI implements F via the three-stage framework described in Section 4.

4 Method

In this section, we present **MedCPI**, which is illustrated in Figure 1. The framework consists of three stages: (i) *Construct*, which builds a task-specific Concept MKG from the unified MKG; (ii) *Personalize*, which constructs and encodes the PKG for each patient by combining the patient EHR subgraph with the Concept MKG; and (iii) *Integrate*, which fuses the PKG and EHR representations to obtain the final prediction.

4.1 Construct: Task-Specific Concept MKG

The *Construct* stage presents a pipeline to construct a task-specific Concept MKG G_τ^{concept} from the unified MKG G. The resulting G_τ^{concept} serves as a shared, task-relevant knowledge pool for the subsequent *Personalize* stage.

Task-Guided Schema Induction. A task-guided relation schema is induced conditioned on the task description T_τ. Relations in the unified MKG are first grouped into semantic clusters shared across tasks. For a given task τ, each cluster is evaluated with respect to the task semantics by an instruction-tuned LLM, which determines relevance and assigns a canonical relation name to each retained cluster; irrelevant clusters are discarded. After cross-cluster inconsistencies are resolved, the induced schema is represented by a set of canonical relation types R_τ and the task-specific mapping

$$f_\tau : R \rightarrow R_\tau \cup \{\text{IGNORE}\}, \tag{2}$$

which maps each original relation type to a canonical relation in R_τ, and to IGNORE otherwise. The algorithm and prompt template are provided in Appendix B.2.

KG Normalization. Using the induced schema R_τ, the unified MKG $G = (V, R, \mathcal{E})$ is normalized to obtain a task-normalized MKG G_τ. Each triple $(h, r, t) \in \mathcal{E}$ is processed by applying f_τ to its relation type: triples mapped to IGNORE are removed, while retained triples are rewritten using their canonical relations. This results in the task-normalized triple set

$$\mathcal{E}_\tau = \{(h, f_\tau(r), t) \mid (h, r, t) \in \mathcal{E}, \; f_\tau(r) \neq \text{IGNORE}\}. \tag{3}$$

and the corresponding task-normalized MKG $G_\tau = (V, R_\tau, \mathcal{E}_\tau)$.

Concept MKG Extraction (Global 1-Hop). Based on G_τ, the task-specific Concept MKG is constructed by extracting a concept-centric subgraph around EHR concepts. For each task, patients are first split into training, validation, and test sets, and the Concept MKG is constructed using only concepts observed in the training split. A global 1-hop expansion is then performed from these training EHR concepts over G_τ, and only triples whose endpoints are of clinically meaningful types S_{med} (defined in Appendix B.3) are retained. The resulting subgraph is denoted as

$$G_\tau^{\text{concept}} = (V_\tau^{\text{concept}}, R_\tau, \mathcal{E}_\tau^{\text{concept}}). \tag{4}$$

4.2 Personalize: PKG Construction and Encoding

The *Personalize* stage constructs a PKG for each patient from their EHR and the task-specific Concept MKG, and encodes it into a representation for the subsequent *Integrate* stage.

PKG Construction. For each patient p, the set of medical concepts C_p appearing in the EHR is identified. A PKG for patient p and task τ is constructed by augmenting the patient EHR subgraph (defined in Section 3) with task-relevant knowledge from G_τ^{concept} and performing path search in G_τ. Specifically, we attach the 1-hop neighborhood of each concept $c \in C_p$ from G_τ^{concept}, and further insert short knowledge paths between concept pairs that co-occur in the patient's EHR. We select the maximum path length on the validation set and use 2-hop paths in all experiments. To obtain a compact and controllable PKG, constraints are applied to path selection and graph size (details are provided in Appendix B.3). The resulting PKG is denoted as

$$G_{p,\tau}^{\text{pkg}} = (V_{p,\tau}^{\text{pkg}},\, R_\tau \cup R^{\text{ehr}},\, \mathcal{E}_{p,\tau}^{\text{pkg}}), \tag{5}$$

which captures both patient-specific clinical events and task-specific medical knowledge.

PKG Encoder. The PKG is encoded using a multi-layer relational graph convolutional network (R-GCN) (Schlichtkrull et al., 2018). Each node $v \in V_{p,\tau}^{\text{pkg}}$ is initialized with an embedding $\mathbf{h}_v^{(0)}$, and message passing is performed over the relations in the PKG. At layer $l{+}1$, node representations are updated by

$$\mathbf{h}_v^{(l+1)} = \sigma\left(\sum_r \sum_{u \in \mathcal{N}_r(v)} \frac{1}{c_{v,r}} \mathbf{W}_r^{(l)} \mathbf{h}_u^{(l)} + \mathbf{W}_0^{(l)} \mathbf{h}_v^{(l)}\right), \tag{6}$$

where $\mathcal{N}_r(v)$ denotes the neighbors of v under relation r. After L layers, the resulting node representations $\{\mathbf{h}_v^{(L)} \mid v \in V_{p,\tau}^{\text{pkg}}\}$ are passed to the EHR–PKG fusion module described in Section 4.3.

4.3 Integrate: EHR–PKG Fusion and Prediction

Given the EHR sequence and the PKG node representations for each patient and task, the Integrate stage encodes the longitudinal EHR into time-aware representations and fuses them with PKG node representations via cross-attention for prediction.

EHR Encoder. The EHR sequence $\mathbf{E}_p$ (defined in Section 3) is encoded using a Hierarchical Time-aware Transformer (HTT), inspired by the hierarchical and time-aware attention mechanism in HiTANet (Luo et al., 2020), to produce time-aware visit representations. HTT performs intra-visit concept aggregation followed by a time-aware Transformer over visits. Within each visit, concepts are aggregated to form a visit embedding, and visit embeddings are then processed by the time-aware Transformer that incorporates visit order and time intervals. Let $\{\mathbf{h}_{p,1}, \ldots, \mathbf{h}_{p,L_p}\}$ denote the final hidden states for all visits. These representations are time-aware in the sense that temporal order and inter-visit intervals are encoded in $\mathbf{h}_{p,t}$.

For patient-level tasks, attention pooling over visits is applied to obtain a single patient-level EHR representation:

$$\mathbf{z}_p^{\text{ehr}} = \sum_{t=1}^{L_p} \alpha_{p,t} \mathbf{h}_{p,t}, \tag{7}$$

where $\alpha_{p,t}$ are normalized attention weights. For visit-level tasks, each visit constitutes a prediction instance, and the corresponding representation $\mathbf{h}_{p,t}$ is used as the EHR context for that instance, without global pooling.

EHR–PKG Fusion. Given an EHR context representation and the PKG node representations $\{\mathbf{h}_v^{(L)}\}_{v \in V_{p,\tau}^{\text{pkg}}}$ from Section 4.2, a cross-attention mechanism is applied that uses the EHR context to attend over PKG nodes, integrating longitudinal clinical information with structured medical knowledge. The time-aware property of this integration arises from the query representation, which encodes temporal information via HTT.

Both task types are described under a unified prediction-instance formulation. Let $\mathcal{I}_\tau$ denote the set of prediction instances for task τ. For patient-level tasks, each instance corresponds to a patient and its EHR context representation is $\mathbf{q}_i = \mathbf{z}_p^{\text{ehr}}$. For visit-level tasks, each instance corresponds to a visit indexed by (p, t), and its EHR context representation is $\mathbf{q}_i = \mathbf{h}_{p,t}$. The PKG is constructed at the patient level for each task τ. For visit-level tasks, when constructing the PKG for an instance (p, t), only concepts observed up to visit t are used, ensuring no future information is introduced. The resulting PKG is shared across visits of the same patient up to the corresponding prediction time.

The cross-attended knowledge summary for instance i is computed as

$$\mathbf{c}_i = \text{CrossAttn}\big(\mathbf{q}_i,\, \{\mathbf{h}_v^{(L)}\}_{v \in V_{p,\tau}^{\text{pkg}}}\big). \tag{8}$$

For visit-level tasks, cross-attention weights are indexed by prediction instances (p, t), while for patient-level tasks they are indexed by patient p.

The fused representation is then obtained by

$$\mathbf{z}_i^{\text{fuse}} = \phi\big([\mathbf{q}_i; \mathbf{c}_i]\big), \tag{9}$$

where $\phi(\cdot)$ is a two-layer MLP.

Prediction. For binary tasks, the predicted probability for instance i is given by

$$\hat{y}_i = \sigma\big(\mathbf{w}_\tau^\top \mathbf{z}_i^{\text{fuse}} + b_\tau\big), \tag{10}$$

where $(\mathbf{w}_\tau, b_\tau)$ are task-specific learnable parameters shared across instances.

The model is trained by minimizing the task loss over prediction instances:

$$\mathcal{L}_\tau = \sum_{i \in \mathcal{I}_\tau} \ell_\tau\big(y_i, \hat{y}_i\big), \tag{11}$$

where $\ell_\tau(\cdot, \cdot)$ denotes the binary cross-entropy loss (Appendix B.4.2).

5 Experiments

In this section, we empirically evaluate **MedCPI** on multiple clinical prediction tasks to answer the following research questions:

- **RQ1**: How does MedCPI perform compared with representative baselines across different datasets and prediction tasks?
- **RQ2**: How do the three stages of MedCPI (*Construct*, *Personalize*, *Integrate*) contribute to predictive performance?
- **RQ3**: How do key design choices and model hyperparameters influence the performance and behavior of MedCPI?

5.1 Experimental Setup

Prediction Tasks. We evaluate MedCPI on four clinical prediction tasks, grouped into general-purpose and disease-specific categories. The general-purpose tasks include *in-hospital mortality* and *30-day readmission*, which are defined at the visit level: each visit forms a prediction instance, using the patient's prior visits as input. The disease-specific tasks include *type 2 diabetes mellitus (T2DM) onset* and *coronary artery disease (CAD) onset*, which are defined at the patient level: each patient contributes a single instance, labeled by whether the disease occurs for the first time in the future. Detailed definitions are provided in Appendix A.2.

Statistic	MIMIC-III	MIMIC-IV
#Patients	7,537	100,163
#Visits	19,993	422,739
Avg. diagnoses per visit	13.02	12.26
Avg. procedures per visit	3.85	1.45
Avg. medications per visit	27.98	23.99
Avg. visits per patient	2.65	4.22
Avg. length of stay (days)	10.72	4.89

Table 1: Summary statistics of the MIMIC-III and MIMIC-IV datasets (patients with at least two visits).

Data and Evaluation Metrics. We conduct experiments on two public EHR datasets, MIMIC-III (Johnson et al., 2016) and MIMIC-IV (Johnson et al., 2023). We retain patients with at least two hospital visits to ensure sufficient longitudinal histories. After filtering, MIMIC-III contains 7,537 patients and 19,993 visits, and MIMIC-IV contains 100,163 patients and 422,739 visits. Table 1 summarizes the processed cohorts and visit-level summary statistics. For all tasks, patients are split into training, validation, and test sets with an 8:1:1 ratio to avoid patient overlap. We evaluate binary prediction performance using the area under the ROC curve (AUROC) and the area under the precision–recall curve (AUPRC).

Baselines. We compare MedCPI against three categories of baseline methods:

- **EHR-only models.** These methods rely solely on structured EHR data without incorporating external medical knowledge. We include general sequence models such as GRU (Cho et al., 2014) and Transformer (Vaswani et al., 2017), as well as EHR-oriented models such as RETAIN (Choi et al., 2016) and HiTANet (Luo et al., 2020).
- **Static KG-enhanced models.** These methods integrate external MKGs to enrich EHR representations during the encoding stage. Representative models include GRAM (Choi et al., 2017) and KAME (Ma et al., 2018).
- **PKG reasoning models.** These methods construct PKGs from EHR and external knowledge sources and perform graph reasoning for clinical prediction. We consider KerPrint (Yang et al., 2023a), KGxDP (Yang et al., 2023b), and GraphCare (Jiang et al., 2024) as representative approaches.

Model	In-hospital Mortality				30-day Readmission			
	MIMIC-III		MIMIC-IV		MIMIC-III		MIMIC-IV	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
EHR-only Models								
GRU (Cho et al., 2014)	0.628(0.014)	0.177(0.009)	0.772(0.009)	0.072(0.006)	0.651(0.007)	0.460(0.012)	0.666(0.004)	0.503(0.004)
RETAIN (Choi et al., 2016)	0.636(0.013)	0.183(0.005)	0.758(0.005)	0.078(0.025)	0.653(0.023)	0.462(0.012)	0.671(0.003)	0.504(0.003)
Transformer (Vaswani et al., 2017)	0.640(0.011)	0.178(0.010)	0.766(0.006)	0.072(0.009)	0.661(0.005)	0.463(0.008)	0.668(0.003)	0.505(0.005)
HiTANet (Luo et al., 2020)	0.647(0.010)	0.202(0.009)	0.785(0.011)	0.105(0.009)	0.665(0.021)	0.464(0.016)	0.681(0.001)	0.515(0.001)
Static KG-enhanced Models								
GRAM (Choi et al., 2017)	0.648(0.011)	0.201(0.006)	0.774(0.009)	0.094(0.007)	0.665(0.012)	0.465(0.011)	0.673(0.005)	0.512(0.001)
KAME (Ma et al., 2018)	0.652(0.007)	0.203(0.004)	0.782(0.008)	0.098(0.013)	0.680(0.007)	0.470(0.014)	0.679(0.002)	0.520(0.009)
PKG Reasoning Models								
KerPrint (Yang et al., 2023a)	0.686(0.010)	0.217(0.011)	0.794(0.008)	0.153(0.013)	0.701(0.006)	0.483(0.003)	0.720(0.008)	0.572(0.007)
KGxDP (Yang et al., 2023b)	0.672(0.009)	0.214(0.004)	0.796(0.003)	0.141(0.009)	0.709(0.003)	0.486(0.012)	0.728(0.001)	0.565(0.009)
GraphCare (Jiang et al., 2024)	0.691(0.003)	0.224(0.019)	0.797(0.006)	0.142(0.002)	0.716(0.005)	0.487(0.011)	0.736(0.004)	0.585(0.008)
MedCPI (Ours)	**0.713**(0.008)	**0.242**(0.018)	**0.809**(0.006)	**0.182**(0.012)	**0.732**(0.005)	**0.498**(0.010)	**0.744**(0.004)	**0.606**(0.007)

Table 2: Performance on two general-purpose clinical prediction tasks, in-hospital mortality (left block) and 30-day readmission (right block), on MIMIC-III and MIMIC-IV, reported in AUROC and AUPRC. Results are reported as mean (standard deviation) over 5 random seeds.

Model	T2DM Onset				CAD Onset			
	MIMIC-III		MIMIC-IV		MIMIC-III		MIMIC-IV	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
EHR-only Models								
GRU (Cho et al., 2014)	0.735(0.024)	0.296(0.043)	0.646(0.012)	0.166(0.016)	0.581(0.044)	0.292(0.042)	0.750(0.008)	0.328(0.015)
RETAIN (Choi et al., 2016)	0.750(0.033)	0.301(0.012)	0.671(0.001)	0.191(0.003)	0.598(0.012)	0.351(0.021)	0.743(0.002)	0.341(0.005)
Transformer (Vaswani et al., 2017)	0.751(0.010)	0.327(0.032)	0.670(0.006)	0.192(0.018)	0.650(0.037)	0.347(0.054)	0.746(0.015)	0.336(0.023)
HiTANet (Luo et al., 2020)	0.753(0.024)	0.380(0.035)	0.673(0.012)	0.194(0.017)	0.661(0.026)	0.355(0.023)	0.760(0.006)	0.356(0.002)
Static KG-enhanced Models								
GRAM (Choi et al., 2017)	0.746(0.014)	0.309(0.018)	0.657(0.012)	0.192(0.008)	0.618(0.009)	0.314(0.016)	0.755(0.007)	0.337(0.006)
KAME (Ma et al., 2018)	0.751(0.023)	0.312(0.009)	0.668(0.001)	0.197(0.003)	0.658(0.004)	0.352(0.009)	0.747(0.001)	0.352(0.005)
PKG Reasoning Models								
KerPrint (Yang et al., 2023a)	0.757(0.012)	0.418(0.014)	0.719(0.011)	0.231(0.023)	0.665(0.014)	0.365(0.018)	0.763(0.011)	0.391(0.014)
KGxDP (Yang et al., 2023b)	0.772(0.003)	0.411(0.008)	0.696(0.006)	0.223(0.003)	0.689(0.003)	0.372(0.008)	0.775(0.001)	0.395(0.004)
GraphCare (Jiang et al., 2024)	0.754(0.014)	0.413(0.007)	0.711(0.001)	0.286(0.003)	0.695(0.005)	0.382(0.007)	0.772(0.004)	0.401(0.004)
MedCPI (Ours)	**0.798**(0.002)	**0.425**(0.006)	**0.763**(0.001)	**0.313**(0.002)	**0.705**(0.003)	**0.405**(0.006)	**0.797**(0.001)	**0.418**(0.003)

Table 3: Performance on two disease prediction tasks, T2DM (left block) and CAD (right block), on MIMIC-III and MIMIC-IV, reported in AUROC and AUPRC. Results are reported as mean (standard deviation) over 5 random seeds.

Implementation Details. UMLS (Bodenreider, 2004) is used as the external medical knowledge source. Diagnosis, procedure, and medication codes are aligned to UMLS concepts using SapBERT (Liu et al., 2021). Task-guided schema induction is performed offline using GPT-5 (OpenAI, 2025), with relation texts embedded by SapBERT and clustered by agglomerative clustering. Hyperparameters are tuned on the validation set for each dataset and task, and results are reported as mean±standard deviation over five random seeds. Experiments are conducted on NVIDIA GH200 GPUs. The project repository is available at `https://github.com/hw146/MedCPI`. Additional implementation details are provided in the Appendix.

5.2 Overall Results (RQ1)

Tables 2 and 3 report the overall performance on four clinical prediction tasks across MIMIC-III and MIMIC-IV. Overall, MedCPI achieves the best results in all datasets and tasks, demonstrating consistent gains across both general-purpose outcomes and disease onset prediction.

The results suggest three key observations. First, temporal modeling plays an important role: a time-aware encoder such as HiTANet can match or outperform static knowledge-enhanced baselines in several settings, showing that strong longitudinal sequence modeling can effectively improve performance. Second, static KG-enhancement provides limited improvements, and the gains vary across tasks and datasets. Third, PKG reasoning methods are generally more robust than static enhancement, suggesting that medical knowledge is more effective when it is selected and used based on the patient context, rather than being added in a fixed way.

Building on these insights, MedCPI integrates both complementary aspects: it constructs task-relevant, patient-specific knowledge and performs time-aware integration along the evolving patient trajectory. This joint design enables consistent im-

provements over both advanced EHR-only models and prior PKG reasoning approaches, supporting MedCPI as an effective framework for clinical prediction.

5.3 Ablation Studies (RQ2)

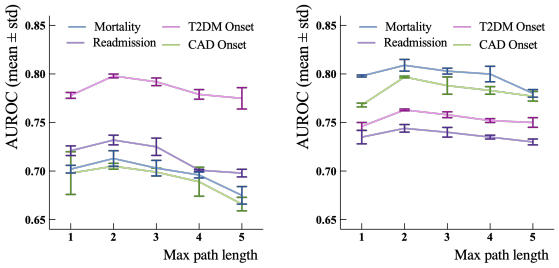
We conduct ablation studies to examine the role of each stage in MedCPI. Specifically, we consider the following variants: (i) *w/o Construct*, which removes task-guided schema induction and Concept MKG construction; (ii) *w/o Personalize*, which disables patient-specific PKG construction and uses only the shared Concept MKG; (iii) *w/o Integrate (PKG-only)*, which makes predictions using PKG representations only; and (iv) *w/o Integrate (EHR-only)*, which relies solely on longitudinal EHR representations. Results on in-hospital mortality and T2DM onset are shown in Table 4; full results are in Appendix C.1.

Removing *Construct* leads to consistent performance drops across datasets and tasks, suggesting that task-guided schema induction and Concept MKG construction help organize and filter task-relevant medical knowledge. Removing *Personalize* also degrades performance, with more noticeable drops on T2DM onset, suggesting that patient-specific knowledge selection provides additional benefits beyond shared knowledge. For the *Integrate* stage, both PKG-only and EHR-only variants underperform the full model. This shows that neither structured knowledge nor longitudinal EHR modeling alone is sufficient. Overall, the full MedCPI achieves the best results, demonstrating that the three stages contribute complementary information.

5.4 Additional Analysis (RQ3)

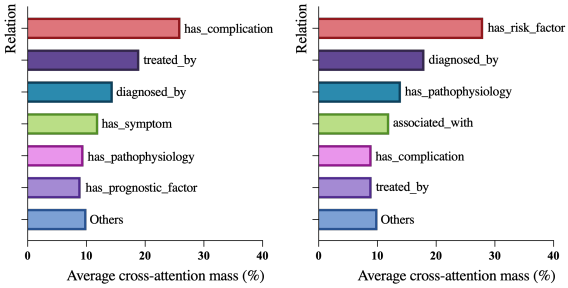
We provide additional analyses to examine key design choices and to better understand how MedCPI leverages structured medical knowledge. We study (i) sensitivity to the maximum path length in PKG construction and (ii) cross-attention distributions over relation types for interpretability.

Effect of Maximum Path Length. Figure 2 reports performance under different maximum path lengths used in PKG construction. Across both MIMIC-III and MIMIC-IV, performance improves when increasing the maximum path length from 1 to 2, and then gradually declines as longer paths are allowed. This trend is consistent across tasks, indicating that short knowledge paths are generally sufficient for effective knowledge enrichment. Based on this observation, we set the maximum path length to 2 in all experiments.



(a) MIMIC-III (b) MIMIC-IV

Figure 2: Performance (AUROC) under different maximum path lengths on MIMIC-III and MIMIC-IV.



(a) In-hospital Mortality (b) T2DM Onset

Figure 3: Cross-attention distribution over relation types for in-hospital mortality and T2DM onset.

Cross-attention over Relation Types. Figure 3 shows the distribution of cross-attention mass over relation types for in-hospital mortality and T2DM onset. The attention mass for each relation type is obtained by averaging cross-attention weights over all prediction instances in the test set, where prediction instances follow the task-specific definition described in Section A.2. The attention patterns differ across tasks: mortality prediction assigns higher weights to relations such as `has_complication` and `treated_by`, whereas T2DM onset emphasizes `has_risk_factor` and `diagnosed_by`. These results suggest that MedCPI selectively attends to task-relevant medical relations, providing an interpretable view of how structured knowledge is leveraged for different tasks.

6 Conclusion

This paper presents **MedCPI**, a unified Construct–Personalize–Integrate framework for knowledge-enhanced clinical prediction. MedCPI addresses key limitations of existing KG-enhanced approaches by (1) constructing a task-specific Con-

Variant	In-hospital Mortality				T2DM Onset			
	MIMIC-III		MIMIC-IV		MIMIC-III		MIMIC-IV	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
w/o Construct	0.684(0.007)	0.219(0.010)	0.792(0.005)	0.155(0.012)	0.771(0.006)	0.412(0.008)	0.728(0.004)	0.276(0.006)
w/o Personalize	0.689(0.006)	0.223(0.011)	0.797(0.004)	0.162(0.010)	0.778(0.004)	0.418(0.006)	0.736(0.003)	0.284(0.005)
w/o Integrate (PKG-only)	0.660(0.008)	0.205(0.012)	0.770(0.006)	0.138(0.010)	0.755(0.006)	0.400(0.010)	0.707(0.005)	0.255(0.008)
w/o Integrate (EHR-only)	0.655(0.010)	0.205(0.012)	0.775(0.008)	0.110(0.010)	0.748(0.012)	0.372(0.018)	0.680(0.009)	0.205(0.014)
Full MedCPI	**0.713**(0.008)	**0.242**(0.018)	**0.809**(0.006)	**0.182**(0.012)	**0.798**(0.002)	**0.425**(0.006)	**0.763**(0.001)	**0.313**(0.002)

Table 4: Ablation results of MedCPI for in-hospital mortality and T2DM onset on MIMIC-III and MIMIC-IV.

cept MKG through task-guided schema induction and normalization, (2) building patient-specific PKGs via controlled local expansion and short path search, and (3) integrating structured medical knowledge with longitudinal EHR representations through time-aware encoding and cross-attention. Experiments on MIMIC-III and MIMIC-IV across four clinical prediction tasks demonstrate that MedCPI consistently outperforms EHR-only models, static KG-enhanced methods, and prior PKG reasoning approaches. Ablation studies confirm that the three stages contribute complementary benefits, while additional analyses provide further insights into effective knowledge utilization. Overall, MedCPI provides a principled and effective approach for incorporating task-relevant and patient-specific medical knowledge into clinical prediction, highlighting the importance of task-aware construction and time-aware integration of knowledge.

Limitations

We discuss several limitations that mainly reflect the scope of this work and suggest directions for future extensions. First, we focus on task-guided construction of Concept MKGs and PKGs based on structured relations in the underlying MKG. Extending the framework to incorporate richer knowledge sources, such as additional ontologies and curated resources, and to support alternative schema design choices is an important direction. Second, our Personalize stage adopts fixed, interpretable constraints to keep PKG construction controllable and efficient. An important direction is to develop adaptive personalization strategies that adjust these constraints based on task requirements or patient context, while maintaining interpretability and efficiency. Third, we primarily study structured EHRs together with KGs. Incorporating unstructured clinical text and other modalities is a natural extension, and may further improve performance in settings where such information is critical. Finally, we evaluate MedCPI on MIMIC-III and MIMIC-IV. Broader validation on additional datasets, care settings, and healthcare systems will be necessary to further assess generalization and practical applicability.

Ethical Considerations

This work follows the ACL Code of Ethics. We obtained access to MIMIC-III and MIMIC-IV through PhysioNet[1] by completing the required human-subjects training and signing the corresponding data use agreements, and we adhere to all PhysioNet requirements for data access, storage, and use. The datasets are de-identified; we do not attempt re-identification or linkage to external sources, and we do not share any patient-level data with external APIs or third-party services. Our method is intended for retrospective research and should not be used as a standalone clinical decision-making tool; any real-world deployment would require careful external validation, monitoring for subgroup disparities, and appropriate clinical oversight.

Acknowledgments

This work is partially supported by the UKRI SLAIDER project and the MRC SLAIDER-QA project. The authors acknowledge the use of resources provided by the Isambard-AI National AI Research Resource (AIRR). Isambard-AI is operated by the University of Bristol and is funded by the UK Government's Department for Science, Innovation and Technology (DSIT) via UK Research and Innovation, and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023] (McIntosh-Smith et al., 2024).

References

Ali Amirahmadi, Mattias Ohlsson, and Kobra Etminani. 2023. Deep learning prediction models based on ehr trajectories: A systematic review. *Journal of biomedical informatics*, 144:104430.

[1] https://physionet.org/

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 2017. Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 787–795.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29.

Edward Choi, Cao Xiao, Walter Stewart, and Jimeng Sun. 2018. Mime: Multilevel medical embedding of electronic health records for predictive healthcare. *Advances in neural information processing systems*, 31.

Edward Choi, Zhen Xu, Yujia Li, Michael Dusenberry, Gerardo Flores, Emily Xue, and Andrew Dai. 2020. Learning the graphical structure of electronic health records with graph convolutional transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 606–613.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Xueren Ge, Abhishek Satpathy, Ronald Dean Williams, John Stankovic, and Homa Alemzadeh. 2024. Dkec: domain knowledge enhanced multi-label classification for diagnosis prediction. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 12798–12813.

Pengcheng Jiang, Cao Xiao, Adam Cross, and Jimeng Sun. 2024. Graphcare: Enhancing healthcare predictions with personalized knowledge graphs. In *The Twelfth International Conference on Learning Representations*.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Balston, Jack Ross, Esther Idowu, and 1 others. 2024. Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *The Lancet Digital Health*, 6(4):e281–e290.

Isotta Landi, Benjamin S Glicksberg, Hao-Chih Lee, Sarah Cherng, Giulia Landi, Matteo Danieletto, Joel T Dudley, Cesare Furlanello, and Riccardo Miotto. 2020. Deep representation learning of electronic health records to unlock patient stratification at scale. *NPJ digital medicine*, 3(1):96.

Jin Li, Benjamin J Cairns, Jingsong Li, and Tingting Zhu. 2023. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. *NPJ digital medicine*, 6(1):98.

Yikuan Li, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Shishir Rao, Abdelaali Hassaine, Dexter Canoy, Thomas Lukasiewicz, and Kazem Rahimi. 2022. Hi-behrt: hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *IEEE journal of biomedical and health informatics*, 27(2):1106–1117.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 4228–4238.

Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. 2020. Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 647–656.

Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 2017. 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*, pages 1903–1911.

Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. 2018. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 743–752.

Liantao Ma, Chaohe Zhang, Yasha Wang, Wenjie Ruan, Jiangtao Wang, Wen Tang, Xinyu Ma, Xin Gao, and

Junyi Gao. 2020. Concare: Personalized clinical feature embedding via capturing the healthcare context. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 833–840.

Chengsheng Mao, Liang Yao, and Yuan Luo. 2022. Medgcn: Medication recommendation and lab test imputation via graph convolutional networks. *Journal of Biomedical Informatics*, 127:104000.

Simon McIntosh-Smith, S. R. Alam, and Christopher Woods. 2024. Isambard-ai: a leadership class supercomputer optimised specifically for artificial intelligence. *arXiv preprint arXiv:2410.11199*.

OpenAI. 2025. Gpt-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, and 1 others. 2018. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer.

Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. 2019a. Pre-training of graph augmented transformers for medication recommendation. *arXiv preprint arXiv:1906.00346*.

Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. 2019b. Gamenet: Graph augmented memory networks for recommending medication combination. In *proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1126–1133.

Nenad Tomašev, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, and 1 others. 2019. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116–119.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Cao Xiao, Edward Choi, and Jimeng Sun. 2018. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10):1419–1428.

Yongxin Xu, Xu Chu, Kai Yang, Zhiyuan Wang, Peinie Zou, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2023. Seqcare: Sequential training with external medical knowledge graph for diagnosis prediction in healthcare data. In *Proceedings of the ACM Web Conference 2023*, pages 2819–2830.

Kai Yang, Yongxin Xu, Peinie Zou, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2023a. Kerprint: local-global knowledge graph enhanced diagnosis prediction for retrospective and prospective interpretations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5357–5365.

Zongbao Yang, Yuchen Lin, Yinxin Xu, Jinlong Hu, and Shoubin Dong. 2023b. Interpretable disease prediction via path reasoning over medical knowledge graphs and admission history. *Knowledge-Based Systems*, 281:111082.

Zhihao Yu, Chu Xu, Yujie Jin, Yasha Wang, and Junfeng Zhao. 2024. Smart: Towards pre-trained missing-aware model for patient health status prediction. *Advances in Neural Information Processing Systems*, 37:63986–64009.

Muhan Zhang, Christopher R King, Michael Avidan, and Yixin Chen. 2020. Hierarchical attention propagation for healthcare representation learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 249–256.

A Data and Tasks

A.1 EHR Data Preprocessing and Cohort Construction

We build patient-level inputs from raw MIMIC records by aggregating events within each visit into a visit, ordering visits by admission time, and removing duplicate codes within a visit. We exclude patients with fewer than two visits and apply task-specific cohort criteria. The patient EHR subgraph G_p^{ehr} follows the definition in Section 3 and is illustrated in Figure 4.

After preprocessing, diagnosis, procedure, and medication codes are aligned to UMLS concepts using SapBERT, and the aligned concepts are used as anchors for subsequent KG construction. Table 6 summarizes the resulting aligned EHR statistics on MIMIC-III and MIMIC-IV. These counts correspond to the final aligned EHR views used in the Construct stage.

A.2 Prediction Tasks and Label Construction

We derive task labels from visit timestamps (admission and discharge times, and mortality timestamps) and diagnosis ICD codes. For all tasks, input features are constructed strictly from records available before the prediction time to avoid label leakage.

General-purpose tasks (visit-level). For each visit, we create one prediction instance using the patient's prior visits as input. For *in-hospital mortality*, the label is positive if death occurs during the target visit, determined by comparing the mortality timestamp with the admission and discharge timestamps. For *30-day readmission*, the label is positive if the patient has a subsequent visit within 30 days after discharge from the target visit; visits with no subsequent visits within 30 days are labeled negative.

Disease-specific tasks (patient-level). Each patient contributes one instance with an observation window of 12 months followed by a prediction window of 12 months. We perform incident prediction by excluding patients with any record of the target disease before the end of the observation window, and assigning a positive label if the target disease first appears within the prediction window. T2DM is identified by ICD codes, namely ICD-9 $250.x0$ and $250.x2$ in MIMIC-III, and ICD-10 $E11.*$ in MIMIC-IV. CAD is identified by ICD code ranges, namely ICD-9 410–414 in MIMIC-III, and ICD-10 $I20$–$I25$ in MIMIC-IV.

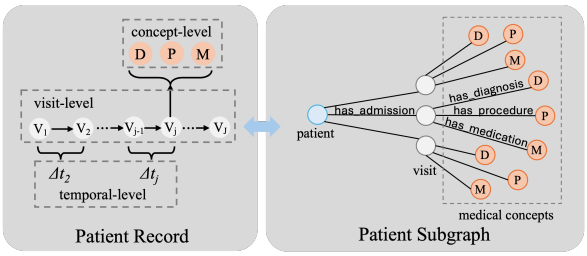


Figure 4: Illustration of the patient EHR subgraph G_p^{ehr}.

Table 5 further reports the split-wise instance counts for the final task-specific prediction instances used in all experiments.

B Technical Details

B.1 Notation

Table 7 summarizes the notation used throughout the paper. We use calligraphic letters (e.g., $\mathcal{P}, \mathcal{T}$) to denote sets, bold lowercase letters (e.g., $\mathbf{z}$) to denote vector representations, and superscripts to distinguish different graph variants (e.g., unified MKG, task-specific Concept MKG, and PKGs). Unless otherwise specified, all learnable parameters are shared across patients and are task-specific only when explicitly indexed by τ.

B.2 Schema Induction

B.2.1 Algorithm

We provide the end-to-end algorithm for inducing a task-specific canonical relation schema R_τ and the mapping $f_\tau : R \to R_\tau \cup \{\texttt{IGNORE}\}$ used in Section 4.1.

Relation Embedding and Clustering. We embed each relation $r \in R$ by encoding a short text x_r that includes the relation name and a small set of representative triples (Algorithm 1, lines 2–5). Relation texts are encoded with SapBERT (`SapBERT-from-PubMedBERT-fulltext`; max length 128; encoder batch size 32). Specifically, we sample up to $S = 5$ triples instantiating r from $\mathcal{E}$, remove duplicates, and serialize each triple as a compact `head-relation-tail` snippet appended to the relation name. We then apply agglomerative clustering on the normalized relation embeddings using cosine distance and average linkage, with distance threshold 0.15, to obtain global semantic clusters $\{K_i\}_{i=1}^m$ shared across tasks. For each cluster K_i, we form a summary s_i by listing

Dataset	Task	Train (total)	Valid (total)	Test (total)
MIMIC-III	In-hospital mortality	9,992	1,216	1,248
	30-day readmission	9,992	1,216	1,248
	T2DM onset	1,419	177	178
	CAD onset	1,214	151	153
MIMIC-IV	In-hospital mortality	258,983	31,460	32,133
	30-day readmission	258,983	31,460	32,133
	T2DM onset	38,367	4,795	4,797
	CAD onset	38,612	4,826	4,827

Table 5: Train/validation/test instance counts for the final task-specific prediction instances. All splits are patient-level disjoint.

Statistic	MIMIC-III	MIMIC-IV
#Patients	7,537	100,163
#Visits	19,993	422,739
Mapped diagnoses	253,991	5,078,670
Mapped procedures	76,463	607,773
Mapped medications	558,225	10,132,439
Diagnosis missing mapping	6,264	104,963
Procedure missing mapping	582	3,197
Medication non-drug supply	1,144	9,145

Table 6: Aligned EHR statistics after concept normalization. Counts are reported on the final patient-level aligned EHR views used for downstream Construct and Personalize stages.

its member relation names and up to $M = 12$ representative triples sampled from the union of triples covered by relations in K_i (lines 6–9).

Prompting and Canonical Relation Naming. For each task τ, we prompt an instruction-tuned LLM with the task description T_τ and each cluster summary s_i to decide task relevance and assign a canonical relation name to relevant clusters (Algorithm 1, lines 10–15). We use the prompt template in Fig. 5, which constrains the output to a compact JSON object with a binary `keep` decision and a single `canonical_name`.

Consistency Resolution and Task-Specific Mapping. We normalize canonical names by lowercasing and enforcing `snake_case`, and merge clusters that receive the same normalized name to resolve cross-cluster inconsistencies. The task schema is then defined as $R_\tau = \{\text{name}_\tau[i] \mid \text{name}_\tau[i] \neq \texttt{IGNORE}\}$, and the task-specific mapping f_τ assigns each original relation r to the canonical name of its cluster, or to `IGNORE` otherwise (Algorithm 1, lines 16–22).

B.3 KG Construction

Concept MKG construction (global 1-hop). Given the task-normalized KG $G_\tau = (V, R_\tau, \mathcal{E}_\tau)$, the Concept MKG G_τ^{concept} is constructed via a global 1-hop expansion from EHR concepts. For each task τ, patients are first split into training, validation, and test sets at the patient level. The Concept MKG is constructed using only medical concepts observed in the training split (including diagnoses, procedures, and medications after code normalization).

Diagnosis, procedure, and medication codes in the EHR are aligned to UMLS concepts using SapBERT (Liu et al., 2021) (`SapBERT-from-PubMedBERT-fulltext`) to enable concept-level linking to the MKG. All triples in G_τ whose head or tail matches these training concepts are collected. Among them, only triples whose endpoints belong to clinically meaningful semantic types S_{med} are retained, where S_{med} is implemented using three coarse UMLS semantic groups: *Disorders*, *Procedures*, and *Chemicals & Drugs*. In the local UMLS SemGroups resource used by our pipeline, findings and signs/symptoms are covered by *Disorders*. Duplicate edges after canonical-relation rewriting are removed to obtain G_τ^{concept}, which is shared across patients for task τ and serves as the knowledge pool for personalization.

Table 8 reports the resulting Construct-stage statistics. For each dataset–task pair, we report the retained schema size, the number of training-split anchor concepts, and the final size of the extracted Concept MKG. As described in Section 4.1, the relation schema is task-specific but shared across datasets. For the onset tasks, we use a more conservative relation mapping so that diagnosis- and treatment-adjacent relation families are represented with fewer label-proximal variants.

PKG construction. For each patient p and task τ, we construct a PKG $G_{p,\tau}^{\text{pkg}}$ by selecting a patient-centric subgraph from G_τ^{concept} and augmenting it with task-relevant multi-hop paths. For visit-level tasks (in-hospital mortality and 30-day readmission), we use the concepts observed up to the

Symbol	Description
Patients, tasks, prediction instances, and EHR sequences	
$\mathcal{P}$, p	Patient set and patient index.
$\mathcal{T}$, τ, T_τ	Task set, task index, and task description.
$\mathcal{I}_\tau$, i	Prediction instance set for task τ and instance index.
$\mathbf{E}_p = [V_{p,1}, \ldots, V_{p,L_p}]$, L_p	Patient p's time-ordered visit sequence and its length.
$V_{p,t} = (\mathcal{D}_{p,t}, \mathcal{P}_{p,t}, \mathcal{M}_{p,t})$	Visit t with diagnosis/procedure/medication code sets.
$\mathcal{D}_{p,t}$, $\mathcal{P}_{p,t}$, $\mathcal{M}_{p,t}$	Diagnosis, procedure, and medication code sets at visit t.
y_i, $\hat{y}_i$	Ground-truth and predicted outcome/probability for prediction instance i.
EHR graph and patient concepts	
$G_p^{\text{ehr}} = (V_p^{\text{ehr}}, R^{\text{ehr}}, \mathcal{E}_p^{\text{ehr}})$	Patient EHR subgraph with visit/concept nodes and visit-concept edges.
V_p^{ehr}, R^{ehr}, $\mathcal{E}_p^{\text{ehr}}$	Node set, EHR edge types (e.g., `has_diagnosis`/`has_procedure`/`has_medication`), and edges.
C_p	Medical concepts appearing in patient p's EHR.
Unified MKG, schema induction, and normalization	
$G = (V, R, \mathcal{E})$	Unified medical knowledge graph (MKG) integrated from multiple sources.
V, R, $\mathcal{E}$	Concept set, relation-type set, and triple set $\mathcal{E} \subseteq V \times R \times V$ with triples (h, r, t).
R_τ, f_τ	Task-induced canonical relation schema and relation mapping.
$f_\tau : R \to R_\tau \cup \{\texttt{IGNORE}\}$	Maps each original relation to a canonical one (or `IGNORE`).
$G_\tau = (V, R_\tau, \mathcal{E}_\tau)$, $\mathcal{E}_\tau$	Task-normalized MKG and its normalized triples after applying f_τ.
Concept MKG and PKG	
S_{med}	Clinically meaningful concept types used to filter Concept MKG triples.
$G_\tau^{\text{concept}} = (V_\tau^{\text{concept}}, R_\tau, \mathcal{E}_\tau^{\text{concept}})$	Task-specific Concept MKG extracted by global 1-hop expansion around EHR concepts.
$G_{p,\tau}^{\text{pkg}} = (V_{p,\tau}^{\text{pkg}}, R_\tau \cup R^{\text{ehr}}, \mathcal{E}_{p,\tau}^{\text{pkg}})$	PKG for patient p and task τ.
$V_{p,\tau}^{\text{pkg}}$, $\mathcal{E}_{p,\tau}^{\text{pkg}}$	Node set and edge/triple set of the PKG.
PKG encoder (R-GCN)	
$\mathbf{h}_v^{(0)}$, $\mathbf{h}_v^{(l)}$, $\mathbf{h}_v^{(L)}$	Initial, layer-l, and final node representations for node v in the PKG.
$\mathcal{N}_r(v)$, $c_{v,r}$	Neighbors of v under relation r, and normalization constant.
$\mathbf{W}_r^{(l)}$, $\mathbf{W}_0^{(l)}$, $\sigma(\cdot)$	Relation-specific/self-loop transforms and activation in R-GCN.
EHR encoder (HTT) and pooling	
$\mathbf{h}_{p,t}$, $\alpha_{p,t}$, $\mathbf{z}_p^{\text{ehr}}$	Visit hidden state, visit attention weight, and pooled patient-level EHR representation.
EHR–PKG fusion and prediction	
$\mathbf{q}_i$	EHR context representation for prediction instance i.
$\mathbf{c}_i$	Cross-attended PKG summary for instance i.
$\mathbf{z}_i^{\text{fuse}}$	Fused EHR–PKG representation for instance i.
$\phi(\cdot)$, $\mathbf{w}_\tau$, b_τ	Fusion MLP and task-specific prediction head parameters.
$\ell_\tau(\cdot, \cdot)$, $\mathcal{L}_\tau$	Pointwise loss and overall training loss for task τ.
Key hyperparameters	
K_{nbr}, K_{path}, L	Max 1-hop neighbors per anchor, max 2-hop paths per anchor pair, and R-GCN layers.

Table 7: **Summary of notation for MedCPI.**

current visit as patient anchors; for patient-level onset tasks (T2DM onset and CAD onset), we use the concepts observed within the task-defined observation window as anchors. Starting from these anchor concepts, we take the induced 1-hop patient subgraph from G_τ^{concept} and then insert additional multi-hop paths returned by the path selection procedure described below. The final $G_{p,\tau}^{\text{pkg}}$ is the union of the 1-hop patient subgraph and the selected paths, with duplicated edges removed.

Path selection and graph size control. To keep $G_{p,\tau}^{\text{pkg}}$ compact, we control both the 1-hop expansion and the insertion of multi-hop paths. For the 1-hop induced subgraph, we cap the number of 1-hop neighbors per anchor concept by K_{nbr} and remove duplicate edges. For multi-hop augmentation, we only insert 2-hop *bridging* paths between anchor concepts that co-occur in the patient EHR. Let C_p be the anchor concepts of patient p. For each anchor pair (c_i, c_j) with $c_i, c_j \in C_p$, we enumerate 2-hop paths on G_τ, i.e., $c_i \to u \to c_j$ (or the reverse), remove duplicates, and keep at most K_{path} paths per anchor pair. To ensure comparable PKG sizes across tasks and datasets, we fix

Dataset	Task	#Rel.	Train Anchor Concepts	#Nodes	#Edges
MIMIC-III	In-hospital mortality	10	6,729	21,231	78,312
	30-day readmission	8	6,729	25,500	99,450
	T2DM onset	9	3,455	16,912	36,200
	CAD onset	8	3,459	16,859	35,971
MIMIC-IV	In-hospital mortality	10	35,426	27,419	123,887
	30-day readmission	8	35,426	33,679	152,536
	T2DM onset	9	18,818	25,955	64,870
	CAD onset	8	19,097	25,885	65,475

Table 8: Construct-stage statistics for the final task-specific Concept MKGs. For each dataset–task pair, we report retained schema size, training-split anchor concepts, and final graph size.

Hyperparameter	Candidates
Embedding dim d	{64, 128, 256}
Hidden dim	{64, 128, 256}
#HTT layers N	{1, 2, 3}
#R-GCN layers L	{1, 2, 3}
Learning rate	{1e−4, 3e−4, 1e−3}
Weight decay	{0, 1e−5, 1e−4}
Dropout	{0.0, 0.1, 0.3}
Batch size	{32, 64, 128}
Max path length (PKG)	{1, 2, 3, 4, 5}

Table 9: Hyperparameter search space for tuning MedCPI on each dataset–task pair (selected by validation AUROC).

$K_{\text{nbr}} = 30$ and $K_{\text{path}} = 2$ in all experiments.

B.4 Model and Training

B.4.1 Model Architecture Summary

MedCPI consists of (i) a time-aware EHR encoder (HTT) that maps the visit sequence to a patient representation, (ii) a PKG encoder implemented as an L-layer R-GCN with relation-specific message passing, and (iii) a cross-attention fusion module that uses the EHR representation as the query and PKG node representations as keys/values, followed by a two-layer MLP for prediction. Padding masks are applied to handle variable-length visit sequences and variable-size PKGs.

B.4.2 Training Protocol and Implementation Details

All tasks are binary prediction. For task τ, given the logit s_i and probability $\hat{y}_i$, we minimize the binary cross-entropy loss and implement it with `BCEWithLogitsLoss` for numerical stability. We train a separate MedCPI model for each dataset–task pair. Hyperparameters are tuned on the validation set and selected by validation AUROC. We optimize all parameters end-to-end using AdamW, apply early stopping based on validation AUROC (patience 10; max 30 epochs), and report mean±std over five random seeds. Dropout is applied in the EHR encoder, R-GCN layers, and the fusion MLP.

B.4.3 Hyperparameter Search for MedCPI

For each dataset–task pair, we tune MedCPI hyperparameters on the validation set and select the best configuration by validation AUROC; the selected configuration is then retrained and evaluated on the test set following the protocol in Appendix B.4.2. Table 9 summarizes the search space. The maximum PKG path length is selected on the validation set, and we use 2-hop paths in the final model. All other PKG construction rules are kept fixed (Appendix B.3) to ensure comparable graph sizes.

C Additional Experiments and Analysis

C.1 Additional Ablations

Table 10 reports the same ablation variants as Table 4 in the main text, but on the remaining two tasks (30-day readmission and CAD onset) across MIMIC-III and MIMIC-IV. We follow the same training and evaluation protocol and report mean (standard deviation) over five random seeds.

C.2 Schema-Induction Robustness and Construct Statistics

We further examine the robustness of the schema-induction procedure used in the *Construct* stage. Specifically, we compare the GPT-5-based pipeline with alternative LLM backbones and two non-LLM schema-induction baselines while keeping the clustering and post-processing pipeline fixed. Table 11 summarizes schema-level agreement and retained schema size across the four tasks.

C.3 Alternative Integration Strategies

To quantify the contribution of the current *Integrate* design, we compare the default cross-attention fusion with three alternative strategies while keeping *Construct*, *Personalize*, and the training protocol fixed. The alternatives include: (i) *Late fusion*, which concatenates the pooled EHR and PKG representations followed by the same prediction head; (ii) *Gated fusion*, which learns a task-dependent

Variant	30-day Readmission				CAD Onset			
	MIMIC-III		MIMIC-IV		MIMIC-III		MIMIC-IV	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
w/o Construct	0.710(0.002)	0.482(0.040)	0.730(0.003)	0.578(0.009)	0.680(0.003)	0.372(0.008)	0.770(0.004)	0.395(0.003)
w/o Personalize	0.713(0.006)	0.484(0.011)	0.733(0.005)	0.581(0.007)	0.684(0.001)	0.375(0.007)	0.773(0.008)	0.398(0.002)
w/o Integrate (PKG-only)	0.655(0.008)	0.440(0.012)	0.670(0.006)	0.500(0.005)	0.655(0.006)	0.350(0.010)	0.740(0.005)	0.375(0.011)
w/o Integrate (EHR-only)	0.662(0.010)	0.461(0.014)	0.678(0.008)	0.512(0.010)	0.658(0.012)	0.352(0.014)	0.757(0.002)	0.353(0.009)
Full MedCPI	**0.732**(0.005)	**0.498**(0.010)	**0.744**(0.004)	**0.606**(0.007)	**0.705**(0.003)	**0.405**(0.006)	**0.797**(0.001)	**0.418**(0.003)

Table 10: Ablation results of MedCPI for 30-day readmission and CAD onset on MIMIC-III and MIMIC-IV.

Method	Jaccard(R_τ) ↑	Keep Agree ↑	Map Agree ↑	#Ret. Rel.
LLM backbones				
GPT-5 (default)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	8.75 (8)
GPT-4o	0.926 (0.889)	0.949 (0.917)	0.934 (0.889)	8.25 (7)
Llama 3.1-70B-Instruct	0.861 (0.778)	0.891 (0.847)	0.872 (0.806)	7.50 (6)
Qwen2.5-72B-Instruct	0.887 (0.833)	0.913 (0.876)	0.904 (0.861)	8.00 (7)
Non-LLM alternatives				
Embedding-based clustering	0.655 (0.455)	0.533 (0.474)	0.431 (0.312)	6.00 (6)
Statistical association filtering	0.542 (0.364)	0.473 (0.404)	0.382 (0.290)	5.00 (5)

Table 11: Schema-induction robustness and method comparison, reported as mean (minimum) over the four tasks. We compare canonical relation sets, cluster-level KEEP/IGNORE decisions, relation-level mappings, and retained schema size.

Method	Mort. AUROC	Mort. AUPRC	T2DM AUROC	T2DM AUPRC
Cross-attn. (default)	0.713	0.242	0.798	0.425
Late fusion	0.707	0.234	0.783	0.401
Gated fusion	0.711	0.240	0.794	0.418
FiLM modulation	0.709	0.237	0.789	0.409

Table 12: Comparison of integration strategies on MIMIC-III in-hospital mortality and T2DM onset.

gate over the EHR and PKG representations before prediction; and (iii) *FiLM-style modulation*, which uses the PKG representation to generate feature-wise modulation parameters for the EHR representation. Table 12 reports predictive performance on two representative MIMIC-III tasks: in-hospital mortality and T2DM onset.

Algorithm 1 Task-guided schema induction.

Input: relation types R, triples $\mathcal{E}$; task descriptions $\{T_\tau\}_{\tau \in \mathcal{T}}$

Parameter: samples S; cluster examples M; $\mathrm{Embed}(\cdot)$; $\mathrm{Cluster}(\cdot)$

Output: $\{(R_\tau, f_\tau)\}_{\tau \in \mathcal{T}}$, where $f_\tau : R \to R_\tau \cup \{\texttt{IGNORE}\}$

(Shared) relation embedding and clustering
for $r \in R$ **do**
  $x_r \leftarrow \mathrm{Summarize}(r, \mathcal{E}, S)$
  $\mathbf{e}_r \leftarrow \mathrm{Embed}(x_r)$
end for
$\{K_i\}_{i=1}^{m} \leftarrow \mathrm{Cluster}(\{\mathbf{e}_r\}_{r \in R})$
for $i = 1$ to m **do**
  $s_i \leftarrow \mathrm{Summarize}(K_i, \mathcal{E}, M)$
end for
(Per-task) canonical naming and mapping construction
for $\tau \in \mathcal{T}$ **do**
  for $i = 1$ to m **do**
    $(\mathsf{keep}, \mathsf{name}) \leftarrow \mathrm{LLM}(T_\tau, s_i)$
    $\mathrm{name}_\tau[i] \leftarrow \mathsf{name}$ if keep else $\texttt{IGNORE}$
  end for
  $\mathrm{name}_\tau \leftarrow \mathrm{ResolveConsistency}(\mathrm{name}_\tau)$
  $R_\tau \leftarrow \{\mathrm{name}_\tau[i] \mid \mathrm{name}_\tau[i] \neq \texttt{IGNORE}\}$
  for $r \in R$ **do**
    $i \leftarrow \mathrm{cluster}(r)$
    $f_\tau(r) \leftarrow \mathrm{name}_\tau[i] \quad f_\tau(r) = \texttt{IGNORE}$ *if cluster i is discarded*
  end for
end for
return $\{(R_\tau, f_\tau)\}_{\tau \in \mathcal{T}}$

You are an expert in clinical informatics and medical knowledge graphs. Your task is to assess whether a cluster of relation types is useful for a given clinical prediction task. If the cluster is useful, assign a single canonical relation name.
Inputs:
(1) **Task description:** {TASK_DESCRIPTION T_τ}
(2) **Relation cluster summary:**
{CLUSTER_SUMMARY s_i}
Decision criteria:

- Set keep=true only if the relations in the cluster provide task-relevant clinical knowledge that can plausibly support prediction for the given task (directly or via short clinical reasoning).
- Otherwise set keep=false and output canonical_name="IGNORE".

Naming rules (only if keep=true):

- Use **lower_snake_case**.
- Use a **verb-like** relation name (e.g., treated_by, diagnosed_by, has_risk_factor, has_complication, has_symptom).
- Use at most **four words**.
- Prefer specific, clinically meaningful names when possible; avoid overly vague names unless a broader association label is needed to capture the retained relation family.
- Output **one** canonical name only.

Output (return **only** a JSON object; no extra text):
{ "keep": true/false, "canonical_name": "<name or IGNORE>" }

Figure 5: Prompt template for task-guided relevance filtering and canonical relation naming.