THCM-CAL: Temporal-Hierarchical Causal Modelling with Conformal Calibration for Clinical Risk Prediction

Xin Zhang[1], **Qiyu Wei**[1], **Yingjie Zhu**[2], **Fanyi Wu**[1], **Sophia Ananiadou**[1]
[1]The University of Manchester [2]Harbin Institute of Technology
{xin.zhang-41@postgrad., qiyu.wei@postgrad., fanyi.wu@}manchester.ac.uk
sophia.ananiadou@manchester.ac.uk, zhuyj@stu.hit.edu.cn

Abstract

Automated clinical risk prediction from electronic health records (EHRs) demands modeling both structured diagnostic codes and unstructured narrative notes. However, most prior approaches either handle these modalities separately or rely on simplistic fusion strategies that ignore the directional, hierarchical causal interactions by which narrative observations precipitate diagnoses and propagate risk across admissions. In this paper, we propose THCM-CAL, a Temporal-Hierarchical Causal Model with Conformal Calibration. Our framework constructs a multimodal causal graph where nodes represent clinical entities from two modalities: Textual propositions extracted from notes and ICD codes mapped to textual descriptions. Through hierarchical causal discovery, THCM-CAL infers three clinically grounded interactions: intra-slice same-modality sequencing, intra-slice cross-modality triggers, and inter-slice risk propagation. To enhance prediction reliability, we extend conformal prediction to multi-label ICD coding, calibrating per-code confidence intervals under complex co-occurrences. Experimental results on MIMIC-III and MIMIC-IV demonstrate the superiority of THCM-CAL.

1 Introduction

Accurate clinical risk prediction from Electronic Health Records (EHRs) (Evans, 2016) is essential for enabling timely clinical interventions and improving treatment effects (Choi et al., 2016; Miotto et al., 2016). EHRs comprise two complementary data modalities: Structured diagnostic codes drawn from the International Classification of Diseases (ICD) (Choi et al., 2017; Bodenreider, 2004) and Unstructured narrative notes that chronicle patient observations and interventions over time (Huang et al., 2019). Leveraging both modalities can enhance automated code assignment (Sun et al., 2024), risk stratification (Tsai et al., 2025), and downstream decision support (Lu et al., 2021; Xu et al., 2023).

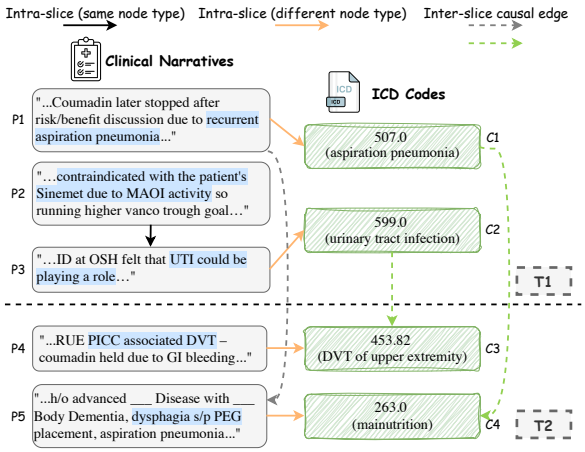

Figure 1: An illustrative two-slice causal graph over textual propositions and ICD nodes. Three directed edge types are learned: intra-slice sequencing ($P_2 \rightarrow P_3$), intra-slice cross-modality triggers ($P_1 \rightarrow C_1$), and inter-slice propagation ($C_1 \rightarrow C_4$).

Previous approaches as shown in Table 1 fall into two broad categories. On one hand, text-centric models such as CAML (Mullenbach et al., 2018) and ZAGNN (Rios and Kavuluru, 2018) leverage label-wise attention over narrative notes but entirely ignore structured code context, while code-focused methods like RETAIN (Choi et al., 2016) and GRAM (Choi et al., 2017) attend only to historical ICD sequences and overlook the rich semantics of free-text observations. On the other hand, recent transformer-based and text-driven frameworks including Chet (Lu et al., 2022), DistilBioBERT (Rohanian et al., 2024), BioMedLM (Bolton et al., 2024), GatorTron (Yang et al., 2022) and DKEC (Ge et al., 2024) improve representation power or inject external knowledge yet still treat text and codes as static co-occurrences. They remain task-agnostic during pretraining, neglect fine-grained narrative-to-code triggers and

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 916–928
November 4-9, 2025

Table 1: Overview of representative clinical risk prediction methods, comparing their input modalities, support for causal structure discovery, uncertainty estimation capabilities, and temporal hierarchical modeling.

Method	Modalities	Causal Discovery	Uncertainty	Temporal Hi
CAML (Mullenbach et al., 2018)	Text	✗	✗	✗
ZAGNN (Rios and Kavuluru, 2018)	Text	✗	✗	✗
DistilBioBERT (Rohanian et al., 2024)	Text	✗	✗	✗
BioMedLM (Bolton et al., 2024)	Text	✗	✗	✗
GatorTron (Yang et al., 2022)	Text	✗	✗	✗
DKEC (Ge et al., 2024)	Code+Text	✗	✗	✗
Chet (Lu et al., 2022)	Code+Text	Dynamic graph	✗	✗
CDANs (Ferdous et al., 2023)	Codes	Instantaneous & lagged edges	✗	✓
CACHE (Xu et al., 2022)	Codes	Hypergraph-based	✗	✗
COMPOSER (Shashikumar et al., 2021)	Codes	✗	Single-task CP	✗
THCM-CAL (Ours)	**Code+Text**	**Intra/inter-slice, cross-modality**	**Multi-label CP**	✓

rely on fixed graph structures that cannot adapt to patient-specific time-varying causal relationships.

Taking the patient trajectory in Figure 1 as an example. In the first admission (T_1), the note reports a medication contraindication (P_2) that prompts a reassessment of the urinary tract infection etiology (P_3), capturing an intra-slice sequencing dependency. The description "recurrent aspiration pneumonia" (P_1) directly precipitates the assignment of ICD code J69.0 (C_1), exemplifying an intra-slice cross-modality trigger. Finally, the diagnosis C_1 in T_1 increases the likelihood of a related complication (C_4) in the subsequent admission (T_2), demonstrating inter-slice risk propagation. These patterns highlight three critical dimensions of clinical causality that current models overlook: how narrative observations temporally trigger diagnoses, how events in one hospitalization propagate risk to the next, and how causal dependencies span hierarchical temporal scales. Existing approaches either operate on a single slice or treat text–code interactions as undirected associations, and thus fail to capture these directed, modality-aware causal mechanisms.

To address these gap, we propose THCM-CAL, a Temporal–Hierarchical Causal Model with Conformal Calibration for Clinical Risk Prediction. Our framework proceeds in four stages: First, we segment each admission's narrative into diagnostically relevant sections; Second, we encode both propositions and code descriptions with BERT and project them into a shared embedding space to form textual and code nodes. Third, we sample intra-slice same-modality, intra-slice cross-modality, and inter-slice propagation edges using Gumbel–Softmax (Jang et al., 2016) with acyclicity constraints; then fuse these edges via graph (Rossi et al., 2020) message passing to produce per-admission embeddings. Finally, we apply split conformal prediction (Shafer and Vovk, 2008) to the multi-label probabilities, generating valid confidence sets at a desired coverage level even under complex code co-occurrences. Our key contributions are:

- We propose a causal framework to unify intra-visit sequencing, cross-modality triggers, and cross-visit propagation in a hierarchical graph.
- We develop a split-conformal calibration method that provides distribution-free uncertainty guarantees on the prediction of diagnostic codes.
- We demonstrate that explicit causal modeling of multimodal interactions yields obvious gains in performance, interpretability, and robustness, with ablations showing each module contributes.

2 Related Work

We categorize prior work into three areas: text-centric diagnosis prediction, temporal causal discovery, and uncertainty quantification.

2.1 Text-Centric Diagnosis Prediction

Automated diagnosis-code prediction from free-text clinical narratives is commonly formulated as a multi-label text classification task mapping notes to sets of ICD codes. Early label-wise attention networks such as CAML (Mullenbach et al., 2018) and ZAGNN (Rios and Kavuluru, 2018) assign each code its own attention mechanism to highlight relevant text spans and capture label co-occurrence

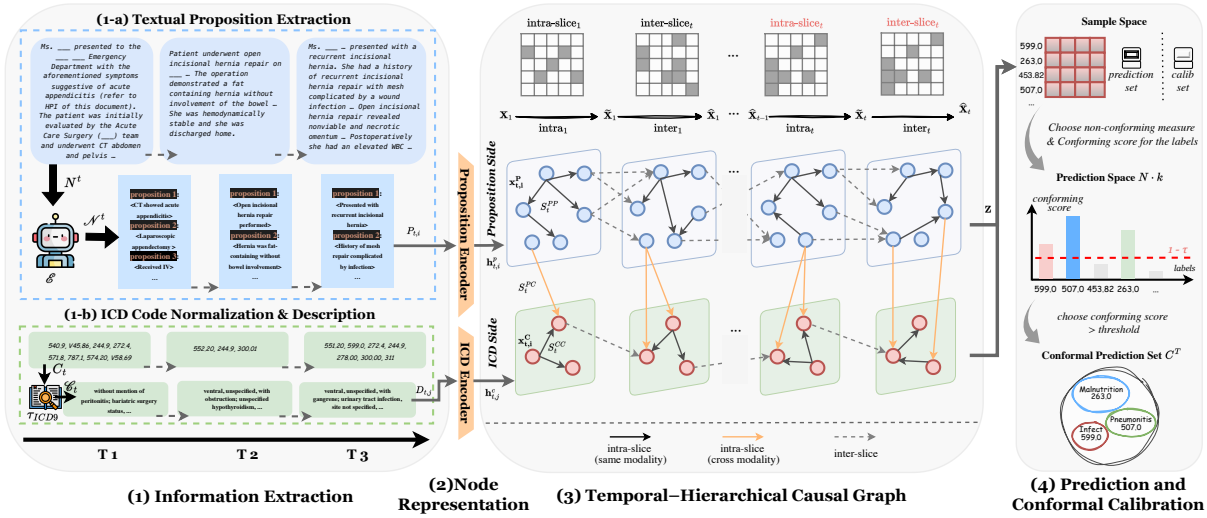

Figure 2: Overview of THCM-CAL. A four-stage pipeline for clinical risk prediction, which consists of: (1) Extracting diagnostic propositions and normalize ICD descriptions, (2) Embeding nodes with BERT, (3) Building and fuse a temporal–hierarchical causal graph via Gumbel–Softmax and message passing, and (4) Appling split conformal prediction for calibrated multi-label ICD coding.

and hierarchy. More recent transformer-based approaches including DistilBioBERT (Rohanian et al., 2024), BioMedLM (Bolton et al., 2024) and GatorTron (Yang et al., 2022) learn rich contextual embeddings via large-scale pre-training but remain task-agnostic and often disregard explicit code dependencies at fine-tuning. Domain-knowledge frameworks such as DKEC (Ge et al., 2024) integrate external ontologies to inform prediction but rely on static graph structures that may not reflect patient-specific interactions between text and codes. Despite their strengths, these methods overlook directional triggers between narrative findings and subsequent codes.

2.2 Causal Discovery in Medical

Causal discovery in EHRs has mainly focused on structured codes. SemDBN (Wang et al., 2018) employs ontology-augmented Bayesian networks for sepsis onset prediction yet excludes unstructured narratives. FGES-based techniques (Shen et al., 2020) improve edge orientation in static cohorts, and CDANs (Ferdous et al., 2023) extend causal discovery to time series by modeling lagged dependencies. CACHE (Xu et al., 2022) applies hypergraph learning and counterfactual inference to structured codes but does not capture how textual propositions precipitate specific diagnoses. As a result, these methods cannot reveal cross-modal, directional relationships between narrative observations and ICD assignments.

2.3 Uncertainty Quantification in Clinical Prediction

Conformal prediction provides finite-sample coverage guarantees but has been applied mainly to single-label clinical forecasting. COMPOSER (Shashikumar et al., 2021) and NeuroSep CP (Zhou et al., 2025a) generate confidence intervals for sepsis onset and temporal risk trajectories, respectively, under an assumption of label independence. Sepsyn OLCP (Zhou et al., 2025b) adapts conformal methods to online monitoring but does not address multi-label ICD coding, where code co-occurrences induce complex uncertainty dependencies. Consequently, existing frameworks do not yield valid per-code confidence sets for multi-label diagnosis prediction.

3 THCM-CAL

3.1 Task Definition

We consider a cohort of S patients, each with up to T chronological hospital admissions $A^1, A^2, \ldots, A^T$. For the t-th admission, we denote the raw data by $A^t = (\tau^t, N^t, C^t)$, where $\tau^t \in \mathbb{R}$ is the admission timestamp, N^t is the free-text clinical note, and C^t is the set of recorded ICD-9 codes. After Stage 1 ("Information Extraction"), each admission is transformed into $\mathcal{A}^t = (\tau^t, \mathcal{C}^t, \mathcal{N}^t, \mathcal{P}^t, \mathcal{D}^t)$, where $\mathcal{N}^t = \mathcal{E}(N^t)$ is the cleaned clinical narrative, $\mathcal{C}^t = \tau_{\text{ICD9}}(C^t)$ is the normalized set of ICD-9 codes, $\mathcal{P}^t$ is the extracted propositions and $\mathcal{D}^t$ is the

labels. Given the history of the first $T-1$ admissions, $\{\mathcal{A}^1, \ldots, \mathcal{A}^{T-1}\}$, our objective is to predict the ICD-9 code set at the next admission: $\hat{\mathcal{C}}^T = f_{\text{ICD}}(\mathcal{A}^1, \ldots, \mathcal{A}^{T-1})$. We train and evaluate f_{ICD} under the standard multi-label framework, comparing $\hat{\mathcal{C}}^T$ to the ground-truth $\mathcal{C}^T$.

3.2 Overview of THCM-CAL

Figure 2 summarizes the four connected modules that compose THCM-CAL. Starting from raw EHRs, we first segment each admission's free-text note into diagnostically meaningful sections and invoke a large language model to extract a set of atomic propositions. In parallel, all recorded ICD codes are normalized to ICD-9 and replaced with their canonical textual descriptions. In the second module, every proposition and code description is fed through BERT and projected into a common embedding space, yielding a rich set of node feature vectors. These vectors form the inputs to our third, core component: we assemble a temporal-hierarchical causal graph by sampling three kinds of directed relationships—within-admission links among propositions and among codes, cross-modal links from propositions to codes, and across-admission propagation edges. Edge selection is performed via Gumbel-Softmax under acyclicity constraints, and the resulting multi-slice graph is combined through a graph fusion that propagates messages and pools node representations into a compact embedding per admission. Finally, the fourth stage applies split conformal prediction to the model's multi-label outputs, producing calibrated confidence sets that respect a user-specified coverage level. By chaining these modules—information extraction, node encoding, causal graph construction with fusion, and conformal calibration—THCM-CAL delivers interpretable risk estimates with rigorous uncertainty quantification.

3.3 Stage 1: Information Extraction

We start from the raw clinical note N^t and the raw ICD set C^t, and extract the cleaned narrative $\mathcal{N}^t$ together with normalized ICD-9 codes $\mathcal{C}^t$ via $\mathcal{N}^t = \mathcal{E}(N^t)$ and $\mathcal{C}^t = \tau_{\text{ICD9}}(C^t)$.

Text segmentation & scoring Let $\mathcal{H} = \{h_1, \ldots, h_{|\mathcal{H}|}\}$ be the set of recognized section headings (e.g., "Chief Complaint", "History of Present Illness"). We split N^t into sections $\mathcal{S} = \{ s_i = \text{SEGMENT}(N^t, h_i, h_{i+1}) \mid h_i, h_{i+1} \in \mathcal{H}\}$. If $\mathcal{S} = \emptyset$, let $\mathcal{S} = \{N^t_{[1:4000]}\}$. Let $\text{count}_{\texttt{mg}}(s)$ denote the number of occurrences of the substring "mg" in s, $\mathcal{K}$ denotes a predefined set of medical keywords (e.g., "fever", "cough", "pain") that are strongly indicative of diseases or diagnoses. For each segment $s_i \in \mathcal{S}$, we extract a feature vector $\mathbf{f}(s_i) = [f_1(s_i),\, f_2(s_i),\, f_3(s_i),\, f_4(s_i)]^\top$, where

$$
\begin{aligned}
f_1(s_i) &= \big|\mathcal{K} \cap s_i\big|, \\
f_2(s_i) &= \mathbb{I}\{\texttt{"Diagnosis:"} \in \text{Text}(s_i)\}, \\
f_3(s_i) &= \min\big(\text{count}_{\texttt{mg}}(s_i),\, 9\big), \\
f_4(s_i) &= \mathbb{I}\{\text{SentCount}(s_i) > 2\}.
\end{aligned}
$$

We define the linear scoring function $\text{score}(s_i) = \mathbf{w}^\top \mathbf{f}(s_i)$, $\mathbf{w} = (2,\, 5,\, 1,\, 3)^\top$, and then sort $\mathcal{S}$ by $\text{score}(\cdot)$ in descending order. The top-K segments are retained: $\mathcal{S}^* = \big\{ s_i \in \mathcal{S} \mid \text{rank}_{\mathcal{S}}\big(\text{score}(s_i)\big) \le K\big\}$,where $K = 3$.

Atomic propositions extraction Let $\mathcal{E}\colon s \mapsto \mathcal{P}_s$ denote the GPT-3.5–based extractor that maps a text segment s to a set of "atomic propositions." We then define

$$
\mathcal{P}^t = \text{uniq}\Big(\bigcup_{s \in \mathcal{S}^*} \mathcal{P}_s\Big) = \{\, P_{t,1},\, P_{t,2},\, \ldots,\, P_{t,n_t}\},
$$

where $\text{uniq}(\cdot)$ removes duplicates and enforces an fixed ordering. These $P_{t,i}$ serve as the proposition-node set for Stage 2.

ICD code normalization We normalize the raw ICD codes C^t via $\mathcal{C}^t = \tau_{\text{icd9}}\big(C^t\big)$, and retrieve their human-readable labels by $\mathcal{D}^t = \big\{\delta(c) \mid c \in \mathcal{C}^t\big\} = \{D_{t,1}, \ldots, D_{t,J_t}\}$, where τ_{icd9} is a standard ICD-9 mapping, δ is the lookup from code to description and $D_{t,j}$ is the ICD-description strings. Consequently, the complete output of Stage 1 for admission t is

$$
\mathcal{A}^t = \big(\tau^t,\, \mathcal{C}^t,\, \mathcal{N}^t,\, \mathcal{P}^t,\, \mathcal{D}^t\big).
$$

3.4 Stage 2: Node Representation

BERT Encodings. Let $\text{Enc}_{\text{BERT}} : \mathcal{T} \to \mathbb{R}^d$ denote the BERT(Devlin et al., 2019) mapping from any text token sequence to its $[CLS]$ embedding of dimension d. We write $\mathbf{h}^P_{t,i} = \text{Enc}_{\text{BERT}}\big(P_{t,i}\big)$, $\mathbf{h}^C_{t,j} = \text{Enc}_{\text{BERT}}\big(D_{t,j}\big)$. Thus $\mathbf{h}^P_{t,i},\, \mathbf{h}^C_{t,j} \in \mathbb{R}^d$.

Modality-Specific Projections. To bring propositions and code descriptions into a shared d'-dimensional space, we apply two learnable linear

projections with nonlinearity:

$$\mathbf{x}^P_{t,i} = \mathrm{Proj}_P(\mathbf{h}^P_{t,i}) = \phi\big(W_P\,\mathbf{h}^P_{t,i} + b_P\big),$$
$$\mathbf{x}^C_{t,j} = \mathrm{Proj}_C(\mathbf{h}^C_{t,j}) = \phi\big(W_C\,\mathbf{h}^C_{t,j} + b_c\big),$$

where $W_P,\ W_C \in \mathbb{R}^{d'\times d},\ b_P,\ b_C \in \mathbb{R}^{d'}$, and $\phi(\cdot)$ denotes an element-wise activation (e.g. ReLU) optionally combined with dropout.

Feature Matrix Assembly. Finally, we concatenate all proposition- and description-level vectors into the admission-level feature matrix

$$\mathbf{X}^t = \big[\mathbf{x}^P_{t,1}, \ldots, \mathbf{x}^P_{t,I_t},\ \mathbf{x}^C_{t,1}, \ldots, \mathbf{x}^C_{t,J_t}\big] \in \mathbb{R}^{d'\times ND_t},$$

where $ND_t = I_t + J_t$. This $\mathbf{X}^t$ serves as the node-feature input to Stage 3.

3.5 Stage 3: Temporal–Hierarchical Causal Graph

Inspired by the temporal causal graph construction in RealTCD (Li et al., 2024) framework , we further extend their approach by building three types of within-admission causal graphs and modeling across-admission temporal dependencies, then propagate information via graph convolutions to obtain per-admission embeddings.

Constructing causal adjacency matrices. To capture causal and temporal dependencies, we construct a slice-wise causal graph at each admission t. Define the Gumbel-Softmax sampling operator $\mathrm{GS}\big(E;\tau\big) = \mathrm{softmax}\big((E + G)/\tau\big), G_{ij} \sim \mathrm{Gumbel}(0,1)$. For each admission t, we parameterize three intra-slice logit matrices $\mathbf{E}^{PP}_t, \mathbf{E}^{PC}_t, \mathbf{E}^{CC}_t \in \mathbb{R}^{ND_t\times ND_t}$ (with $ND_t = I_t + J_t$, represents the total number of nodes at time t) and obtain adjacency blocks by $\mathbf{S}^{PP}_t = \mathrm{GS}\big(\mathbf{E}^{PP}_t;\tau\big), \mathbf{S}^{PC}_t = \mathrm{GS}\big(\mathbf{E}^{PC}_t;\tau\big), \mathbf{S}^{CC}_t = \mathrm{GS}\big(\mathbf{E}^{CC}_t;\tau\big)$. We then form the block-structured intra-slice adjacency $\mathbf{A}_t = \begin{pmatrix} \mathbf{S}^{PP}_t & \mathbf{S}^{PC}_t \\ \mathbf{0} & \mathbf{S}^{CC}_t \end{pmatrix}$.

To ensure that each intra-slice adjacency $\mathbf{A}_t$ defines a directed acyclic graph, we adopt the continuous acyclicity penalty from NOTEARS (Zheng et al., 2018): $h(\mathbf{A}_t) = \mathrm{tr}\big(\exp(\mathbf{A}_t \circ \mathbf{A}_t)\big) - ND_t$, where $\mathbf{A}_t \circ \mathbf{A}_t$ is the Hadamard square of $\mathbf{A}_t$ and $ND_t = I_t + J_t$ is the total number of nodes. We add $\lambda_{\mathrm{acyc}}\ h(\mathbf{A}_t)$ to the overall loss, which vanishes if and only if $\mathbf{A}_t$ is acyclic. In parallel, to promote sparse and clinically interpretable graphs, we include an ℓ_1 penalty on the sampled adjacency blocks: $\lambda_{\ell_1} \sum_{x,y\in\{P,C\}} \big\|\mathbf{S}_t\big\|_1$. Similarly, inter-slice logits $\mathbf{E}^{\mathrm{inter}}_t$ produce $\mathbf{S}^{\mathrm{inter}}_t = \mathrm{GS}\big(\mathbf{E}^{\mathrm{inter}}_t;\tau\big)$, which models influences from slice t to slice $t+1$.

Graph Fusion and Temporal Message Passing. Let $\mathbf{X}_t \in \mathbb{R}^{d'\times ND_t}$ be the matrix whose rows are $\{\mathbf{x}^P_{t,i}\}_{i=1}^{I_t}$ followed by $\{\mathbf{x}^C_{t,j}\}_{j=1}^{J_t}$. Define two graph-convolution operators:

$$\mathrm{GC}_{\mathrm{intra}}(\mathbf{S},\mathbf{X};\mathbf{W}) = \mathrm{ReLU}(\mathbf{S}\,\mathbf{X}\,\mathbf{W}) + \mathbf{X},$$
$$\mathrm{GC}_{\mathrm{inter}}(\mathbf{S},\mathbf{X};\mathbf{W}) = \mathbf{S}\,\mathbf{X}\,\mathbf{W} + \mathbf{X}.$$

We perform:

$$\widetilde{\mathbf{X}}_t = \mathrm{GC}_{\mathrm{intra}}\big(\mathbf{A}_t,\ \mathbf{X}_t;\ \mathbf{W}^{(1)}\big),$$
$$\widehat{\mathbf{X}}_{t+1} = \mathrm{GC}_{\mathrm{inter}}\big(\mathbf{S}^{\mathrm{inter}}_t,\ \widetilde{\mathbf{X}}_t;\ \mathbf{W}^{(2)}\big).$$

Finally, we pool the updated intra-slice features and project: $\mathbf{Z}_t = \mathrm{MLP}\Big(\big[\mathrm{Mean}(\widetilde{\mathbf{X}}^P_t); \mathrm{Mean}(\widetilde{\mathbf{X}}^C_t)\big]\Big)$, where $\mathrm{Mean}(\cdot)$ averages row-vectors and $\widetilde{\mathbf{X}}^P_t, \widetilde{\mathbf{X}}^C_t$ denote the proposition- and code-node partitions of $\widetilde{\mathbf{X}}_t$. Concatenating $\{\mathbf{Z}_t\}_{t=1}^T$ yields the trajectory embedding $\mathbf{Z} \in \mathbb{R}^{T\times k}$, which feeds into the downstream prediction and calibration modules.

3.6 Stage 4: Prediction and Conformal Calibration

Prediction Let $\mathbf{Z} \in \mathbb{R}^{T\times k}$ be the trajectory embedding, where T is the number of admissions and k is the embedding dimension. Define $\mathbf{W}_o \in \mathbb{R}^{k\times L},\ \mathbf{b}_o \in \mathbb{R}^L,$, where $\mathbf{1} \in \mathbb{R}^T$)is all-ones vector. where L is the number of target labels. We compute the logit matrix $\mathbf{Y} = \mathbf{Z}\,\mathbf{W}_o + \mathbf{1}\,\mathbf{b}_o^\top \in \mathbb{R}^{T\times L}$, and apply the element-wise sigmoid $\sigma(x) = \frac{1}{1+e^{-x}} \implies \hat{\mathbf{P}} = \sigma(\mathbf{Y}) \in (0,1)^{T\times L}$, with $\hat{P}_{t,j} = \sigma(Y_{t,j})$.

Training loss We train by minimizing the focal loss over all admissions t and labels j:

$$\mathcal{L}_{\mathrm{FL}} = -\sum_{t=1}^{T}\sum_{j=1}^{L}\Big[\alpha\, y_{t,j}\,(1-\hat{P}_{t,j})^\gamma \log \hat{P}_{t,j} + (1-\alpha)\,(1-y_{t,j})\,\hat{P}^\gamma_{t,j}\log\big(1-\hat{P}_{t,j}\big)\Big],$$

where α and γ are the focal-loss balancing and focusing parameters. The full objective also includes the Stage 3 acyclicity penalty and sparsity regularization:

$$\mathcal{L} = \mathcal{L}_{\mathrm{FL}} + \lambda_{\mathrm{acyc}}\sum_{t=1}^{T} h\big(\mathbf{A}_t\big) + \lambda_{\ell_1}\sum_{t=1}^{T}\sum_{x,y\in\{p,c\}}\big\|\mathbf{S}^{xy}_t\big\|_1,$$

where $h(\mathbf{A}_t)$ is the NOTEARS acyclicity term introduced in Stage 3.

Conformal Calibration We partition the data into a proper training set and a calibration set $\mathcal{D}_{\text{calib}}$. For each calibration pair $(t, j) \in \mathcal{D}_{\text{calib}}$, define the nonconformity score $\alpha_{t,j} = 1 - \hat{P}_{t,j}$. Let $\{\alpha_{(1)} \leq \cdots \leq \alpha_{(N)}\}$ be these $N = |\mathcal{D}_{\text{calib}}|$ scores in ascending order, and set $\tau = \alpha_{\lceil (N+1)(1-\epsilon) \rceil}$. At test time for admission t, the conformal prediction set is

$$\hat{C}_t = \{ j : 1 - \hat{P}_{t,j} \leq \tau \},$$

which guarantees the marginal coverage $\Pr_{(t,j)\sim\text{test}}(y_{t,j} \in \hat{C}_t) \geq 1 - \epsilon$.

4 Experiments

To rigorously evaluate THCM-CAL, we conduct comprehensive experiments across several baseline clinical language models and benchmark its diagnostic-prediction performance against state-of-the-art methods.

4.1 Experimental Setup

Dataset We evaluate THCM-CAL on two standard EHR benchmarks. For MIMIC-III (Johnson et al., 2016), we include all patients with at least two hospitalizations (7192 patients, 11980 admissions), and for MIMIC-IV (Johnson et al., 2023), due to the risk of label leakage from discharge summaries in the original MIMIC-IV notes, we merge the original MIMIC-IV dataset with MIMIC-IV-Ext-BHC[1], which is a meticulously cleaned and standardized corpus of clinical notes (Labeled Clinical Notes Dataset for Hospital Course Summarization), and apply the same two hospitalization filter (17526 patients, 38346 admissions). We explicitly omit discharge summaries in the ICD task to prevent label leakage. Our objective is multi label ICD-9 prediction, where the first $N - 1$ visits are used to predict the full set of ICD-9 codes at visit N. Each cohort is split by patient into 70% train, 10% validation, and 20% test sets. In the Conformal Calibration stage we use the validation set for calibration.

4.2 Baselines

We compare THCM-CAL against the following state-of-the-art models for multi-label ICD prediction: **GatorTron** (Yang et al., 2022): a 345M-parameter transformer pretrained on large-scale EHR narratives to capture clinical language nuances. **DistilBioBERT** (Rohanian et al., 2024): a compact 66M-parameter BERT distilled on biomedical text, offering a lightweight yet effective encoder for clinical notes. **BioMedLM** (Bolton et al., 2024): a 2.7B-parameter language model trained on diverse biomedical corpora, providing rich domain knowledge for downstream classification. **CAML** (Mullenbach et al., 2018): employs per-label convolutional filters and attention to highlight text spans most relevant to each ICD code. **ZAGCNN** (Rios and Kavuluru, 2018): integrates the ICD code hierarchy via graph convolutions to enable zero- and few-shot code prediction. **Chet** (Lu et al., 2022): models multi-label diagnosis prediction as a sequence-to-set generation task using a transformer augmented with clinical event encodings. **DKEC** (Ge et al., 2024): incorporates external medical ontologies and domain rules into a multi-label classifier to enforce code consistency and improve rare code recall.

Parameter Setup Our methods use BERT for medical text embedding with hidden dimensions $d = 768$ for both proposition and ICD code representations. Our hierarchical temporal causal model employs a 2-layer architecture with residual connections and layer normalization. We train using Adam optimizer (learning rate 1×10^{-4}, weight decay 1×10^{-5}) with batch size 16 and dropout 0.1. Early stopping is applied with patience of 5 epochs over a maximum of 50 epochs. Each admission is represented with a maximum of 50 propositions and 30 ICD codes. The causal structure employs Gumbel-Softmax temperature annealing from 1.0 to 0.1 to enforce DAG constraints. Implementation uses PyTorch on NVIDIA A100 GPUs.

4.3 Main Results

Table 2 reports AUROC, Precision@10, Recall@10, Precision@20 and Recall@20 for the multi-label ICD-9 prediction task on MIMIC-III and MIMIC-IV. We highlight three key findings:

(1) On MIMIC-III, THCM-CAL achieves 30.02% Precision@10 and 24.04% Recall@10, representing gains of 5.50 and 5.22 percentage points over the strongest baseline (Chet: 24.52% / 18.82%). On MIMIC-IV, THCM-CAL attains 28.83% Precision@10 and 37.03% Recall@10, improvements of 6.24 and 7.20 points over the best baseline (DistilBioBERT: 22.59% / 29.83%). These results underscore the benefit

[1] https://physionet.org/content/labelled-notes-hospital-course/1.1.0/

Model	MIMIC-III					MIMIC-IV				
	AUROC	Precision@10	Recall@10	Precision@20	Recall@20	AUROC	Precision@10	Recall@10	Precision@20	Recall@20
CAML	93.43	20.61	18.15	14.98	25.74	95.49	20.61	27.38	14.03	35.52
ZAGCNN	89.88	17.50	15.51	12.61	21.63	93.96	21.10	28.13	14.33	36.17
GatorTron	**95.02***	20.53	17.72	15.03	25.66	95.90	22.87	30.19	15.47	38.81
DistilBioBERT	94.99	20.18	17.85	14.98	25.62	**95.99***	22.59	29.83	15.33	38.50
Chet	60.17	24.52	18.82	18.61	27.71	60.79	19.06	26.65	12.62	34.11
DKEC	94.47	20.64	17.92	14.55	24.47	94.34	20.03	26.50	13.61	34.59
BioMedLM	93.66	20.56	17.84	14.90	25.38	95.14	20.72	27.68	14.09	35.32
THCM-CAL	92.07	**30.02***	**24.04***	**21.47***	**33.16***	94.91	**28.83***	**37.03***	**18.66***	**46.04***
– w/o BERT	86.54	20.26	17.97	14.34	25.07	89.24	20.24	25.44	13.98	33.66
– w/o ICD	91.16	23.17	17.58	17.38	25.82	92.61	20.99	27.67	14.39	36.58
– w/o Proposition	91.97	26.07	20.55	19.27	29.25	93.87	22.43	28.99	15.44	38.49
– w/o ConCalib	91.84	26.03	20.68	19.29	29.61	94.33	25.36	32.63	16.94	42.03

Table 2: Comparison of multi-label ICD code prediction performance for representative baselines and THCM-CAL variants on MIMIC-III and MIMIC-IV. Metrics reported are AUROC, Precision@10, Recall@10, Precision@20, and Recall@20. Cells highlighted in pink denote our THCM-CAL method, and boldface values marked with "*" indicate the best performance across all methods, "w/o" means without.

of our narrative-to-code attention coupled with hierarchical causal modeling, which elevates the most critical diagnoses into the top-K predictions even under distribution shift.

(2) THCM-CAL delivers 33.16% Recall@20 on MIMIC-III (versus 25.74% for CAML and 24.47% for DKEC), and 46.04% Recall@20 on MIMIC-IV (versus 35.52% for CAML and 34.59% for DKEC), yielding improvements of 7.42 and 10.52 points, respectively. This boost is driven by our dynamic causal graph component, which explicitly captures cross-modal and temporal triggers, ensuring that less frequent but clinically important comorbidities are correctly retrieved.

(3) The gap in AUROC between THCM-CAL and other methods is relatively small, with THCM-CAL maintaining 92.07% on MIMIC-III and 94.91% on MIMIC-IV. This stability reflects the effect of our conformal calibration in adjusting confidence estimates for domain shifts, together with the causal structure that guards against spurious correlations.

Together, these results confirm that combining rich contextual embeddings, fine-grained proposition extraction, and explicit causal modeling can produce more effective code rankings and broader coverage of clinically relevant diagnoses.

4.4 Ablation Study

To quantify the contribution of each component, we perform an ablation study (Table 2) by removing: (i) contextual BERT embeddings in favor of simple indexing ("w/o BERT"), (ii) the ICD-side causal graph ("w/o ICD"), (iii) the proposition-side causal graph ("w/o Proposition"), and (iv) the conformal calibration module ("w/o ConCalib"). In every case, omitting a component degrades performance across both datasets. First, replacing BERT with one-hot indexing ("w/o BERT") yields the largest drop: on MIMIC-III, Precision@20 falls from 21.47% to 14.34%, and on MIMIC-IV from 18.66% to 13.98%. This underscores the necessity of rich contextual text representations for capturing fine-grained clinical nuances. Second, removing the ICD-side causal graph ("w/o ICD") reduces Recall@20 on MIMIC-III from 33.16% to 25.82% and on MIMIC-IV from 46.04% to 36.58%, indicating that explicit modeling of inter-code dependencies is critical to retrieving less frequent but clinically important comorbidities. Third, ablating the proposition-side graph ("w/o Proposition") lowers Precision@10 from 30.02% to 26.07% on MIMIC-III and Recall@10 from 37.03% to 28.99% on MIMIC-IV, demonstrating the value of narrative-to-code triggers for prioritizing key diagnoses. Finally, skipping conformal calibration ("w/o ConCalib") causes AUROC to drop modestly from 92.07% to 91.84% on MIMIC-III and from 94.91% to 94.33%

Dataset: MIMIC-III			
Configuration	Cov.↑	MIW↓	IE↑
THCM-CAL	0.9006	0.1087	9.1990
– w/o Proposition	0.8980	0.1137	8.7979
– w/o ICD	0.8937	0.1144	8.7412
Dataset: MIMIC-IV			
Configuration	Cov.↑	MIW↓	IE↑
THCM-CAL	0.8973	0.0820	12.1923
– w/o Proposition	0.8976	0.0975	10.2582
– w/o ICD	0.8991	0.0889	11.2478

Table 3: Conformal Prediction Metrics on MIMIC-III and MIMIC-IV (Cov.↑: higher is better; MIW↓: lower is better; IE↑: higher is better)

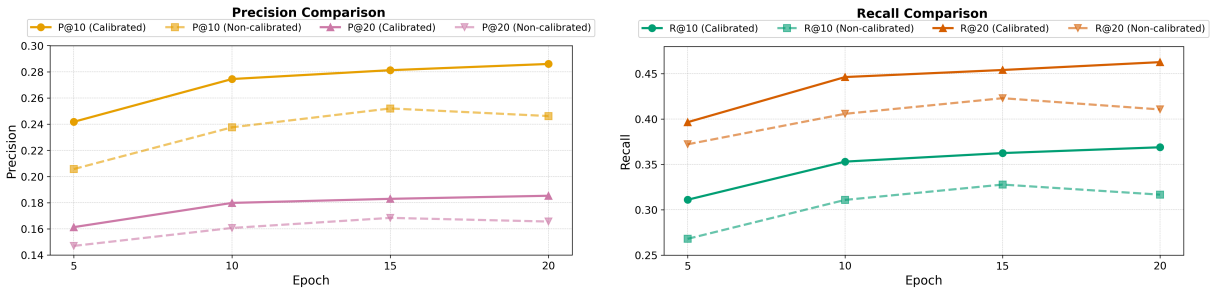

Figure 3: Comparison of metrics with and without split conformal calibration on MIMIC-IV.

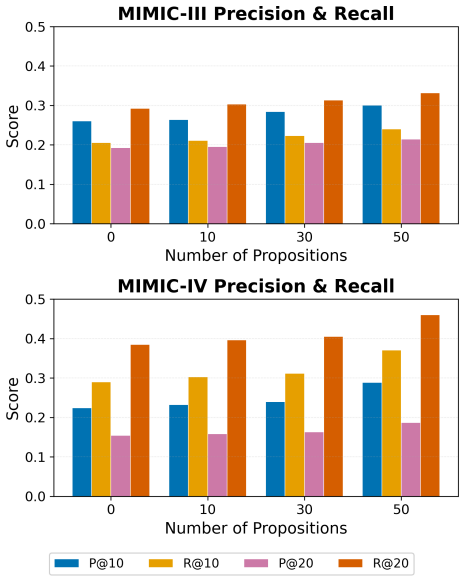

Figure 4: Effect of the number of extracted propositions for multi-label ICD-9 prediction.

on MIMIC-IV, confirming that score recalibration is important for maintaining stable discrimination under dataset shift.

4.5 Analysis on Conformal Prediction

We evaluate conformal prediction using three complementary metrics. Coverage (Cov) measures the fraction of true ICD codes captured by the prediction set. Mean interval width (MIW) quantifies the average size of these sets, with smaller widths indicating tighter intervals. Interval efficiency (IE), defined as the reciprocal of MIW, directly reflects interval compactness. Table 3 presents these metrics, averaged over all epochs, for our full THCM-CAL model and two ablations on both MIMIC-III and MIMIC-IV.

As shown in Table 3, THCM-CAL attains the highest coverage and the narrowest intervals on both datasets, leading to superior efficiency. On MIMIC-III the full model achieves a coverage of 0.9006 alongside a mean interval width of 0.1087 (IE = 9.1990), whereas removing the proposition side reduces coverage to 0.8980 and widens intervals to an average width of 0.1137, and omitting the ICD side further lowers coverage to 0.8937 with intervals of width 0.1144. Comparable patterns emerge on MIMIC-IV, where THCM-CAL delivers efficiency of 12.1923 compared to 10.2582 without the proposition side and 11.2478 without the ICD side. These results demonstrate that the proposition side is crucial for concentrating uncertainty into tighter sets and that the ICD side preserves nominal coverage under label sparsity. Neither ablation matches the performance of the complete model, confirming that both sides contribute to the final calibration quality.

4.6 Effect of Extracted Propositions

Figure 4 examines how varying the number of extracted propositions per admission affects prediction performance. As the proposition count increases from 0 to 50, both precision and recall at $K = 10$ and $K = 20$ steadily improve on MIMIC-III and MIMIC-IV. On MIMIC-III, Precision@10 rises from 0.26 to 0.30 and Recall@20 from 0.29 to 0.33; on MIMIC-IV, Precision@10 climbs from 0.22 to 0.29 and Recall@20 from 0.38 to 0.46. These gains reflect that each additional proposition injects new, fine-grained clinical observations into our causal graph, enabling more accurate identification of both the highest-priority codes (boosting Precision@10) and the broader set of relevant comorbidities (boosting Recall@20). With more narrative nodes, the inter-slice propagation and cross-modality trigger edges become better grounded, leading to smoother per-admission embeddings and reduced variance in performance across patients. Notably, the marginal benefit from adding propositions begins to plateau beyond 30 propositions, suggesting a point of diminishing returns where most salient information has already been captured. In practical, setting $K = 30$

propositions strikes a favorable balance between extraction cost and predictive performance.

4.7 Effect of Calibration

Figure 3 compares models trained with versus without our split conformal calibration on MIMIC-IV. Across all four metrics and at every checkpoint (epochs 5, 10, 15, 20), the calibrated model achieves higher scores than the non-calibrated baseline, illustrating that calibration not only provides valid confidence intervals but also leads to more accurate top-K code retrieval. The larger gap at early epochs indicates that calibration rapidly corrects over- and under-confidence in raw probabilities, yielding stable and reliable predictions as training progresses. Beyond accuracy, conformal calibration also ensures that the empirical coverage of each code's prediction set aligns with the desired level, even under non-stationary label co-occurrence patterns in MIMIC-IV. Importantly, calibration mitigates the impact of skewed code frequencies by adaptively adjusting thresholds per label, thereby reducing overprediction of common codes and underprediction of rare ones. These results confirm that split conformal calibration serves as a good post-hoc uncertainty quantifier.

5 Conlcusion

In this work, we introduced THCM-CAL, a unified Temporal–Hierarchical Causal Model with Conformal Calibration for Clinical Risk Prediction. By constructing a multi-slice causal graph that jointly captures intra-visit proposition sequencing, intra-visit cross-modality triggers, and inter-visit risk propagation, our model uncovers clinically meaningful relationships between narrative observations and diagnostic codes. We further adapt split conformal prediction to the multi-label setting, providing finite-sample guarantees on per-code. Extensive experiments on MIMIC-III and MIMIC-IV benchmarks demonstrate that THCM-CAL substantially outperforms state-of-the-art baselines, while offering calibrated uncertainty estimates. Ablation studies confirm that each causal edge type and the conformal calibration module are critical to performance gains.

Acknowledgement

We would like to thank the anonymous reviewers for their insightful comments and suggestions. This work was supported by the British Heart Foundation Manchester Research Excellence Award (RE/24/130017). We also acknowledge the CSF3 at the University of Manchester for providing GPU resources. Xin is supported by the UoM-CSC Joint Scholarship.

Limitations

While THCM-CAL demonstrates strong gains in accuracy, interpretability, and uncertainty calibration, it has several limitations.

- First, the reliance on large language models (e.g., GPT-3.5) for atomic proposition extraction introduces additional computational overhead and may propagate errors when the extractor misidentifies or omits clinically relevant statements.
- Second, although we validate on two MIMIC datasets mapped to ICD-9, extending THCM-CAL to richer coding systems (e.g., ICD-10) or to non-English clinical corpora will require careful adaptation of both the proposition extraction and code description modules. Addressing these challenges is our future work.

Despite these considerations, our approach provides a solid foundation for predicting ICD, and we believe that further refinements in these areas can further enhance its applicability.

References

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

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. 2024. Biomedlm: A 2.7b parameter language model trained on biomedical text. *CoRR*, abs/2403.18421.

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, Halifax, NS, Canada, August 13 - 17, 2017*, pages 787–795. ACM.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter F. Stewart. 2016. 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, December 5-10, 2016, Barcelona, Spain*, pages 3504–3512.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

R Scott Evans. 2016. Electronic health records: then, now, and in the future. *Yearbook of medical informatics*, 25(S 01):S48–S61.

Muhammad Hasan Ferdous, Uzma Hasan, and Md. Osman Gani. 2023. Cdans: Temporal causal discovery from autocorrelated and non-stationary time series data. In *Machine Learning for Healthcare Conference, MLHC 2023, 11-12 August 2023, New York, USA*, volume 219 of *Proceedings of Machine Learning Research*, pages 186–207. PMLR.

Xueren Ge, Abhishek Satpathy, Ronald D. Williams, John A. 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, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 12798–12813. Association for Computational Linguistics.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *CoRR*, abs/1904.05342.

Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

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.

Peiwen Li, Xin Wang, Zeyang Zhang, Yuan Meng, Fang Shen, Yue Li, Jialong Wang, Yang Li, and Wenwu Zhu. 2024. Realtcd: temporal causal discovery from interventional data with large language model. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4669–4677.

Chang Lu, Tian Han, and Yue Ning. 2022. Context-aware health event prediction via transition functions on dynamic disease graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4567–4574.

Chang Lu, Chandan K. Reddy, Prithwish Chakraborty, Samantha Kleinberg, and Yue Ning. 2021. 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, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3529–3535. ijcai.org.

Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. 2020. 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, Virtual Event, CA, USA, August 23-27, 2020*, pages 647–656. ACM.

Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. 2016. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1):26094.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1101–1111. Association for Computational Linguistics.

Tuan Nguyen, Thanh Trung Huynh, Minh Hieu Phan, Quoc Viet Hung Nguyen, and Phi Le Nguyen. 2024. CARER - clinical reasoning-enhanced representation for temporal health risk prediction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 10392–10407. Association for Computational Linguistics.

Anthony Rios and Ramakanth Kavuluru. 2018. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3132–3142. Association for Computational Linguistics.

Omid Rohanian, Mohammadmahdi Nouriborji, Hannah Jauncey, Samaneh Kouchaki, Farhad Nooralahzadeh, Lei A. Clifton, Laura Merson, and David A. Clifton. 2024. Lightweight transformers for clinical natural language processing. *Nat. Lang. Eng.*, 30(5):887–914.

Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. 2020. Temporal graph networks for deep learning on dynamic graphs. *arXiv preprint arXiv:2006.10637*.

Glenn Shafer and Vladimir Vovk. 2008. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3).

Supreeth P. Shashikumar, Gabriel Wardi, Atul Malhotra, and Shamim Nemati. 2021. Artificial intelligence sepsis prediction algorithm learns to say "i don't know". *npj Digit. Medicine*, 4.

Xinpeng Shen, Sisi Ma, Prashanthi Vemuri, M. Regina Castro, Pedro J. Caraballo, and György J. Simon. 2020. A novel method for causal structure discovery from EHR data, a demonstration on type-2 diabetes mellitus. *CoRR*, abs/2011.05489.

Yaoqian Sun, Lei Sang, Dan Wu, Shilin He, Yani Chen, Huilong Duan, Han Chen, and Xudong Lu. 2024. Enhanced icd-10 code assignment of clinical texts: A summarization-based approach. *Artificial Intelligence in Medicine*, 156:102967.

Ming-Lung Tsai, Kuan-Fu Chen, and Pei-Chun Chen. 2025. Harnessing electronic health records and artificial intelligence for enhanced cardiovascular risk prediction: A comprehensive review. *Journal of the American Heart Association*, 14(6):e036946.

Tony Wang, Tom Velez, Emilia Apostolova, Tim Tschampel, Thuy L. Ngo, and Joy Hardison. 2018. Semantically enhanced dynamic bayesian network for detecting sepsis mortality risk in ICU patients with infection. *CoRR*, abs/1806.10174.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Bowen Jin, May Dongmei Wang, Joyce C. Ho, and Carl Yang. 2024. RAM-EHR: retrieval augmentation meets clinical predictions on electronic health records. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024 - Short Papers, Bangkok, Thailand, August 11-16, 2024*, pages 754–765. Association for Computational Linguistics.

Ran Xu, Yue Yu, Chao Zhang, Mohammed K. Ali, Joyce C. Ho, and Carl Yang. 2022. Counterfactual and factual reasoning over hypergraphs for interpretable clinical predictions on EHR. In *Machine Learning for Health, ML4H 2022, 28 November 2022, New Orleans, Lousiana, USA & Virtual*, volume 193 of *Proceedings of Machine Learning Research*, pages 259–278. PMLR.

Yongxin Xu, Kai Yang, Chaohe Zhang, Peinie Zou, Zhiyuan Wang, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2023. Vecocare: Visit sequences-clinical notes joint learning for diagnosis prediction in healthcare data. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 4921–4929. ijcai.org.

Xi Yang, Nima M. Pournejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harle, Gloria P. Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, and Yonghui Wu. 2022. Gatortron: A large clinical language model to unlock

patient information from unstructured electronic health records. *CoRR*, abs/2203.03540.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. 2018. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31.

Anni Zhou, Raheem Beyah, and Rishikesan Kamaleswaran. 2025a. Neurosep-cp-lcb: A deep learning-based contextual multi-armed bandit algorithm with uncertainty quantification for early sepsis prediction. *CoRR*, abs/2503.16708.

Anni Zhou, Beyah Raheem, Rishikesan Kamaleswaran, and Yao Xie. 2025b. Sepsyn-olcp: An online learning-based framework for early sepsis prediction with uncertainty quantification using conformal prediction. *CoRR*, abs/2503.14663.

A Thresholding Strategies for F1 Score Calculation

In our experiments we do not report F1 scores because each baseline employs a different method to convert model scores into binary labels. This variation makes direct comparison of F1 values problematic. Instead, we provide AUROC, P@10, R@10, P@20 and R@20 which do not depend on a fixed threshold and thus allow a fair comparison. For reference we summarize below the thresholding strategies used in recent ICD code prediction works.

Specifically, **Chet** (Lu et al., 2022) counts the number of true labels for each instance and selects that many codes with the highest predicted scores. **CARER** (Nguyen et al., 2024) for multi-label tasks uses a dynamic threshold equal to the number of true labels and for binary tasks applies a fixed cutoff of 0.5. **DKEC** (Ge et al., 2024) applies a fixed cutoff of 0.5 and marks all codes with probability at least 0.5 as positive and if none meets this threshold it selects the single code with the highest probability. **CACHE** (Xu et al., 2022) uses the same 0.5 cutoff but allows this value to be adjusted at runtime. **RAM-EHR** (Xu et al., 2024) applies a 0.5 threshold independently for each code and reports per-label metrics. **HiTANet** (Luo et al., 2020) does not use any threshold and instead selects exactly one code by choosing the label with the maximum model score.

As shown above some models use dynamic thresholds matching each sample's true label count some use a fixed threshold of 0.5, and one uses maximum-probability selection. This heterogeneity in thresholding makes F1 scores difficult to compare on equal terms. We therefore omit F1 from our evaluations and rely on threshold-independent metrics to ensure a fair assessment of all methods.