Original Paper

A Disease Trajectory Representation Framework With Risk-Aware Pseudotime Analysis for Electronic Health Records: Application to Heart Failure Risk Stratification

Jhen-Li Huang[1], MS; Cing-Han Yang[1], MS; Tun-Wen Pai[1,2], PhD

[1]Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung, Taiwan

[2]Department of Computer Science and Information Engineering, National Taipei University of Technology, Taipei, Taiwan

Corresponding Author:
Tun-Wen Pai, PhD
Department of Computer Science and Information Engineering
National Taipei University of Technology
No. 1, Sec. 3, Zhongxiao E. Rd.
Taipei, 106344
Taiwan
Phone: 886 2 2771 2171 ext 4222
Email: twp@ntut.edu.tw

Abstract

Background: Heart failure develops through complex longitudinal interactions among multiple chronic diseases, making longitudinal disease trajectories an important source of clinical information. Longitudinal disease trajectories contain rich structural information that reflects disease evolution and comorbidity interactions. However, this information is often lost when electronic health record (EHR) data are summarized using conventional pooling strategies or sparse *International Classification of Diseases (ICD*) representations.

Objective: This study aimed to develop an interpretable trajectory representation that preserves structural characteristics of longitudinal disease evolution for heart failure risk stratification using EHRs.

Methods: Longitudinal *ICD* trajectories were represented using Node2Vec embeddings and summarized with information gain (IG)–weighted geometric features, including historical means, directional shifts, dispersion, and deviation relative to high-risk disease representations. The resulting maximum IG-weighted deviation geometric (mIGwDG) representation was evaluated using 10-fold cross-validation against conventional embedding aggregation strategies, sparse *ICD* one-hot encoding, and a long short-term memory (LSTM) model. To facilitate interpretation of longitudinal disease evolution, we further developed a risk-aware pseudotime framework to organize patient trajectories along a model-inferred risk continuum and used a framework combining IG with the positive pointwise mutual information (PPMI) framework to characterize clinically relevant comorbidity patterns.

Results: The proposed framework achieved an out-of-fold (OOF) area under the precision-recall curve (AUPRC) of 0.667 and an area under the curve (AUC) of 0.888, outperforming conventional embedding aggregation methods. Compared with sparse *ICD* one-hot encoding, the proposed framework reduced feature dimensionality by approximately 75% while maintaining comparable predictive performance. Combining mIGwDG with sparse *ICD* one-hot features further improved both logistic regression (LR) and extreme gradient boosting (XGBoost), suggesting that the proposed representation captures complementary longitudinal disease trajectory information. In addition, the risk-aware pseudotime framework organized heterogeneous patient trajectories along a model-inferred risk continuum, while IG×PPMI analysis highlighted clinically relevant comorbidity patterns that supported the interpretation of longitudinal disease organization.

Conclusions: The proposed framework provides a compact and interpretable representation of longitudinal disease trajectories while maintaining comparable predictive performance. By combining geometric trajectory characterization with risk-aware pseudotime visualization, the framework facilitates the representation of longitudinal EHRs, the visualization of disease evolution, and risk stratification.

(JMIR Med Inform 2026;14:e95929) doi: 10.2196/95929

XSL•FO
RenderX

KEYWORDS

heart failure; electronic health records; clinical risk identification; comorbidity trajectories; representation framework; interpretability

Introduction

Heart failure (HF), chronic kidney disease (CKD), and type 2 diabetes mellitus (T2DM) represent an escalating global health burden and frequently manifest as highly interrelated comorbid conditions. These conditions interact synergistically to form the cardiovascular-kidney-metabolic (CKM) syndrome [1,2]. The clinical course of HF is characterized by multisystem involvement, in which hypertension and metabolic disorders evolve into complex, time-varying comorbidity patterns [3]. Consequently, HF is increasingly regarded as a terminal convergence point of multiple chronic diseases, representing the culmination of long-term interactions that ultimately lead to systemic functional deterioration and decompensation [4,5]. Unlike conventional static risk models, clinical trajectories capture the longitudinal interplay between chronic risk accumulation and acute episodes across the care continuum in patients with HF [6-8]. Given that these evolving comorbidity structures are potent predictors of hospitalization and mortality, leveraging EHRs to develop risk stratification models that encapsulate such longitudinal dynamics remains important for clinical phenotyping [9,10].

In EHR-based research, *International Classification of Diseases* (*ICD*) codes are among the most standardized data elements for disease risk prediction due to their structured representation and universal adoption [11]. Conventional approaches typically encode *ICD* codes as independent features, using one-hot encoding or frequency-based representations such as term frequency-inverse document frequency (TF-IDF) [12]. These methods are computationally efficient and readily interpretable but implicitly assume diagnostic independence, thereby overlooking the rich comorbidity structures that underlie complex chronic diseases. Because clinical risk often emerges from interactions among multiple coexisting conditions rather than isolated diagnoses, recent studies have shifted toward distributed disease representations by embedding *ICD* codes into continuous vector spaces [13-15]. Graph embedding techniques, such as Node2Vec and DeepWalk, can learn richer representations by constructing disease co-occurrence networks and preserving semantic similarity [16,17]. More recently, graph neural network (GNN)–based approaches have integrated longitudinal EHR information with comorbidity connectivity and demonstrated improved disease risk prediction using heterogeneous graph attention mechanisms and joint modeling of temporal clinical records and disease networks [18]. Despite these advances, graph embedding techniques primarily capture global network topology but may overlook fine-grained comorbidity configurations and individualized disease evolution. Moreover, contemporary GNNs generally rely on increasingly sophisticated architectures, yet their learned representations remain difficult to interpret clinically. Consequently, developing embedding spaces that preserve both clinically meaningful comorbidity structures and individualized longitudinal characteristics remains an important research challenge [19].

Deep learning architectures, including recurrent neural networks (RNNs) and Transformer-based models, excel at capturing temporal patterns in longitudinal EHRs. Nevertheless, patient histories are typically encoded into latent representations or simplified via aggregation, obscuring the underlying comorbidity dynamics [20]. Furthermore, many representation learning approaches primarily capture static, population-level relationships and may struggle to distinguish acute exacerbations from long-term disease burden within heterogeneous trajectories [21,22]. Although attention-based representations improve predictive performance, they generally provide limited insight into the evolving geometry of diagnostic relationships, making it difficult to provide the structural interpretability needed for clinical decision support [23,24]. Developing models that integrate high predictive performance, explicit structural interpretability, and temporally informed representations of admission data therefore remains a central objective [10] and motivates the proposed risk-aware geometric trajectory framework.

In this study, we proposed a geometric trajectory representation framework for disease risk stratification to address these limitations. We modeled each patient's diagnostic history over 3 years as a dynamic trajectory within a latent *ICD* embedding space to stratify the risk of HF during their current hospitalization. Our approach was designed to capture target-centered relational structures and interdisease interactions. We derived geometric structural features to characterize the geometric characteristics and evolution of disease trajectories. This representation enabled predictive discrimination and structural interpretability, facilitating interpretation of clinically significant comorbidities in a transparent manner.

Methods

Data Sources and Preprocessing

This study used clinical diagnosis data from the Medical Information Mart for Intensive Care IV (MIMIC-IV; version 2.2) [25-27]. MIMIC-IV is a comprehensive, publicly available electronic health record (EHR) database developed by the Massachusetts Institute of Technology (MIT) and Beth Israel Deaconess Medical Center (BIDMC), encompassing inpatient medical records of patients with critical illness admitted to BIDMC between 2008 and 2019. Using individual hospital admissions as the primary unit of analysis, we collected hospitalization records for patients aged 20 years and older, organized chronologically by admission time. To construct a structured dataset capturing both cross-sectional and longitudinal features, historical hospitalization information from the preceding 3-year window was integrated for each index admission. Consequently, each record was treated as an independent sample, comprising a sequence of *ICD* codes from hospitalizations within the prior 3 years, and the corresponding study label.

The final cohort consisted of 71,212 patients and 226,992 hospital admissions, covering 1663 unique *International*

XSL•FO
RenderX

Classification of Diseases, 10th Revision, Clinical Modification (*ICD-10-CM*) diagnosis codes. Each admission contained a mean of 12.36 (SD 6.99) diagnosis codes, providing a heterogeneous representation of patients' comorbidity profiles. Baseline demographic and clinical characteristics were summarized in Table 1, including age, sex, length of stay, mortality, and admission statistics. To characterize the availability of longitudinal information for each index admission, we summarized the distribution of historical hospital visits within the 3-year observation window. Each index admission was associated with a mean of 4.82 (SD 8.14) historical hospital visits (median 2, IQR 1-5). Among all index admissions, 33.8% had only one historical visit available, whereas 66.2%, 28.3%, and 12.1% had at least 2, 5, and 10 historical visits, respectively. These statistics demonstrate that the cohort included both sparse and dense longitudinal trajectories, enabling evaluation of the proposed framework across varying amounts of historical information.

Table 1. Baseline demographic and clinical characteristics of the study cohort. The table summarizes patient demographics, hospitalization characteristics, and the distribution of historical hospital visits available for each index admission within the 3-year observation window.

Characteristic	All	Male	Female
Number of patients, n (%)	71,212	33,047 (46.4)	38,165 (53.6)
Number of admissions	226,992	111,376	115,616
Age (years), mean (SD)	58.98 (19.01)	60.44 (17.46)	57.71 (20.17)
Age (years), median (IQR)	61 (45-74)	62 (49-73.5)	59 (40-74)
Length of stay (days), mean (SD)	4.24 (6.92)	4.37 (7.24)	4.13 (6.60)
Mortality rate, n/N (%)	4,346/71,212 (6.1)	2,330/33,047 (7.1)	2,016/38,165 (5.3)
Historical admissions within the 3-year observation window, mean (SD)	4.82 (8.14)	5.40 (9.90)	4.26 (5.91)
Historical admissions within the 3-year observation window, median (IQR)	2.00 (1.00-5.00)	2.00 (1.00-6.00)	2.00 (1.00-5.00)

Since the data collection period of MIMIC-IV coincided with a transition in the US disease coding system, both the *ICD, Ninth Revision, Clinical Modification* (*ICD-9-CM*) and *ICD-10-CM* coding systems coexist within the dataset. To standardize the feature format, the primary preprocessing step involved integrating these 2 coding systems. We downloaded the General Equivalence Mappings (GEMs) from the official website of the Centers for Medicare & Medicaid Services (CMS) [28] to convert all mappable *ICD-9-CM* codes into their corresponding *ICD-10-CM* codes. Specifically, the mapping logic was applied as follows:

- If the mapping table indicated that an *ICD-9-CM* code had no corresponding *ICD-10-CM* equivalent, the code was omitted.
- If a single *ICD-9-CM* code mapped to multiple *ICD-10-CM* codes, all corresponding *ICD-10-CM* codes were recorded.
- If an *ICD-9-CM* code was missing from the official GEMs, we traced upward through the hierarchical tree to identify its most recent common ancestor (MRCA) node to obtain an alternative code.

Finally, to balance feature granularity and interpretability, all resulting *ICD-10-CM* codes were truncated to their first 3 characters. Furthermore, we excluded Z codes, which represent health care encounters unrelated to active diseases or injuries. Unless specified otherwise, all subsequent mentions of *ICD* codes in this study refer to the *ICD-10-CM* system.

Target-specific analyses were conducted for HF (*ICD-10-CM* code I50). For outcome labeling, hospitalizations were categorized as positive (label=1) if the target *ICD* code was present in either the primary or secondary hospital discharge diagnosis list; otherwise, they were labeled as negative (label=0). These discharge diagnoses were assigned by certified professional medical coders based on comprehensive clinical documentation using standardized *ICD-10-CM* coding practices. Although administrative diagnosis codes may be subject to coding inaccuracies, *ICD*-based outcome definitions have been widely adopted in large-scale EHR studies, thereby providing a standardized and reproducible approach for outcome ascertainment.

To evaluate the representational capacity of the model and prevent information leakage, a robust feature masking protocol was implemented. Specifically, the target code (I50) was removed from all diagnosis sequences prior to feature construction. Furthermore, to avoid reliance on direct clinical manifestations or established etiologies, we excluded highly correlated codes, including hypertensive heart disease (I11, I13), cardiomyopathies (I42, I43), complications and ill-defined descriptions of heart disease (I51), and pulmonary edema (J81). This approach encourages the model to learn latent risk representations from the underlying comorbidity structure and longitudinal disease history. The masked diagnosis sequences were used for all predictive feature construction, model training, and performance evaluation, including the risk stratification and architectural ablation analyses. In contrast, unmasked diagnostic records were used solely for post hoc interpretation and visualization, including the IG and positive pointwise mutual information (PPMI) comorbidity analysis and risk-aware pseudotime visualization. These unmasked records were analytically separated from the predictive modeling pipeline and not used for feature construction, model training, or performance evaluation.

XSL•FO
RenderX

Only historical *ICD* records preceding the index admission were used as input features, whereas diagnosis codes from the index admission were reserved strictly for outcome definition. This design ensured that risk stratification relied solely on historical medical information available prior to the index admission while reflecting a clinically realistic setting. Patients with at least one prior hospitalization within the 3-year look-back period were included in the trajectory construction. Patients without any historical hospitalizations during this period lacked a longitudinal trajectory and were therefore excluded from the trajectory-based classification analysis. For patients with a single admission, trajectory-derived features were set to zero, while pooled representations remained available. Consequently, these admissions were represented by a single-state embedding rather than being excluded from the analysis.

Ethical Considerations

This study was a secondary analysis of a publicly available, deidentified MIMIC-IV database. The creation and maintenance of the MIMIC-IV database were approved by the Institutional Review Boards (IRBs) at MIT and BIDMC. Access to the database was granted to the author (JLH) after successful completion of the "Data or Specimens Only Research" training (CITI Program, Record ID: 61470359) and the execution of a formal data use agreement. All patient records contained in the MIMIC-IV database were rigorously deidentified prior to release in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor regulations. Consequently, the requirement for individual patient informed consent was waived by the IRBs. There was no direct contact with human subjects, and participants received no compensation. Furthermore, no identifying information or clinical images of individual participants were included in this manuscript or its supplementary materials.

Information Gain for ICD Feature Discrimination

In machine learning, information gain (IG) is a fundamental metric for quantifying a feature's contribution to class discrimination and is based on information-theoretic entropy [29]. In this study, IG was employed to assess the discriminative power of individual *ICD* codes for target disease identification, measuring the reduction in uncertainty regarding the outcome given the presence or absence of a specific code. For each *ICD* code in the training dataset, the IG value was calculated using equation 1:

$$IG(ICD) = H(Y) - H(Y|ICD)$$

where $H(Y)$ denotes the marginal entropy of the target outcome and $H(Y|ICD)$ represents the conditional entropy. This metric captures the extent to which each *ICD* code facilitates effective classification, where higher IG values signify superior discriminative capacity. Since the target outcome was HF (*ICD* code I50), the resulting IG scores specifically quantified the contribution of each *ICD* code to the identification of HF. These disease-specific IG scores were subsequently used in a weighted pooling mechanism to emphasize *ICD* diagnosis codes with higher clinical relevance for identifying the target disease.

PPMI-Based ICD Co-Occurrence Analysis

To capture the underlying comorbidity structures within hospitalization records, we constructed an *ICD* code co-occurrence network from the training data using pointwise mutual information (PMI) [30]. PMI is widely used in machine learning and natural language processing (NLP) to quantify co-occurrence strength between discrete entities. In our framework, each hospitalization was treated as a "document" comprising all assigned *ICD* codes. A co-occurrence event was defined by the concurrent appearance of 2 *ICD* codes within the same admission. Based on the marginal probabilities of individual *ICD* codes and their joint probabilities estimated from the training dataset, the PMI between a pair of *ICD* codes (i, j) was defined as equation 2:

$$PMI(i,j) = \log\frac{P(i,j)}{P(i)P(j)}$$

where $P(i)$ denotes the occurrence probability of code i, and $P(i,j)$ represents their joint probability. To reduce noise from negative associations, which may lack clinical interpretability and be mathematically unstable, we excluded negative PMI values. Instead, we adopted PPMI, as defined in equation 3, to retain only diagnostically meaningful positive associations between *ICD* codes. Finally, *ICD* codes were modeled as nodes, with PPMI values serving as edge weights, thereby constructing a weighted, undirected *ICD* co-occurrence graph.

$$PPMI(i,j)=\max(PMI(i,j),0)$$

Following the construction of the *ICD* co-occurrence graph, the Node2Vec algorithm was implemented to learn complex relational patterns among *ICD* nodes [16]. Node2Vec employs second-order random walks that interpolate between breadth-first search (BFS) and depth-first search (DFS) through the return parameter (p) and in-out parameter (q). In this study, the default values (p=q=1) were adopted, providing an unbiased exploration of the disease network. Accordingly, no explicit bias toward either BFS- or DFS-like exploration was imposed during random walk generation. Through this approach, the high-dimensional comorbidity graph was mapped into a low-dimensional continuous vector space. This ensured that *ICD* codes with strong co-occurrence patterns, or those sharing similar structural roles within the network despite lacking direct connections, were positioned in proximity within the embedding space. Node2Vec was implemented using an embedding dimension of 64, a walk length of 10, and 30 random walks per node. The complete hyperparameter configuration was summarized in Multimedia Appendix 1.

Longitudinal Diagnosis History Representation Architecture

Overview

To encode the irregularly sampled, longitudinal EHR sequence into a dense, time-aware feature space $X_{admission} \in \mathbb{R}^d$, we organized feature construction into a 4-stage hierarchical architecture. The hierarchy progressively integrated admission-level disease representations, IG-guided risk profiling, temporal weighting, and trajectory geometry to capture

XSL•FO
RenderX

complementary structural characteristics of disease evolution within a unified representation.

For an index admission, their longitudinal medical history within a 3-year observation window was defined as a sequence of historical hospitalizations $P=\{V_1,V_2,\ldots V_m\}$, where m represented the total count of preceding admissions. Each admission V_m contained a subset of discrete diagnosis codes mapped into the embedding space via a pre-trained Node2Vec model, transforming each code $c \in V_m$ into a dense vector $e(c) \in \mathbb{R}^d$ (where d=64).

Admission Representation Layer

The foundational layer transformed discrete diagnosis codes within a single medical encounter into an admission-level disease representation. Diagnosis embeddings within each admission were aggregated to construct baseline representations that characterized the semantic composition of individual clinical encounters.

Mean Pooling (Centroid Representation)

For each admission $V_m \in \mathbb{R}^d$, mean pooling calculated the average position of all diagnosis embeddings within the latent *ICD* embedding space, yielding a geometric centroid that represented the "average disease state" of that hospitalization within the semantic space (equation 4):

$$V_m^{mean} = \frac{1}{M}\sum_{c \in V_m} e(c)\text{, where } M = |V_m|$$

where *M* denotes the set of diagnosis codes in admission *m*, and *e(c)*represents the embedding vector of diagnosis code *c*. The resulting vector V_m^{mean} represents the central position of the admission diagnoses profile within the embedding space, effectively characterizing the "average disease state" of that hospitalization within the disease semantic space.

Max Pooling (Acute Signal Representation)

Max pooling effectively captured acute diagnostic signals and clinical deterioration indicators (eg, severe complications) by preserving the most prominent latent features across diagnosis embeddings by selecting the maximum value along each embedding dimension (equation 5):

$$V_m^{max} = \max_{c \in V_m} e(c)$$

The resulting vector V_m^{max} represented a vector that highlighted dominant latent features of the admission diagnoses within the embedding space.

Risk Profiling Layer

To capture heterogeneous diagnostic contributions within each admission, this layer enhanced admission-level representations through IG-guided weighting and deviation-based geometric characterization:

Entropy-Based IG Gating

We introduced a max IG-weighted pooling strategy to capture high-impact diagnostic signals while reducing the influence of common but less informative diagnoses. IG values *IG*(*c*) were derived from the training corpus and normalized per training fold to prevent information leakage (equation 6):

$$V_m^{ig} = \max_{c \in V_m}(IG(c)e(c))$$

Compared with conventional mean pooling, which summarizes the overall disease composition of an admission, IG-weighted pooling generates a risk-informed representation by assigning greater contribution to diagnostically informative features. The resulting vector V_m^{ig} captures admission-level disease characteristics weighted according to their model-derived importance.

Pathological Risk Deviation Vector

To characterize the difference between general disease composition and high-information diagnostic patterns, we introduced a deviation vector based on geometric displacement within the *ICD* embedding space.

For each admission V_m within the 3-year history path (PATH), the mean pooling vector V_m^{mean} was used to summarize the overall admission-level disease profile, while the max IG-weighted vector V_m^{ig} designed to accentuate highly informative features while suppressing common, nondiscriminative diagnostic codes. The admission-level deviation vector V_m^{diff} (Equation 7) quantified the geometric displacement within a specific diagnostic combination:

$$V_m^{diff}=V_m^{ig}-V_m^{mean}$$

(In naive ablation models, this is substituted as $V_m^{diff}=V_m^{max}-V_m^{mean}$.)

The resulting vector quantified the magnitude and direction of displacement between the general disease representation and the information-weighted representation. This geometric difference characterized how diagnostic patterns within an admission shifted away from the overall disease context toward regions associated with higher model-estimated risk.

Temporal Aggregation Layer

To incorporate the temporal heterogeneity of longitudinal EHR records, this layer aggregated historical admission-level representations according to their temporal proximity to the index admission. Unlike conventional static aggregation approaches that assigned identical importance to all historical encounters, the proposed temporal weighting strategy emphasized more recent clinical information while retaining the contribution of earlier disease history (shown in equation 8). For an index admission occurring t_m days prior to the prediction target point, a normalized linearly decayed weight w_m was calculated as (equation 8):

$$w_m = \frac{\max(0.01, 1-\frac{t_m}{window})}{\sum_{k=1}^{M}\max(0.01, 1-\frac{t_k}{window})}$$

where w_m represents the relative temporal contribution of admission *m* within the predefined 3-year observation window.

Using these temporal weights, the admission-level representations were aggregated into historical path-level indicators. To establish internal baseline comparisons, we

XSL•FO
RenderX

constructed conventional history-level pooling representations together with the proposed deviation-based trajectory features.

- Baseline history weighted mean vector ($P^{wm} \in \mathbb{R}^d$): the historical weighted mean representation reflects the overall disease composition across all previous admissions within the 3-year observation window (equation 9):

$$P^{wm} = \sum_{m=1}^{M} w_m V_m^{mean} \quad \text{E}$$

In naive internal ablation models, w_m defaults to *1/M*, reducing the calculation to a simple unweighted historical average.

- Baseline history max vector ($P^{max} \in \mathbb{R}^d$): this representation preserves the maximum latent, absolute risk signal encountered at any point across the entire 3-year observation period, which captures prominent diagnostic signals that may be underrepresented by averaging operations (equation 10),

$$P^{max} = \max_{m=1\ldots M}(V_m^{max})$$

The admission-level deviation vector V_m^{diff} reflects the positional transition of comorbidity combinations within the disease graph, capturing the extent to which a patient's state deviated from a baseline profile toward high-risk phenotypes. To translate these discrete admission-level transitions into continuous, longitudinal patient-level descriptors, the sequence of deviation vectors was systematically compressed into the final feature space through 2 distinct perspectives:

- Cumulative deviation burden: the cumulative deviation descriptor ($P^{diffsum} \in \mathbb{R}^d$) was calculated by summing deviation vectors across chronological admissions (equation 11),

$$P^{diffsum} = \sum_{m=1}^{M} V_m^{diff}$$

It served as a measure of sustained disease burden. This feature represented the accumulated magnitude and direction of diagnostic shifts across the longitudinal disease history.

- Recency-weighted mean deviation: the recency-weighted deviation representation ($P^{diffwm} \in \mathbb{R}^d$), shown in equation 12, incorporated the temporal weights w_m derived from equation 8:

$$P^{diffwm} = \sum_{m=1}^{M} w_m V_m^{diff}$$

By assigning greater importance to recent admissions, this feature captured temporally prioritized deviations within the patient's longitudinal trajectory while preserving information from earlier encounters.

Trajectory Geometry Layer

The final layer characterized longitudinal disease trajectories by quantifying geometric changes, dispersion, and directional alignment of multiadmission representations within the embedding space.

Intervisit Trajectory Transition Features

To capture the displacement between consecutive clinical encounters, we calculated displacement vectors, as shown in equation 13, between successive admission representations for patients with $M \geq 2$:

$$\Delta_m = V_{(m+1)}^{mean} - V_m^{mean}$$

Where Δ_m represents the embedding displacement between 2 consecutive admissions.

Two trajectory transition descriptors were derived:

- Cumulative intervisit trajectory transition ($X_{trai_sum} \in \mathbb{R}^d$): the absolute displacement vectors across consecutive admissions were summed to quantify the overall magnitude of longitudinal representation changes.
- Mean trajectory transition ($X_{traj_mean} \in \mathbb{R}^d$): the average displacement vector across consecutive admissions was calculated to summarize the typical direction and magnitude of longitudinal representation changes.

Embedding Space Geometric Variance and Dispersion

To characterize the structural variability of longitudinal disease representations, we quantified the spatial dispersion of admission-level embeddings around each patient's historical centroid.

First, admission-level representations were aggregated to obtain the patient-specific history centroid, denoted as:

$$\alpha = \frac{1}{M}\sum_{m=1}^{M} V_m^{mean}$$

To quantify the structural variability of a patient's disease trajectory, we computed the Euclidean distance between the max IG-weighted embedding (or max-pooled embedding in ablation variants) of each visit V_m^{ig} and their personal history centroid α. The overall trajectory dispersion, denoted as geometric variance ($s^2_{dist} \in \mathbb{R}^1$), was formulated as equation 14:

$$s^2_{dist} = \frac{1}{M}\sum_{m=1}^{M} \|V_m^{ig} - \alpha\|_2^2$$

This metric quantifies the extent to which a patient's longitudinal disease representations are distributed within the embedding space.

Trajectory Angular Coherence and Target-Guided Alignment

Cosine similarity was used to characterize directional relationships among longitudinal disease representations, providing complementary information beyond Euclidean distance.

- Trajectory angular coherence ($\mu_{cos} \in \mathbb{R}^1$, $\sigma_{cos}^2 \in \mathbb{R}^1$): to characterize the consistency and coordination of the patient's internal clinical evolution, we computed the pairwise cosine similarity across all unique combinations of admission-level vectors within the patient's own history (shown in equation 15),

XSL•FO
RenderX

$$s_{dist}^2 = \frac{1}{M}\sum_{m=1}^{M}\|V_m^{ig} - \alpha\|_2^2$$

In naive ablation models, V_m^{ig} is substituted as V_m^{max}. The mean (μ_{cos}) and variance ($\sigma_{\cos}^2$) of pairwise cosine similarities were obtained from the upper triangular elements of the similarity matrix, excluding self-comparisons. These features summarize the directional consistency and variability of longitudinal disease representations.

- Target-guided alignment ($\mu_{cos_pos} \in \mathbb{R}^1$, $\sigma_{\cos}^2 \in \mathbb{R}^1$): to quantify the geometric relationship between patient trajectories and high-risk disease representations, a positive-reference centroid ($\mu_{pos} \in \mathbb{R}^d$) was derived by aggregating the baseline history mean vectors (P^{mean}) of all positive training samples (label=1). The cosine similarity between each patient's admission-level representation $V_m^{mean}$and the high-risk reference vector was calculated as equation 16,

$$sim_{target}(m) = \frac{V_m^{ig}\mu_{pos}}{\|V_m^{ig}\|_2\|\mu_{pos}\|_2}$$

In naive ablation models, V_m^{ig} is substituted as V_m^{mean}. The resulting mean (μ_{cos_pos}) and variance ($\sigma_{\cos_pos}^2$) describe the degree and variability of alignment between individual trajectories and the high-risk reference space.

For patients with a single admission ($M \leq 1$), trajectory transition and dispersion features could not be calculated due to insufficient longitudinal observations. These features were assigned zero values, while alignment-based features were assigned neutral values to maintain consistent feature dimensionality.

Feature Integration and Classification

For each index admission, all geometric features derived from the corresponding longitudinal patient history were concatenated into a 389-dimensional feature vector that served as the input to the downstream classifiers. This architecture preserves both the longitudinal disease context and the immediate diagnostic signals of the index admission, thereby enhancing the model's ability to identify clinical risk states. To prevent data leakage and overestimation of model performance resulting from recurrent patient admissions, we implemented a patient-stratified data splitting strategy during 10-fold cross-validation. Under this framework, all admissions associated with a single patient were assigned to the same fold, ensuring complete independence between training and testing sets. We employed an extreme gradient boosting (XGBoost) classifier for risk identification and used logistic regression (LR) as a baseline to compare XGBoost with a conventional linear approach. The detailed hyperparameter settings for the XGBoost classifier and the LR baseline were summarized in Multimedia Appendix 2. These hyperparameters were determined through preliminary experiments and remained fixed throughout all cross-validation folds, except for the fold-specific scale_pos_weight parameter, which was recalculated according to the class distribution within each training fold. To mitigate class imbalance (18.83% positive prevalence), a cost-sensitive learning strategy was adopted instead of resampling, thereby preserving the original class distribution during model training. Specifically, for the XGBoost model, the scale_pos_weight parameter was determined dynamically for each training fold according to the ratio of negative to positive samples in the corresponding training data. For the LR model, the class_weight parameter was set to balanced. This approach penalized the misclassification of the minority class during loss optimization. Finally, model performance was evaluated using the area under the curve (AUC) and the area under the precision-recall curve (AUPRC) [31]. In addition, 95% CIs for the AUC and AUPRC were estimated using patient-level bootstrap resampling with replacement on the independent test set over 1000 bootstrap iterations. The resulting empirical bootstrap distributions were used to derive the corresponding 95% CIs.

Baseline Models

To benchmark the proposed maximum IG-weighted deviation geometric (mIGwDG) framework, we implemented 3 baseline approaches under the same training, validation, and evaluation protocol. First, to model longitudinal disease trajectories, we implemented a standard long short-term memory (LSTM) network using the same Node2Vec-based admission embeddings as sequential inputs. LSTM is a widely adopted deep learning architecture for modeling longitudinal EHR data and disease progression [32]. Second, we implemented LR and XGBoost using sparse *ICD* one-hot encoded features representing all historical diagnosis codes. These models are widely used baselines for structured EHR-based risk prediction and allow evaluation of the proposed representation without feature compression [33]. Finally, sparse *ICD* one-hot encoded features were concatenated with the proposed mIGwDG representation and used as input to both LR and XGBoost to evaluate whether the learned trajectory representation provided complementary predictive information beyond the original sparse *ICD* one-hot encoded features by incorporating longitudinal and geometric disease trajectory information.

Risk-Based Visualization and Risk-Aware Pseudotime Stratification

This study presented a risk-stratified visualization framework to characterize disease progression patterns based on model-inferred risk estimates. Throughout this study, the term "pseudotime" denoted a risk-ordered representation of index admissions derived from calibrated model-predicted probabilities, rather than an optimization-based trajectory inference method, as commonly used in single-cell transcriptomics. Out-of-fold (OOF) risk probabilities were obtained through 10-fold cross-validation to ensure unbiased risk estimation and mitigate data leakage. Because the raw probability outputs of tree-based models such as XGBoost were not necessarily well calibrated, probability calibration was incorporated into the predictive pipeline before risk stratification. Specifically, Platt scaling was applied within each cross-validation iteration by fitting an LR model to the raw OOF probabilities from the training folds. This fold-wise calibration strategy ensured that probability calibration remained independent of the validation data, thereby preventing

XSL•FO
RenderX

information leakage. Calibration performance was evaluated using reliability diagrams (calibration curves) across 10 decile bins, as shown in Multimedia Appendix 3.

The calibrated probabilities were subsequently used for pseudotime risk stratification. To avoid reliance on arbitrary probability thresholds, we employed a quantile-based binning to divide index admissions into 5 risk strata (Very low, Low, Medium, High, and Very high risk) according to the calibrated probabilities. Formally, the risk stage of the *i*-th index admission was defined using equation 17,

$$R_i = Q(\hat{p}_i)$$

Where

$$\hat{p}_i$$

denotes the calibrated OOF probability, and $Q(\cdot)$ represents the quintile mapping function that partitions the cohort into 5 equal-sized risk strata.

To characterize the accumulation of comorbidities across the inferred risk spectrum, we computed the prevalence of each *ICD* code within every risk stratum. For *ICD* code *j* in risk stage *k*, the prevalence was calculated using equation 18,

$$P_{jk} = \frac{n_{jk}}{N_k}$$

where n_{jk} is the number of index admissions in risk stage k containing *ICD* code *j*, with each *ICD* counted at most once per admission, and N_k is the total number of index admissions assigned to that risk stage.

To reveal the sequential emergence of comorbidities along the risk spectrum, *ICD* codes were ordered according to their weighted average risk stage (evolution score) and shown in equation 19,

$$S_j = \frac{\sum_{k=1}^{5} w_k P_{jk}}{\sum_{k=1}^{5} P_{jk}}$$

Where $w_k \in \{0,1,2,3,4\}$ denotes the ordinal index of the 5 risk strata from Very low to Very high. *ICD* codes were subsequently sorted in ascending order of S_j, thereby arranging diseases according to their predominant stage of occurrence.

We prioritized the top 25-IG×PPMI-weighted comorbid *ICD* codes for visualization. Finally, this prevalence matrix was visualized as a heatmap highlighting the gradual and structured accumulation of clinically relevant comorbidities along the latent risk spectrum. The resulting heatmap organized comorbidities along a pseudo-temporal risk axis that reflects increasing model-inferred risk rather than chronological time. This framework was intended solely as a descriptive, hypothesis-generating representation of disease evolution based on model-inferred risk and does not imply chronological progression, direct temporal causality, or validated clinical disease staging.

Results

Dataset and Validation Strategy

We evaluated our framework on 226,992 hospital admissions with HF (*ICD-10*: I50) as the target condition. Notably, the positive class (Label=1) accounted for 18.83% of the data, representing a significant class imbalance. To ensure the integrity of our findings and prevent data leakage, we implemented a patient-stratified 10-fold cross-validation strategy. This approach guarantees that all encounters belonging to a specific patient remain in either the training or the test set, never both. Furthermore, within each fold, both the disease co-occurrence graphs and the IG-based importance weights were derived exclusively from the training partition. This protocol ensures that the learned representations are validated on entirely independent data, preserving the methodological integrity of our risk stratification evaluation.

Admission-Level Risk Stratification and Architectural Ablation

We first evaluated admission-level risk stratification performance using 10-fold cross-validation across different *ICD* embedding aggregation strategies. As shown in Table 2, XGBoost consistently outperformed LR in both AUC and AUPRC. For conventional static baselines, XGBoost achieved AUPRCs of 0.627 (SD 0.011; mean pooling) and 0.625 (SD 0.017; maximum pooling), compared with a mean of 0.538 (SD 0.013) and a mean of 0.527 (SD 0.029) for LR, respectively. In addition to the proposed framework and its internal ablation variants, 3 commonly used baseline approaches (LR, XGBoost with *ICD* one-hot encoding, and LSTM) were evaluated under the same experimental protocol.

To further quantify the contribution of each feature module, we conducted an additional feature-module ablation study by progressively incorporating risk deviation, intervisit trajectory transition, and geometric variance/dispersion features (Multimedia Appendix 4). Starting from the risk deviation representation, trajectory change and geometric features resulted in only marginal differences in predictive performance, with changes generally within the range of cross-validation variability. In contrast, IG weighting produced a more pronounced improvement compared with maximum-based aggregation. The complete mIGwDG representation achieved the best overall performance, although the incremental contributions of the individual trajectory-change and geometric modules were not significantly obtained.

XSL•FO
RenderX

Table 2. Global predictive performance and OOF[a] stability at the admission level under different feature designs (N=226,992 admissions).

Model	10-fold cross-validation		Admission level OOF	
	AUC[b]	AUPRC[c]	AUC (95% CI)	AUPRC (95% CI)
Deep learning baseline				
LSTM[d]	0.877 (0.003)	0.642 (0.011)	0.877 (0.873-0.880)	0.639 (0.628-0.651)
Embedding approach				
LR[e]-mean	0.856 (0.005)	0.538 (0.013)	0.856 (0.853-0.860)	0.538 (0.528-0.548)
LR-max	0.842 (0.010)	0.527 (0.029)	0.842 (0.838-0.845)	0.527 (0.516-0.537)
LR-maxDG[f]	0.881 (0.003)	0.641 (0.013)	0.881 (0.877-0.884)	0.641 (0.629-0.653)
LR-mIGwDG[g]	0.881 (0.004)	0.647 (0.013)	0.881 (0.878-0.885)	0.646 (0.635-0.658)
XGBoost[h]-mean	0.875 (0.003)	0.627 (0.011)	0.875 (0.871-0.879)	0.627 (0.617-0.639)
XGBoost-max	0.874 (0.005)	0.625 (0.017)	0.874 (0.871-0.877)	0.624 (0.615-0.636)
XGBoost-maxDG	0.886 (0.003)	0.660 (0.014)	0.886 (0.883-0.889)	0.660 (0.650-0.670)
XGBoost-mIGwDG	0.888 (0.003)	0.667 (0.016)	0.888 (0.885-0.891)	0.667 (0.657-0.677)
One-hot encoding				
LR	0.883 (0.004)	0.665 (0.016)	0.883 (0.879-0.886)	0.665 (0.655-0.676)
XGBoost	0.886 (0.004)	0.666 (0.016)	0.886 (0.883-0.889)	0.667 (0.656-0.677)
One-hot and embedding approach (mIGwDG)				
LR	0.888 (0.003)	0.674 (0.014)	0.888 (0.884-0.891)	0.674 (0.665-0.684)
XGBoost	0.892 (0.003)	0.679 (0.017)	0.892 (0.889-0.895)	0.679 (0.669-0.689)

[a]OOF: out-of-fold.

[b]AUC: area under the curve.

[c]AUPRC: area under the precision-recall curve.

[d]LSTM: long short-term memory.

[e]LR: logistic regression.

[f]maxDG: maximum-pooling deviation geometric.

[g]mIGwDG: maximum information gain-weighted deviation geometric.

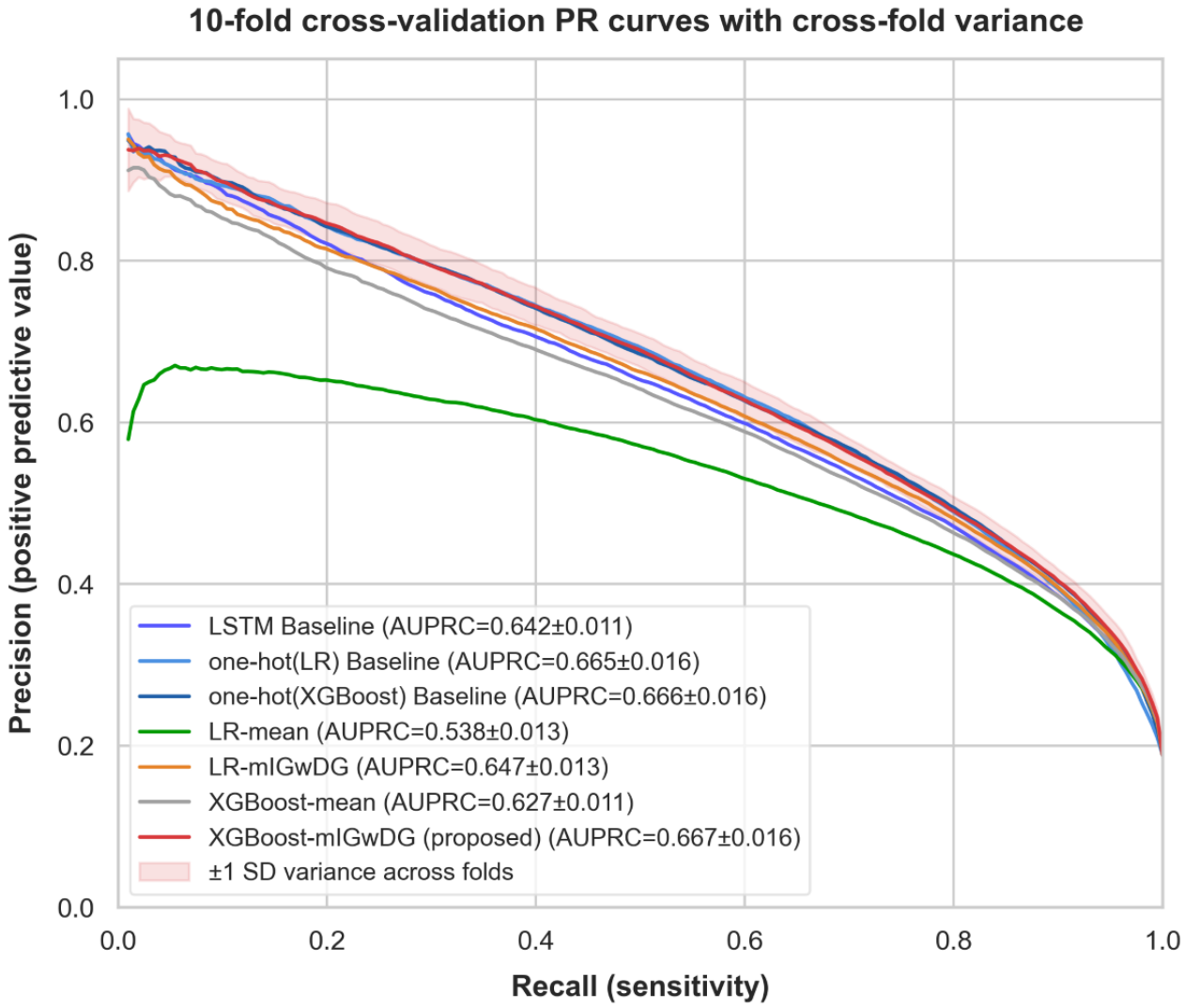
[h]XGBoost: extreme gradient boosting.

The complete mIGwDG framework achieved an AUC of 0.888 (SD 0.003) and an AUPRC of 0.667 (SD 0.016) under XGBoost in 10-fold cross-validation while maintaining comparable predictive performance to sparse *ICD* one-hot encoding with substantially lower feature dimensionality. Under LR, mIGwDG achieved an AUC of 0.881 (SD 0.004) and an AUPRC of 0.647 (SD 0.013). Importantly, the proposed hierarchical representation substantially narrowed the performance gap between the linear and nonlinear classifiers, suggesting that it transforms longitudinal disease trajectories into a more discriminative feature space that can be effectively exploited even by a linear classifier. Figure 1 presents precision-recall curves across the evaluated models. Solid lines indicate the mean precision at each recall level across the 10 folds, and the shaded area represents the standard deviation of the proposed XGBoost-mIGwDG model. Mean AUPRC values and cross-validation variability are reported in the figure legend.

To ensure an unbiased evaluation, we generated OOF risk predictions using 10-fold cross-validation and calculated AUPRC based on aggregated OOF predictions. The proposed mIGwDG representation combined with XGBoost achieved an aggregated OOF AUPRC of 0.677 (95% CI 0.657-0.677). The *ICD* one-hot XGBoost baseline achieved a comparable cross-validated AUPRC of 0.666 (SD 0.016). Although the one-hot representation retained the full *ICD* feature space, it resulted in a substantially higher-dimensional feature representation. In contrast, mIGwDG achieved similar predictive performance using a compact trajectory-based representation.

XSL•FO
RenderX

Figure 1. Ten-fold cross-validation precision-recall (PR) curves with cross-fold variance. AUPRC: area under the precision-recall curve; LR: logistic regression; LSTM: long short-term memory; mlGwDG: maximum information gain-weighted deviation geometric; XGBoost: extreme gradient boosting.



Patient-Level Risk Stratification via Max-Pooled OOF Aggregation

To evaluate patient-level risk stratification performance, we aggregated admission-level OOF predictions into patient-level risk scores using a max-pooling strategy across each patient's longitudinal sequence. This resulted in 71,212 patient-level risk profiles derived from 226,992 admissions (19.52% positive prevalence).

As shown in Table 3, patient-level aggregation improved performance across all evaluated models compared with admission-level evaluation. The proposed XGBoost-mIGwDG model achieved the best performance, with an OOF-AUC of 0.902 (95% CI 0.899-0.905) and an OOF-AUPRC of 0.730 (95% CI 0.722-0.737). The LSTM baseline achieved an OOF-AUPRC of 0.702 (95% CI 0.694-0.710), whereas LR using the full feature set achieved an OOF-AUPRC of 0.705 (95% CI 0.697-0.713).

XSL•FO
RenderX

Table 3. Patient-level predictive performance based on max-pooled OOF predictions. (number of patients=71,212)

Model	AUC[a] (95% CI)	AUPRC[b] (95% CI)
Deep sequential baseline		
LSTM[c]	0.891 (0.888-0.894)	0.702 (0.694-0.710)
Mean (basic)		
LR[d]	0.861 (0.858-0.864)	0.561 (0.552-0.571)
XGBoost[e]	0.888 (0.886-0.891)	0.684 (0.676-0.692)
mIGwDG[f] (proposed)		
LR	0.897 (0.894-0.900)	0.705 (0.697-0.713)
XGBoost	0.902 (0.899-0.905)	0.730 (0.722-0.737)

[a]AUC: area under the curve.

[b]AUPRC: area under the precision-recall curve.

[c]LSTM: long short-term memory.

[d]LR: logistic regression.

[e]XGBoost: extreme gradient boosting.

[f]mIGwDG: maximum information gain-weighted deviation geometric.

Evaluation in Clinically Vulnerable Subgroups

To evaluate performance within the most clinically vulnerable population, we performed a subgroup analysis using predefined clinical criteria that were independent of model predictions. The Charlson Comorbidity Index (CCI) was calculated for each index admission using all diagnosis codes recorded during the preceding 3-year observation window, thereby reflecting the cumulative burden of comorbidities prior to the index admission. Patients aged ≥75 years and with a CCI ≥4 were classified as the high-risk subgroup, isolating 24,574 high-acuity admissions (10.83%) and 7871 unique patients (11.05%). The predictive performance of all evaluated models within this subgroup is summarized in Table 4.

Table 4. The predictive performance of high-risk cohorts (age ≥75 years and CCIa≥4).

Model	Admission level		Patient level	
	AUC[b] (95% CI)	AUPRC[c] (95% CI)	AUC (95% CI)	AUPRC (95% CI)
Deep learning baseline				
LSTM[d]	0.765 (0.754-0.775)	0.787 (0.771-0.801)	0.817 (0.808-0.826)	0.878 (0.869-0.885)
Embedding approach				
LR[e]-mean	0.739 (0.729-0.750)	0.744 (0.730-0.758)	0.768 (0.757-0.779)	0.815 (0.803-0.828)
LR-mIGwDG[f]	0.765 (0.754-0.776)	0.787 (0.771-0.802)	0.819 (0.810-0.828)	0.873 (0.864-0.881)
XGBoost[g]-mean	0.754 (0.743-0.764)	0.772 (0.756-0.786)	0.800 (0.791-0.810)	0.856 (0.845-0.865)
XGBoost-mIGwDG	0.776 (0.766-0.785)	0.796 (0.783-0.809)	0.831 (0.822-0.839)	0.884 (0.874-0.892)

[a]CCI: Charlson Comorbidity Index.

[b]AUC: area under the curve.

[c]AUPRC: area under the precision-recall curve.

[d]LSTM: long short-term memory.

[e]LR: logistic regression.

[f]mIGwDG: maximum information gain-weighted deviation geometric.

[g]XGBoost: extreme gradient boosting.

We further evaluated model performance within the high-risk subgroup at both the admission level (event rate: 54.98%) and the patient level (event rate: 61.34%). At the admission level, XGBoost models incorporating the proposed trajectory representations consistently outperformed conventional aggregation strategies. The XGBoost-mIGwDG model achieved the best performance, with an AUC of 0.776 and an AUPRC of 0.796, compared with an AUPRC of 0.787 for the LSTM baseline. LR also showed consistent performance improvements after incorporating the proposed trajectory representation. After patient-level aggregation, performance further improved across all evaluated models. The XGBoost-mIGwDG model achieved

XSL•FO
RenderX

an AUPRC of 0.884, comparable to the LSTM baseline (AUPRC=0.878). LR also improved, with AUPRC increasing from 0.815 using mean aggregation to 0.873 with the proposed trajectory representation.

These results demonstrate consistent performance of the proposed representation across admission-level and patient-level high-risk stratification settings. Because the high-risk subgroup has a higher baseline event rate than the overall cohort, these performance metrics should be interpreted within the subgroup rather than directly compared with those of the full population.

Robustness to Incomplete Diagnostic Information

To evaluate the robustness of the proposed disease trajectory representation to incomplete diagnostic information, we conducted an additional analysis by retaining admissions that lacked disease-specific diagnostic information after preprocessing. This expanded cohort comprised 71,323 patient-level risk profiles across 227,328 admissions, with positive prevalence rates of 18.80% and 19.49%, respectively. In the primary analysis, these admissions were excluded because they did not contribute meaningful disease signals to the trajectory representation. In the robustness analysis, these admissions were retained to represent increased structural uncertainty arising from incomplete diagnostic information, a common characteristic of real-world EHR data. Compared with the primary analysis, predictive performance showed only marginal reductions, suggesting that the learned trajectory representation is relatively robust to moderate levels of incomplete diagnostic information. The detailed predictive performance of this robustness analysis is presented in Table 5.

Table 5. Performance comparison between the primary analysis and the robustness analysis with retained admissions lacking disease-specific diagnostic information.

Analysis	Admission level		Patient level	
	AUC[a] (95% CI)	AUPRC[b] (95% CI)	AUC (95% CI)	AUPRC (95% CI)
XGBoost[c]–mIGwDG[d]				
Primary	0.888 (0.885-0.891)	0.667 (0.657-0.677)	0.902 (0.899-0.905)	0.730 (0.722-0.737)
Robustness	0.886 (0.883-0.890)	0.660 (0.650-0.671)	0.900 (0.898-0.904)	0.722 (0.715-0.729)
LR[e]–mIGwDG				
Primary	0.881 (0.878-0.885)	0.646 (0.635-0.658)	0.897 (0.894-0.900)	0.705 (0.697-0.713)
Robustness	0.881 (0.878-0.885)	0.640 (0.628-0.652)	0.897 (0.894-0.899)	0.699 (0.690-0.706)

[a]AUC: area under the curve.

[b]AUPRC: area under the precision-recall curve.

[c]XGBoost: extreme gradient boosting.

[d]mIGwDG: maximum information gain-weighted deviation geometric.

[e]LR: logistic regression.

Analysis of Comorbidity Structure via the IG×PPMI Framework

To interpret model risk identification specifically for HF (*ICD*: I50), we constructed an IG-weighted comorbidity subgraph that integrates statistical co-occurrence structure with IG-based feature relevance. For this post hoc interpretability analysis, the PPMI-based co-occurrence structure and IG values were calculated separately using unmasked diagnostic records. The IG values used to determine node weights in this visualization were derived specifically for the interpretability analysis, and they were not used for predictive feature construction, model training, or performance evaluation. In contrast, the predictive analyses were conducted using the masked diagnosis sequences described above. To synthesize structural relevance with feature relevance, node weights were assigned based on the separately calculated IG values, while edge weights represented the co-occurrence strength measured by PPMI. The resulting IG×PPMI comorbidity subgraph highlights diagnostic codes that were strongly represented in the observed comorbidity structure and informative in the IG-based interpretability analysis. It also provided an interpretable lens into the local comorbidity architecture surrounding the target diagnosis. Notably, this visualization serves as a post hoc interpretability tool, designed to elucidate informative comorbidity configurations rather than to infer direct causal pathways among diseases.

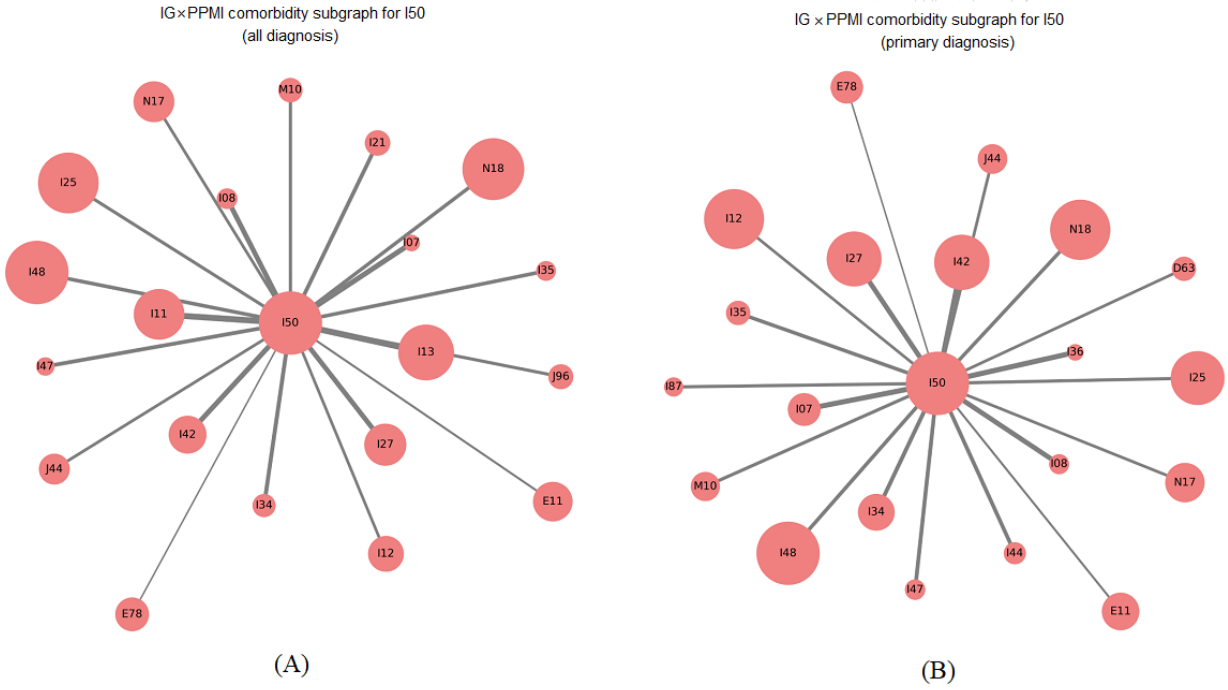
To investigate how diagnostic roles influence comorbidity architectures, we compared I50-centered subgraphs across different contexts (Figure 2). In general diagnostic contexts (Figure 2A), the neighborhood is characterized by a broad spectrum of systemic conditions, including hypertension-related disorders (I11, I13) and metabolic syndromes (E11, E78, and M10), alongside acute critical events such as myocardial infarction (I21). These findings suggest that, in broader clinical contexts, I50 often manifests as a comorbid component within a milieu of chronic systemic burden and acute dysregulation. Conversely, when I50 is designated as the primary diagnosis (Figure 2B), the comorbidity architecture becomes more focused on intrinsic structural and functional pathologies. Notably, conditions such as renal anemia (D63), conduction disorders (I44), and venous congestion (I87) emerge as top-20 components exclusively in this primary diagnostic role. This divergence

XSL•FO
RenderX

demonstrates that the IG×PPMI framework effectively captures nuanced contextual semantics, identifying distinct clinical profiles that transcend simple co-occurrence frequencies.

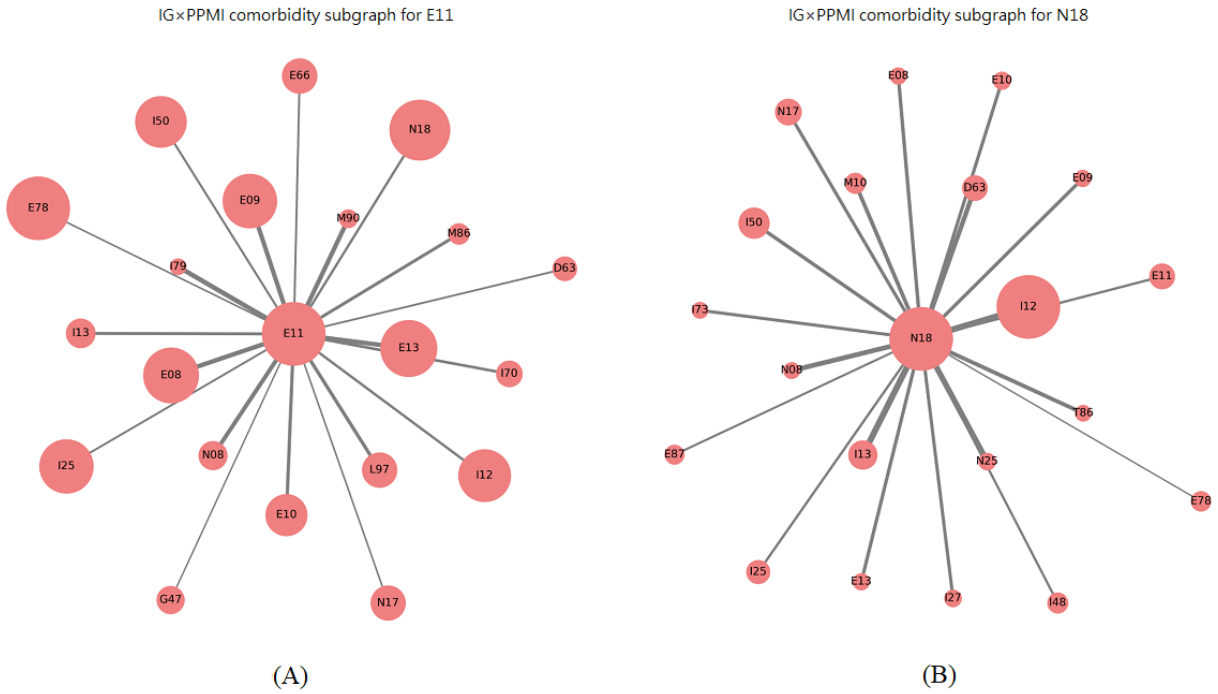
Furthermore, the IG×PPMI framework revealed 2 prominent structural modules centered on E11 (type 2 diabetes) and N18 (CKD), respectively (Figure 3). The E11-centered module was primarily associated with cardiovascular and metabolic diseases, including dyslipidemia (E78) and chronic ischemic heart disease (I25), whereas the N18-centered module was closely associated with cardiorenal complications, including hypertensive heart and CKD (I13), electrolyte imbalance (E87), and hypertensive nephropathy (I12). These findings reveal 2 distinct comorbidity modules surrounding HF, providing a structural basis for subsequent interpretation of heterogeneous disease progression patterns.

Figure 2. Top 20 IG-weighted comorbidity subgraphs centered on HF (I50). (a) Comorbidity subgraph constructed from all diagnostic contexts. (b) Comorbidity subgraph constructed from admissions in which I50 was the primary diagnosis. Node size reflects min-max normalized (IG) values and linearly mapped to the range of 200-3200 for visualization purposes. The target ICD node was assigned a fixed size of 3200. Edge width is proportional to the corresponding PPMI score.



XSL•FO
RenderX

Figure 3. Visualization of the top-20 information gain (IG)–weighted comorbidity subgraphs centered on (A) E11 and (B) N18. Node size reflects minimum-maximum normalized (IG) values and is linearly mapped to the range of 200-3200 for visualization purposes. The target International Classification of Diseases (ICD) node was assigned a fixed size of 3200. Edge width is proportional to the corresponding positive pointwise mutual information (PPMI) score. IG: information gain; PPMI: positive pointwise mutual information.



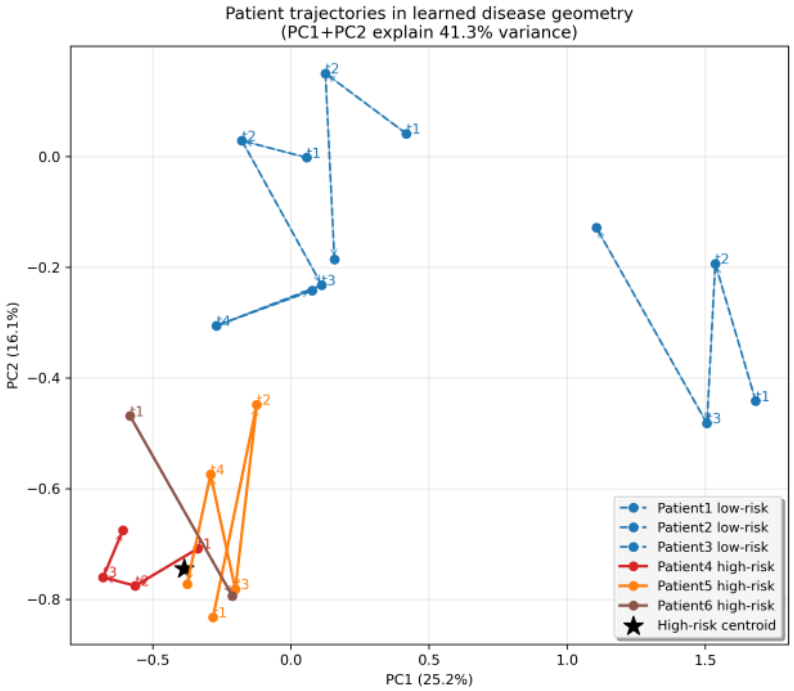
(A) (B)

Longitudinal Trajectory Analysis and Geometric Interpretation

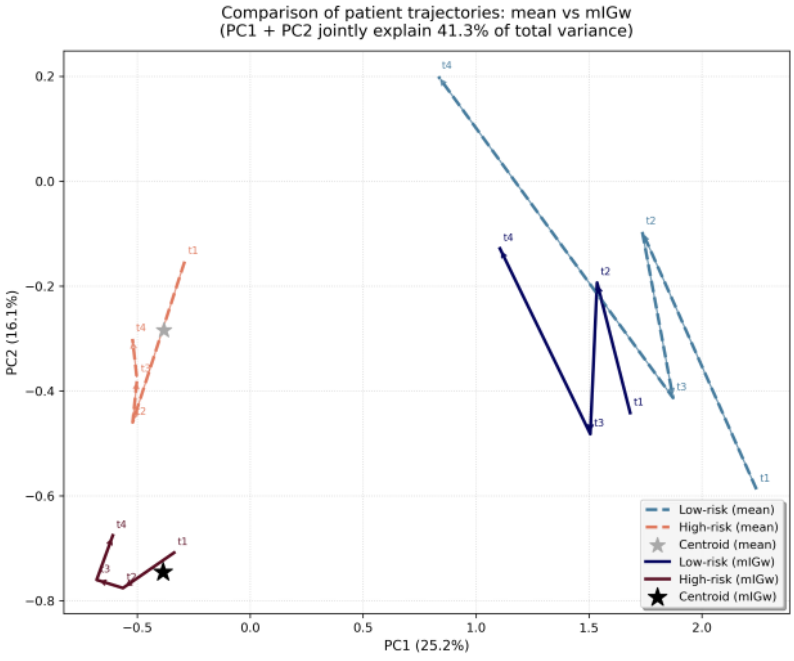
To assess the clinical face validity of our framework, we visualized the disease histories of 3 low-risk and 3 high-risk patients, selected based on their calibrated OOF risk probabilities, using principal component analysis (PCA; Figure 4A). To provide a common visualization space, the PCA basis was fitted on the mean-pooled admission embeddings derived from all admissions in the cohort, and the visit-level maximum information gain-weighted (mIGw) embeddings were subsequently projected onto this shared coordinate system. Here, mIGw denotes the visit-level information-gain-weighted embedding (Equation 6), obtained by applying the IG weights to the embedded *ICD* codes within each hospitalization. These visit-level representations are subsequently aggregated along each patient's longitudinal hospitalization history to construct the mIGwDG path-level representation used for model training and prediction. The first 2 principal components accounted for 41.3% of the total variance (25.2% and 16.1%, respectively). Each point represents an individual admission embedding, and consecutive admissions were connected chronologically with directed arrows to form a patient trajectory, illustrating the evolution of the patient's clinical state over time.

XSL•FO
RenderX

Figure 4. Risk-stratified disease trajectory visualizations in the International Classification of Diseases (ICD) embedding space. (A) Principal component analysis (PCA) projection of trajectories in the learned disease geometry. Each sequential clinical encounter (admission) is labeled chronologically (t1, t2, etc). (B) Comparison of representative trajectories generated by conventional mean pooling and the proposed representation. The representative low-risk patient is shown in blue, the representative high-risk patient in red, and the learned high-risk centroid in black. mIGw: maximum information gain-weighted.



(A)



(B)

As illustrated in Figure 4A, representative high-risk (calibrated OOF>60%) and low-risk (calibrated OOF<0.1%) patients occupy visually distinct regions in the learned disease geometry. The representative high-risk trajectories remain localized near the learned high-risk centroid (marked by the black star), whereas the representative low-risk trajectories consistently remain distal to the high-risk centroid. Although this visualization is illustrative rather than quantitative, it qualitatively shows that patient trajectories are organized into a coherent geometric structure within the learned representation. This alignment between geometric movement and calibrated

XSL•FO
RenderX

model-estimated risk probabilities provides evidence for the interpretability of our latent representation.

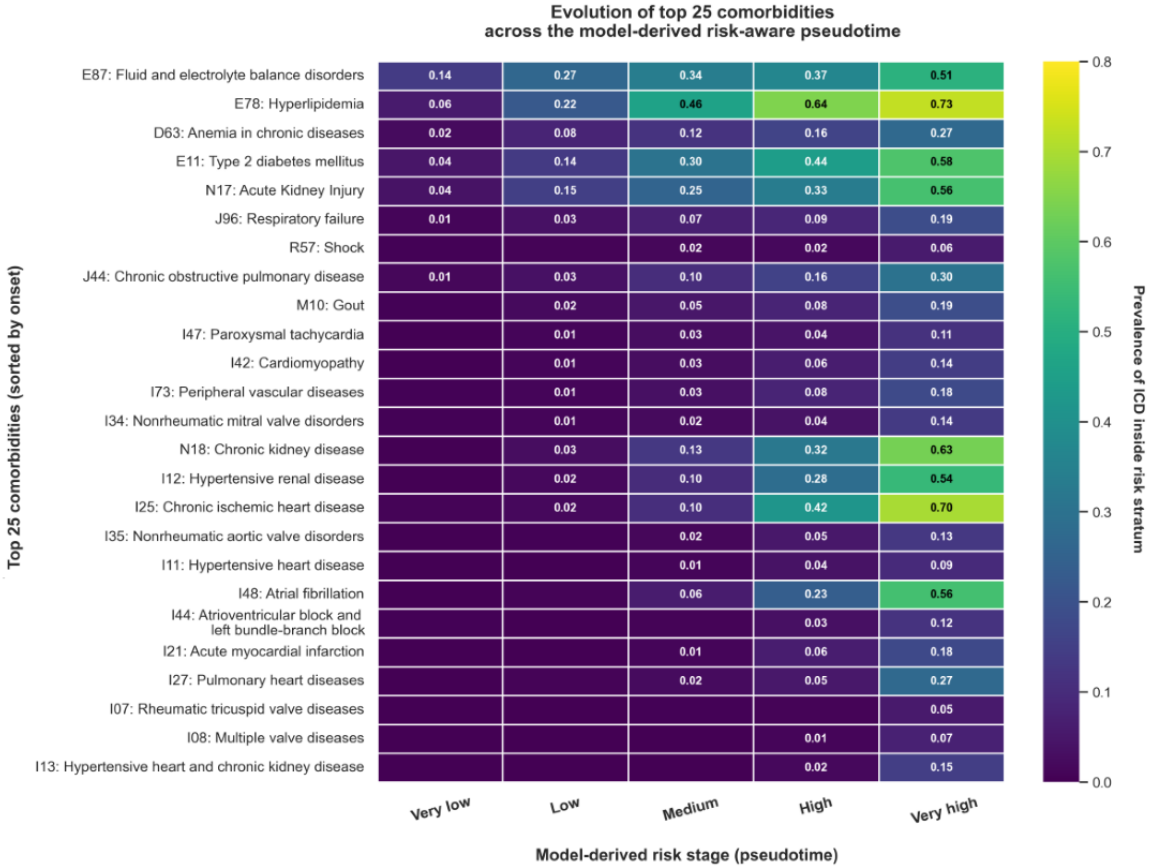
To qualitatively examine how the proposed representation influences the learned disease geometry, representative patient trajectories and the corresponding high-risk centroid were projected into a 2D space using PCA. Figure 4B compares the trajectory representations generated by the conventional mean pooling strategy and the proposed mIGw representation using the same representative for patients with high-risk and low-risk. Both representations were projected onto the same PCA coordinate system to enable direct visual comparison. In both representations, the representative high-risk trajectory remained in the vicinity of the learned high-risk centroid, whereas the representative low-risk trajectory occupied a distinct region of the embedding space.

Compared with the conventional mean pooling strategy, the proposed mIGw representation produced more compact trajectory distributions with clearer separation between the representatives of patients with high-risk and low-risk. The proposed representation also appeared to produce more localized high-risk trajectories, with the representative patient with high-risk remaining closer to the learned high-risk centroid while exhibiting reduced geometric variability around the centroid. Although these visualizations are qualitative rather than statistically conclusive, they provide a geometric perspective that complements the quantitative predictive performance reported for mIGwDG. These qualitative observations suggest that the proposed representation better preserves the local geometric organization of patients with similar disease risk, resulting in more coherent trajectory organization within the embedding space.

Risk-Aware Pseudotime Analysis of Comorbidity Evolution

In clinical practice, disease progression is highly heterogeneous; some patients remain stable over a decade of follow-up, while others experience rapid functional decline within only 2 years. Traditional temporal analyses often struggle to align these disparate trajectories when relying solely on chronological time or visit counts. To overcome this limitation, we visualized the top 25 comorbid *ICD* codes identified through the IG×PPMI framework using unmasked diagnostic records along a model-derived, risk-aware pseudotime axis derived from the masked predictive representation (Figure 5). Individual patient trajectories and their corresponding longitudinal comorbidity patterns were visualized using unmasked diagnostic records, whereas patients were ordered along the pseudotime axis according to mIGwDG risk scores rather than absolute time, thereby treating risk magnitude as the primary metric of progression. This ensures that comorbidity trajectories reflect the pathological evolution of the disease rather than mere temporal accumulation, providing a standardized scale to observe how comorbidity profiles transition across 5 risk strata (from Very low to Very high).

Figure 5. Risk-aware pseudotime heatmap of the top 25 information gain (IG)×positive pointwise mutual information (PPMI) comorbidities. Pseudotime ordering was based on maximum information gain-weighted deviation geometric (mIGwDG) predictions from masked diagnosis sequences, whereas comorbidity patterns were obtained from unmasked diagnostic records for post hoc interpretation. ICD: International Classification of Diseases.



XSL•FO
RenderX

The analysis reveals a clear monotonic accrual of chronic systemic conditions rather than stochastic acute episodes. Foundational metabolic disorders, such as hyperlipidemia (E78) and type 2 diabetes (E11), are markedly elevated even in low-risk stages, with E78 reaching a prevalence of 0.78 at the very high-risk stage. As patients progress, a cardiorenal-metabolic signature emerges with increasing synchrony; CKD (N18), hypertensive kidney disease (I12), and electrolyte disorders (E87) exhibit a pronounced gradient. This trajectory mirrors the pathophysiological shift from localized injury to systemic homeostatic failure.

Distinct temporal patterns were observed across disease categories. Cardiovascular comorbidities, including cardiomyopathy (I42) and atrial fibrillation (I48), show concentrated gains in the moderate-to-high risk range. In contrast, acute conditions like shock (R57) and myocardial infarction (I21) remain infrequent until the very high-risk stage, suggesting they serve as indicators of terminal decompensation rather than early risk markers.

Collectively, the model-inferred risk stratification identifies a progressive accrual of chronic systemic comorbidities rather than stochastic acute episodes. This suggests that the model effectively prioritizes structural disease evolution. Consequently, as patients transition into high-risk strata, our findings advocate for a clinical shift toward integrated multiorgan management, addressing the complex interplay between cardiac, renal, and pulmonary systems.

The y-axis represents *ICD* codes, and the x-axis represents 5 risk strata along the model-inferred pseudotime axis. Color intensity indicates the normalized prevalence of each *ICD* code within each risk stratum. The continuous pseudotime was divided into 5 ordinal risk strata (Very low, Low, Medium, High, and Very high) using quantile-based binning based on the empirical 20th, 40th, 60th, and 80th percentiles of the calibrated pseudotime distribution. This pseudotime representation is intended to summarize model-inferred risk progression and should not be interpreted as a direct temporal or causal ordering of diseases. Instead, it illustrates how clinically relevant comorbidities vary across increasing levels of model-estimated risk.

Discussion

Principal Findings

In this study, we developed a longitudinal *ICD* trajectory representation that incorporates geometric structural information for high-risk patient identification from EHR data. By integrating geometric features, including historical means and maxima, Euclidean distances, directional shifts, and cosine dispersion, with IG-based weighting, the proposed approach summarizes disease histories in a structurally informed manner. Rather than relying solely on conventional pooling or sequential modeling, the proposed approach is designed to preserve structural characteristics of disease trajectories that may reflect progression toward higher model-estimated risk. In addition, we introduced a risk-aware pseudotime framework that organizes heterogeneous patient trajectories along a latent risk continuum inferred from model predictions. Rather than relying on chronological alignment, this approach enables characterization of disease trajectory patterns according to increasing risk profiles, providing a complementary perspective for longitudinal risk stratification.

The proposed framework consistently outperformed conventional embedding aggregation strategies while achieving predictive performance comparable to the sparse *ICD* one-hot representation despite reducing the feature space by approximately 75%, from approximately 1600 *ICD* variables to a 389-dimensional trajectory representation. Beyond prediction, the learned representation facilitates visualization of disease trajectories, risk-aware pseudotime analysis, and geometric characterization of disease evolution, providing additional interpretability through visualization and trajectory characterization that high-dimensional sparse one-hot features alone do not readily provide. Furthermore, combining mIGwDG with the original *ICD* one-hot representation consistently improved the performance of both LR and XGBoost, suggesting that the proposed representation captures complementary information beyond conventional sparse *ICD* encoding.

Methodological and Practical Advantages

While deep learning architectures (such as RNNs and Transformers) offer substantial flexibility for temporal modeling, their performance in clinical risk prediction may depend on large datasets, adequate representation of positive cases, and extensive hyperparameter tuning. In the context of HF, hospitalization data are characterized by high heterogeneity, sparse events, and pronounced class imbalance, conditions under which deep models may struggle with stability and generalizability.

Recent GNN frameworks have demonstrated excellent predictive performance by jointly modeling longitudinal EHRs and comorbidity connectivity through heterogeneous graph attention mechanisms [18]. These approaches represent an important direction in medical representation learning but generally require more sophisticated model architectures and greater computational resources. As a complementary alternative, the proposed framework employs Node2Vec embeddings coupled with geometric trajectory descriptors, providing a compact yet explicit representation of longitudinal comorbidity patterns. This design ensures robust learning and consistent performance even under moderate sample sizes. A key advantage of this approach is its intrinsic interpretability; it decomposes historical hospitalizations into IG-weighted representations and geometric measures (displacement and dispersion), and each component can be independently linked to clinically meaningful patterns.

Furthermore, the computational efficiency of the Node2Vec and XGBoost pipeline facilitates deployment in standard information systems without specialized hardware. Notably, the proposed representation remains highly effective even when coupled with a simple LR classifier. This suggests that the proposed representation reshapes longitudinal clinical representations into a more separable risk space, reducing the reliance on high-capacity, noninterpretable nonlinear models. Recent studies have likewise demonstrated that constructing informative representations from high-dimensional clinical data

XSL•FO
RenderX

prior to supervised learning can substantially improve predictive performance [34]. Although our framework does not employ explicit feature selection through statistical correlation analysis or optimization-based search algorithms, it follows a similar methodological motivation by transforming sparse longitudinal disease histories into compact graph-based geometric representations before classification. Consequently, predictive performance is enhanced through more informative representations rather than by relying solely on more complex predictive models.

Disease Trajectory Representation Framework

Traditional pooling-based embedding approaches, such as mean pooling or max pooling, summarize diagnostic information either by representing an overall central tendency of a patient's disease space or by emphasizing a small number of extreme diagnoses that emphasize the most prominent diagnostic features. While effective for dimensionality reduction, these methods fail to distinguish between stable disease burden and abrupt pathological deviations along the disease course.

In contrast, our framework is designed to jointly capture chronic burden and acute shifts. By employing geometry-based features, we quantify the spatial relationships between diagnostic combinations and their proximity to a high-risk region within the embedding space. The average distance was designed to quantify diagnostic dispersion, indicating whether a patient's profile is concentrated on a single clinical axis or spans multiple comorbidity modules. The variability of these distances characterizes systemic complexity; higher dispersion may reflect greater heterogeneity across comorbidity patterns.

A distinctive innovation is the incorporation of deviation vectors, providing a directional geometric perspective on comorbidity evolution. We observed that high-risk trajectories deviate from low-risk regions, gravitating toward the high-risk centroid. This displacement may reflect systematic changes in disease trajectory organization rather than the influence of isolated diagnostic events. Consequently, this trajectory-level representation provides additional information for risk stratification, enabling the model to characterize the internal architecture of diagnostic combinations. By capturing this structural information, our framework enhances risk discrimination without necessitating the computational overhead of complex deep learning architectures. The primary contribution of these geometric features is to characterize changes in disease trajectories rather than individual diagnostic states.

Diagnostic Interpretability via the IG×PPMI Comorbidity Framework

The IG×PPMI comorbidity analysis provides an additional perspective for interpreting the learned representations. By integrating model-derived attribution (IG) with population-level co-occurrence (PPMI), the framework highlights clinically relevant comorbidity patterns that may not be apparent from either source alone. An advantage of this approach is its ability to identify comorbidities that receive relatively high IG attribution despite only moderate population co-occurrence. These findings suggest that the proposed representation may capture diagnostic patterns that contribute disproportionately to model-estimated risk beyond highly prevalent comorbidities such as hypertension. The combined IG×PPMI framework therefore complements conventional feature importance analyses by jointly considering model-derived predictive attribution and population-level disease co-occurrence.

In the top 20 comorbidities in the HF-centered subgraph (I50), anemia (D63), and electrolyte disorders (E87) exhibit high IG weights, often emerging before overt symptoms like orthopnea or marked edema. Anemia, with a reported prevalence of 35%-70% in patients with HF [35,36], is a well-established marker of disease severity and is strongly associated with increased hospitalization and mortality [37]. Similarly, electrolyte disturbances (E87), including hyponatremia and potassium imbalances, reflect both intrinsic neurohormonal dysregulation and the impact of pharmacological interventions. These imbalances serve as critical indicators of pathophysiological shifts and are robustly linked to adverse outcomes [38,39].

Furthermore, among the identified comorbidity modules, the E11- and N18-centered modules are of particular clinical interest because they correspond to 2 well-established systemic syndromes in HF. The E11-centered module is consistent with the well-established cardiometabolic continuum, whereas the N18-centered module is consistent with a common cardiorenal syndrome in patients with advanced HF. These findings indicate that the proposed IG×PPMI framework could identify clinically meaningful comorbidity structures beyond what can be inferred from diagnostic frequency alone. By integrating feature importance with disease co-occurrence, the framework may reveal potential early warning signals embedded within routine EHRs, thereby highlighting clinically relevant comorbidity patterns associated with HF and facilitating clinical risk stratification.

Risk-Aware Pseudotime Analysis Framework

A second contribution of this study is the introduction of the risk-aware pseudotime framework, which provides a risk-oriented organization of longitudinal disease trajectories. Longitudinal EHR analyses are often challenged by temporal heterogeneity and irregular sampling, making direct alignment of clinical trajectories across individuals difficult. Our framework addresses this challenge by projecting discrete patient histories onto a "model-inferred risk continuum" derived from longitudinal disease representations. This continuum provides a common axis for characterizing heterogeneous patient trajectories according to increasing model-estimated risk rather than chronological timing alone.

The resulting risk-aware organization enables investigation of how comorbidity patterns vary across the inferred risk spectrum. The observed enrichment of cardiorenal-metabolic conditions, including type 2 diabetes (E11) and CKD (N18), along the pseudotime axis highlights their importance in the disease trajectories captured by the model. Furthermore, the striking contrast between the gradual accrual of these chronic comorbidities and the abrupt, infrequent appearance of terminal events like shock (R57) validates the framework's capacity to prioritize structural disease evolution over episodic triggers.

XSL•FO
RenderX

Collectively, these findings suggest that the proposed geometric trajectory framework captures longitudinal patterns of multimorbidity associated with increasing HF risk. Beyond predictive modeling, the risk-aware pseudotime framework provides an interpretable approach for visualizing disease trajectory organization and identifying patterns of comorbidity evolution preceding high-risk states.

Limitations

Real-world longitudinal EHR data are inherently sparse and heterogeneous, often containing irregular follow-up intervals, variable numbers of historical admissions, and incomplete diagnosis coding due to administrative encounters or heterogeneous documentation practices. Such uncertainty may affect the construction of longitudinal disease trajectories by introducing admissions with limited disease-specific information. In our robustness analysis, retaining additional diagnosis-free admissions resulted in only marginal reductions in predictive performance, suggesting that the proposed trajectory representation maintained stable performance under moderate levels of incomplete diagnostic information. One possible explanation is that the representation aggregates disease information across multiple longitudinal encounters rather than relying on individual admissions alone, thereby reducing the impact of isolated incomplete records on the overall trajectory representation.

In addition, the acute-care focus of the MIMIC-IV database may introduce cohort selection bias, as patients receiving frequent or severe hospital-based care may be overrepresented in the available longitudinal records. This may limit the generalizability of the observed disease trajectories and risk patterns to broader populations, particularly those managed predominantly in outpatient or community settings.

Nevertheless, several limitations should be acknowledged. First, although the proposed framework naturally accommodates variable-length trajectories, patients with only a single prior admission inevitably contribute less longitudinal information than patients with multiple readmissions. Consequently, the richness of the learned trajectory representation remains dependent on the amount of historical clinical information available. Second, the present robustness analysis evaluated only one practical form of structural uncertainty. Other sources, including varying degrees of missing diagnoses, systematic coding bias, and missing encounters, warrant further investigation. Finally, this work lacks external validation on independent multicenter cohorts. Although the framework demonstrates strong internal validity, its generalizability across heterogeneous health care systems remains to be established in future studies.

A further limitation of this study is the exclusion of diagnoses that are highly correlated with HF, including hypertensive heart disease and cardiomyopathies. Although this strategy is methodologically rigorous from a machine learning perspective, being adopted to reduce information leakage and ensure that the model learned broader longitudinal comorbidity patterns rather than relying on near-proxy indicators of the outcome, it also results in a simplified representation of the patient's clinical condition. In routine clinical practice, these diagnoses constitute important etiological and prognostic information and are frequently incorporated into clinical decision-making. Consequently, the present framework should be interpreted primarily as an evaluation of the predictive value of latent comorbidity structure rather than a fully comprehensive clinical risk assessment model. Future work should investigate the integration of both disease trajectory features and direct HF-related diagnoses to better reflect real-world clinical decision support scenarios.

Future Directions

Future work should focus on external validation across diverse cohorts, incorporation of multimodal clinical information, and further development of advanced trajectory modeling approaches to better characterize complex disease evolution across clinical domains.

Conclusions

In conclusion, we developed a longitudinal disease trajectory representation that integrates IG-based weighting, disease co-occurrence information, and geometric structural features to transform heterogeneous EHR data into compact and interpretable representations of clinical evolution. The proposed framework captures latent patterns associated with disease trajectories and provides a risk-aware perspective for longitudinal risk stratification in high-noise clinical data. Beyond predictive performance, this approach facilitates visualization and characterization of disease trajectories, providing a framework for further investigation of trajectory-based risk assessment in longitudinal clinical data.

Acknowledgments

The authors would like to thank the Massachusetts Institute of Technology (MIT) Laboratory for Computational Physiology for providing access to the MIMIC-IV database. We also appreciate the efforts of the clinical and technical staff at the Beth Israel Deaconess Medical Center for their data collection and curation efforts. This research was made possible through the use of the PhysioNet platform.

Funding

The article processing fee for this publication was supported by the National Science and Technology Council, Taiwan (Grant No. MOST 111-2221-E-027-113-MY2), and the joint research project between National Taipei University of Technology and Beijing University of Technology (Project No. NTUT-BJUT-111-062), coordinated by Dr. Maoshen Jia of the Faculty of Information Technology, Beijing University of Technology, Beijing, China.

XSL•FO
RenderX

Generative AI Use

The authors used ChatGPT (OpenAI) to assist with language editing, grammar correction, improving clarity of scientific writing, and revising manuscript text. All study design, methodology development, data processing, statistical analyses, interpretation of results, and scientific conclusions were performed solely by the authors, who take full responsibility for the content of the manuscript.

Data Availability

The medical data analyzed in this study were obtained from the Medical Information Mart for Intensive Care IV (MIMIC-IV, version 2.2) database, hosted on PhysioNet. Due to patient privacy concerns and data protection policies, access to the dataset requires credentialed users to complete the required CITI training and agree to the data use terms.

The source code and computational pipelines supporting the findings of this study are publicly available at the following GitHub repository. The repository contains the core implementation of the proposed mIGwDG-based modeling pipeline, including preprocessing, Node2Vec embedding, information gain weighting, XGBoost training, and model evaluation. Original MIMIC-IV data cannot be redistributed because of data-use and privacy restrictions.

Authors' Contributions

Conceptualization: J-LH, C-HY
Data curation: J-LH, T-WP
Formal analysis: J-LH, C-HY
Investigation: J-LH
Methodology: J-LH, C-HY
Project administration: T-WP
Resources: J-LH, T-WP
Supervision: T-WP
Software: J-LH, C-HY
Validation: J-LH, C-HY
Visualization: J-LH, C-HY
Writing – original draft: J-LH, C-HY
Writing – review & editing: T-WP

Conflicts of Interest

None declared.

Multimedia Appendix 1

The detailed hyperparameter settings for the Node2Vec embedding model.
[PDF File (Adobe PDF File), 61 KB-Multimedia Appendix 1]

Multimedia Appendix 2

The complete hyperparameter settings for XGBoost classifier and Logistic Regression model.
[PDF File (Adobe PDF File), 13 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Reliability diagram calibrated based on Platt scaling. The red line represents the probability curve of the original XGBoost model. The blue line represents the curve after Platt scaling calibration. The black dashed line indicates the 45° perfect calibration diagonal.
[PNG File , 217 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Progressive ablation analysis of the proposed feature modules.
[PDF File (Adobe PDF File), 146 KB-Multimedia Appendix 4]

References

XSL•FO
RenderX

1. Ndumele CE, Rangaswami J, Chow SL, Neeland IJ, Tuttle KR, Khan SS, et al. American Heart Association. Cardiovascular-kidney-metabolic health: a presidential advisory from the American Heart Association. Circulation. 2023;148(20):1606-1635. [doi: 10.1161/CIR.0000000000001184] [Medline: 37807924]
2. Shahim B, Kapelios CJ, Savarese G, Lund LH. Global public health burden of heart failure: an updated review. Card Fail Rev. 2023;9:e11. [FREE Full text] [doi: 10.15420/cfr.2023.05] [Medline: 37547123]
3. Gimeno-Miguel A, Gracia Gutiérrez A, Poblador-Plou B, Coscollar-Santaliestra C, Pérez-Calvo JI, Divo MJ, et al. Multimorbidity patterns in patients with heart failure: an observational Spanish study based on electronic health records. BMJ Open. Dec 23, 2019;9(12):e033174. [FREE Full text] [doi: 10.1136/bmjopen-2019-033174] [Medline: 31874886]
4. Chamberlain AM, St Sauver JL, Gerber Y, Manemann SM, Boyd CM, Dunlay SM, et al. Multimorbidity in heart failure: a community perspective. Am J Med. 2015;128(1):38-45. [FREE Full text] [doi: 10.1016/j.amjmed.2014.08.024] [Medline: 25220613]
5. Tomasoni D, Vitale C, Guidetti F, Benson L, Braunschweig F, Dahlström U, et al. The role of multimorbidity in patients with heart failure across the left ventricular ejection fraction spectrum: data from the swedish heart failure registry. Eur J Heart Fail. 2024;26(4):854-868. [FREE Full text] [doi: 10.1002/ejhf.3112] [Medline: 38131248]
6. Mentz RJ, O'Connor CM. Pathophysiology and clinical evaluation of acute heart failure. Nat Rev Cardiol. 2016;13(1):28-35. [doi: 10.1038/nrcardio.2015.134] [Medline: 26370473]
7. Gerhardt T, Gerhardt LMS, Ouwerkerk W, Roth GA, Dickstein K, Collins SP, et al. Multimorbidity in patients with acute heart failure across world regions and country income levels (REPORT-HF): a prospective, multicentre, global cohort study. Lancet Glob Health. 2023;11(12):e1874-e1884. [doi: 10.1016/s2214-109x(23)00408-4]
8. Screever EM, van der Wal MHL, van Veldhuisen DJ, Jaarsma T, Koops A, van Dijk KS, et al. Comorbidities complicating heart failure: changes over the last 15 years. Clin Res Cardiol. 2023;112(1):123-133. [FREE Full text] [doi: 10.1007/s00392-022-02076-1] [Medline: 35976430]
9. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc. 2017;24(1):198-208. [FREE Full text] [doi: 10.1093/jamia/ocw042] [Medline: 27189013]
10. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. 2018;1(1):18. [FREE Full text] [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
11. Wei WQ, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, Gamazon ER, et al. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. PLoS One. 2017;12(7):e0175508. [FREE Full text] [doi: 10.1371/journal.pone.0175508] [Medline: 28686612]
12. Ng K, Steinhubl SR, deFilippi C, Dey S, Stewart WF. Early detection of heart failure using electronic health records: practical implications for time before diagnosis, data diversity, data quantity, and data density. Circ Cardiovasc Qual Outcomes. 2016;9(6):649-658. [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.116.002797] [Medline: 28263940]
13. Jensen AB, Moseley PL, Oprea TI, Ellesøe SG, Eriksson R, Schmock H, et al. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nat Commun. 2014;5(1):4022. [FREE Full text] [doi: 10.1038/ncomms5022] [Medline: 24959948]
14. Choi Y, Chiu CY, Sontag D. Learning low-dimensional representations of medical concepts. AMIA Jt Summits Transl Sci Proc. 2016;2016:41-50. [FREE Full text] [Medline: 27570647]
15. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep. 2016;6:26094. [FREE Full text] [doi: 10.1038/srep26094] [Medline: 27185194]
16. Grover A, Leskovec J. node2vec: Scalable feature learning for networks. KDD. 2016;2016:855-864. [FREE Full text] [doi: 10.1145/2939672.2939754] [Medline: 27853626]
17. Beam AL, Kompa B, Schmaltz A, Fried I, Weber G, Palmer N, et al. Clinical concept embeddings learned from massive sources of multimodal medical data. Pac Symp Biocomput. 2020;25:295-306. [FREE Full text] [Medline: 31797605]
18. Ramamurthy P, Han Z, Delen D, Miao Z, Gin A, Luo X, et al. Graph attention network with comorbidity connectivity embedding for post-traumatic epilepsy risk prediction using sparse time-series electronic health records. Int J Med Inform. 2026;208:106239. [doi: 10.1016/j.ijmedinf.2025.106239] [Medline: 41478164]
19. Rotmensch M, Halpern Y, Tlimat A, Horng S, Sontag D. Learning a health knowledge graph from electronic medical records. Sci Rep. 2017;7(1):5994. [FREE Full text] [doi: 10.1038/s41598-017-05778-z] [Medline: 28729710]
20. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. J Am Med Inform Assoc. 2018;25(10):1419-1428. [FREE Full text] [doi: 10.1093/jamia/ocy068] [Medline: 29893864]
21. Ayala Solares JR, Diletta Raimondi FE, Zhu Y, Rahimian F, Canoy D, Tran J, et al. Deep learning for electronic health records: a comparative review of multiple deep neural architectures. J Biomed Inform. 2020;101:103337. [FREE Full text] [doi: 10.1016/j.jbi.2019.103337] [Medline: 31916973]
22. Landi I, Glicksberg BS, Lee H, Cherng S, Landi G, Danieletto M, et al. Deep representation learning of electronic health records to unlock patient stratification at scale. NPJ Digit Med. 2020;3:96. [FREE Full text] [doi: 10.1038/s41746-020-0301-z] [Medline: 32699826]

XSL•FO
RenderX

23. Jain S, Wallace BC. Attention is not Explanation. In: Minneapolis, Minnesota: Association for Computational Linguistics. June; Minneapolis, Minnesota. Association for Computational Linguistics; 2019. Presented at: NAACL-HLT 2019; June 2–7, 2019:3543-3556; Minneapolis, Minnesota. [doi: 10.18653/v1/N19-1357]
24. Wiegreffe S, Pinter Y. Attention is not not Explanation. Hong Kong, China. Association for Computational Linguistics; 2019. Presented at: EMNLP-IJCNLP 2019; November 3–7, 2019:11-20; Hong Kong, China. [doi: 10.18653/v1/D19-1002]
25. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 2000;101(23):E215-E220. [doi: 10.1161/01.cir.101.23.e215] [Medline: 10851218]
26. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV (version 2.2). PhysioNet. 2023. [doi: 10.13026/6mm1-ek67]
27. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023;10(1):1. [FREE Full text] [doi: 10.1038/s41597-022-01899-x] [Medline: 36596836]
28. ICD-10 Files & News Archive. Centers for Medicare & Medicaid Services (CMS) URL: https://www.cms.gov/medicare/coding-billing/icd-10-codes/icd-10-cm-icd-10-pcs-gem-archive [accessed 2025-06-13]
29. Quinlan JR. Induction of decision trees. Mach Learn. 1986;1(1):81-106. [doi: 10.1007/bf00116251]
30. Church KW, Hanks P. Word association norms, mutual information, and lexicography. Computational Linguistics. 1990;16(1):22-29. [FREE Full text]
31. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Association for Computing Machinery; 2016. Presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug 16:785-794; New York, NY, USA. [doi: 10.1145/2939672.2939785]
32. McGilvray MMO, Heaton J, Guo A, Masood MF, Cupps BP, Damiano M, et al. Electronic health record-based deep learning prediction of death or severe decompensation in heart failure patients. JACC Heart Fail. 2022;10(9):637-647. [doi: 10.1016/j.jchf.2022.05.010]
33. Liu W, Stansbury C, Singh K, Ryan AM, Sukul D, Mahmoudi E, et al. Predicting 30-day hospital readmissions using artificial neural networks with medical code embedding. PLoS One. 2020;15(4):e0221606. [FREE Full text] [doi: 10.1371/journal.pone.0221606] [Medline: 32294087]
34. Tutsoy O, Ozturk HA, Sumbul HE. Big data dimensionality reduction-based supervised machine learning algorithms for NASH diagnosis. BMC Bioinformatics. 2025;26(1):256. [FREE Full text] [doi: 10.1186/s12859-025-06263-5] [Medline: 41120851]
35. Mirghani H, Alshreef AA, Al-Temani HA, Alanazi NK, Algohani A, Alrshidi WM, et al. Updates on the association between anemia and heart failure: a systematic review. Cureus. 2024;16(9):e69101. [doi: 10.7759/cureus.69101] [Medline: 39391392]
36. Puri P, Kachhadia MP, Singla N, Kataria H, Singh J, Singh H, et al. Anemia in heart failure: a systematic review of prevalence and prognostic impact. Cureus. 2025;17(11):e96363. [doi: 10.7759/cureus.96363] [Medline: 41367439]
37. Pan J, Liu M, Huang J, Chen L, Xu Y. Impact of anemia on clinical outcomes in patients with acute heart failure: a systematic review and meta-analysis. Clin Cardiol. 2024;47(2):e24228. [FREE Full text] [doi: 10.1002/clc.24228] [Medline: 38402548]
38. Urso C, Brucculeri S, Caimi G. Acid-base and electrolyte abnormalities in heart failure: pathophysiology and implications. Heart Fail Rev. 2015;20(4):493-503. [FREE Full text] [doi: 10.1007/s10741-015-9482-y] [Medline: 25820346]
39. Sato R, Koziolek MJ, von Haehling S. Translating evidence into practice: managing electrolyte imbalances and iron deficiency in heart failure. Eur J Intern Med. 2025;131:15-26. [FREE Full text] [doi: 10.1016/j.ejim.2024.10.024] [Medline: 39521682]

Abbreviations

AUC: area under the curve
AUPRC: area under the precision-recall curve
BFS: breadth-first search
BIDMC: Beth Israel Deaconess Medical Center
CCI: Charlson Comorbidity Index
CKD: chronic kidney disease
CKM: cardiovascular-kidney-metabolic
CMS: Centers for Medicare & Medicaid Services
DFS: depth-first search
EHR: electronic health record
GEM: General Equivalence Mapping
GNN: graph neural network
HF: heart failure
HIPAA: Health Insurance Portability and Accountability Act
ICD: International Classification of Diseases
ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification

XSL•FO
RenderX

ICD-10-CM: International Classification of Diseases, 10th Revision, Clinical Modification
IG: information gain
IRB: Institutional Review Board
LR: logistic regression
LSTM: long short-term memory
MIMIC-IV: Medical Information Mart for Intensive Care IV
MIT: Massachusetts Institute of Technology
mIGw: maximum information gain-weighted
mIGwDG: maximum information gain-weighted deviation geometric
MRCA: most recent common ancestor
NLP: natural language processing
OOF: out-of-fold
PCA: principal component analysis
PMI: pointwise mutual information
PPMI: positive pointwise mutual information
RNN: recurrent neural network
TF-IDF: term frequency-inverse document frequency
T2DM: type 2 diabetes mellitus
XGBoost: extreme gradient boosting

Edited by A Coristine; submitted 27.Mar.2026; peer-reviewed by O Tutsoy, H Xu; comments to author 17.Jun.2026; revised version received 02.Oct.2026; accepted 06.Oct.2026; published 10.Nov.2026

Please cite as:
Huang J-L, Yang C-H, Pai T-W
A Disease Trajectory Representation Framework With Risk-Aware Pseudotime Analysis for Electronic Health Records: Application to Heart Failure Risk Stratification
JMIR Med Inform 2026;14:e95929
URL: https://medinform.jmir.org/2026/1/e95929
doi: 10.2196/95929
PMID:

©Jhen-Li Huang, Cing-Han Yang, Tun-Wen Pai. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 10.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX