Recent Advances in Predictive Modeling with Electronic Health Records

Jiaqi Wang[1] , **Junyu Luo**[1] , **Muchao Ye**[1] , **Xiaochen Wang**[1] , **Yuan Zhong**[1] , **Aofei Chang**[1] ,
Guanjie Huang[1] , **Ziyi Yin**[1] , **Cao Xiao**[2] , **Jimeng Sun**[3] and **Fenglong Ma**[1*]
[1]Pennsylvania State University
[2]GE Healthcare
[3]University of Illinois Urbana-Champaign
{jqwang, junyu, muchao, xcwang, yuanzhong, aofei, gzh8, ziyiyin, fenglong}@psu.edu,
cao.xiao@gehealthcare.com, jimeng@illinois.edu

Abstract

The development of electronic health records (EHR) systems has enabled the collection of a vast amount of digitized patient data. However, utilizing EHR data for predictive modeling presents several challenges due to its unique characteristics. With the advancements in machine learning techniques, deep learning has demonstrated its superiority in various applications, including healthcare. This survey systematically reviews recent advances in deep learning-based predictive models using EHR data. Specifically, we introduce the background of EHR data and provide a mathematical definition of the predictive modeling task. We then categorize and summarize predictive deep models from multiple perspectives. Furthermore, we present benchmarks and toolkits relevant to predictive modeling in healthcare. Finally, we conclude this survey by discussing open challenges and suggesting promising directions for future research.

1 Introduction

Advancements in healthcare technology have sparked a revolution in patient information storage. Many health institutions and providers are now adopting electronic health record (EHR) systems to digitize patient information. This enables healthcare professionals to extract valuable insights from EHR data, thereby supporting various aspects of healthcare, ranging from enhancing clinical decision-making to predicting diseases, enabling personalized medicine, driving quality improvement initiatives, and contributing to public health surveillance and epidemiological research.

This survey focuses on **predictive modeling** in healthcare, which refers to using machine learning techniques to analyze patients' historical health data along with current observations to support diagnosis or make predictions about future health events. This approach usually leverages large datasets, often derived from EHR systems, to identify patterns, trends, and relationships that can be used to forecast specific health-related events. Predictive modeling has various applications in healthcare, including but not limited to disease risk assessment, hospital discharge potential or readmission risk, treatment outcomes prediction, and health resource allocation. In other words, predictive modeling can enhance clinical decision-making, improve patient care, and contribute to more efficient and cost-effective healthcare delivery.

However, accurate and reliable predictive modeling in healthcare is challenging due to the unique characteristics of EHR data, summarized as follows:

- **Temporal Dynamics**: The sequentially collected EHR data exhibits temporal dynamics. Changes in a patient's health status over time may not follow a linear pattern, wihle capturing temporal dependencies requires sophisticated modeling techniques.
- **High Dimensionality**: Medical codes in EHR data (e.g., diagnosis, medication, procedure codes) are high-dimensional. When encoded using multihot encoding, the codes that are related to the prediction task are often sparse.
- **Multimodalities and Heterogeneity**: Health data comprises of multiple modalities such as clinical notes, laboratory results, imaging reports, and administrative records. The data could also be assembled from different sources. The heterogeneity of the shapes and sources of the data can bring challenges to data integration, which is a foundational task before predictive modeling.
- **Imbalanced Data**: In healthcare, some patient cohorts could be far smaller than certain majority groups, while certain outcomes or events may be rare, leading to imbalances in the distribution of classes. This can impact the performance of predictive models, making them biased toward the majority class.
- **Clinical Explainability**: Healthcare professionals often require predictive models to be explainable. While effective, many advanced machine learning algorithms may lack explainability, making it challenging for clinicians to understand, trust, and utilize the model results.

In this comprehensive survey, we present a systematic overview of recent advancements in deep learning techniques to enhance the performance of predictive models by addressing the aforementioned challenges. The contributions of this paper can be summarized as follows: (1) We introduce a taxonomy based on the techniques employed in existing deep

[*] Corresponding author.

learning-based predictive modeling approaches, offering a detailed analysis of current methods in Section 3. (2) Common benchmark datasets and toolkits are summarized to provide a valuable resource for researchers and practitioners in Section 4. (3) We discuss open challenges and propose future research directions in the field of predictive modeling in Section 5.

2 Background

EHR Data. Let $\mathcal{D} = \{\mathcal{P}_1, \cdots, \mathcal{P}_N\}$ denote the extracted ataset from an EHR system, where $\mathcal{P}_n$ represents the n-th patient information. EHR data consists of a sequence of a patient's visit information associated with a time stamp, which can be represented as $\mathcal{P}_n = \{(\mathcal{V}_1^n, t_1^n), \cdots, (\mathcal{V}_T^n, t_T^n)\}$, where $\mathcal{V}_i^n$ and t_i^n denotes the visit information and the time information, respectively, and T is the number of visits. Due to the heterogeneity of EHR data, each visit $\mathcal{V}_i^n$ usually contains M modalities, denoted as $\mathcal{V}_i^n = \{\mathcal{M}_{i,j}^n\}_{j=1}^M$, where each modality $\mathcal{M}_{i,j}^n$ can be either demographical information, diagnosis codes, medication codes, procedure codes, images, time-series monitoring data, or even a clinical note.

Predictive Modeling. The predictive modeling task is formulated as follows: Given the historical EHR data of the n-th patient as represented below,

$$\begin{aligned}\mathcal{P}_n &= \{(\mathcal{V}_1^n, t_1^n), \cdots, (\mathcal{V}_T^n, t_T^n)\} \\ &= \{([\mathcal{M}_{1,1}^n, \cdots, \mathcal{M}_{1,M}^n], t_1^n), \cdots, ([\mathcal{M}_{T,1}^n, \cdots, \mathcal{M}_{T,M}^n], t_T^n)\},\end{aligned}$$

the goal is to predict the future health event or outcome y_n after a specified time window (e.g., six months or one year after the last timestamp T). The applications of predictive modeling tasks typically involve binary classification. The cross-entropy (CE) loss is employed to train the models:

$$\mathcal{L} = \sum_{n=1}^{N} \text{CE}(f(\mathcal{P}_n; \Theta), y_n),$$

where $f(\cdot)$ represents the model, and Θ denotes the model parameter set. In the following sections, we will summarize recent advancements in the development of accurate deep learning-based predictive models (i.e., $f(\cdot)$) in healthcare.

3 Existing Progress

We summarize the recent work on predictive modeling from different perspectives, as shown in Table 1. In the following subsections, we provide details of each type of existing work.

3.1 Basic Deep Learning-Based Predictive Models

Recurrent Neural Networks-Based Models. Due to the temporality of EHR data, recurrent neural networks (RNN) are the primary structure choice in deep learning for handling temporality. In general, predictive models of this type regard EHR data as text-like sequence data and utilize RNNs such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs) to propagate each visit information to obtain a comprehensive representation. An early attempt at this design named Retain [Choi *et al.*, 2016b] is proposed by Choi et al., which augments RNNs with an attention mechanism and uses those learned attention scores for interpretability. RNN-based predictive models can even act as a doctor and predict the diagnosis and medication for a future visit given the previous input diagnosis and medication codes [Choi *et al.*, 2016a]. Later, Dipole [Ma *et al.*, 2017] is proposed based on the bidirectional RNN, which aggregates the input EHR data and can effectively attend features in the hierarchy of visits for health risk prediction. These successful early attempts show the effectiveness of the design of combining RNNs with attention mechanisms, and other designs in this fashion come after that, including Timeline [Bai *et al.*, 2018], Stage-Net [Gao *et al.*, 2020] and Health-ATM [Ma *et al.*, 2018c].

Transformer-Based Models. Recent advancements have highlighted the Transformer architecture [Vaswani *et al.*, 2017], which leverages the self-attention mechanism to model relationships within sequential data. Pioneering studies using Transformers in healrhcare such as BEHRT [Li *et al.*, 2020] and HiTANet [Luo *et al.*, 2020], among others [Rasmy *et al.*, 2021; Li *et al.*, 2022; Yang *et al.*, 2023b; Huang *et al.*, 2022], exemplify this trend. BEHRT [Li *et al.*, 2020] innovatively utilizes a transformer instead of the conventional RNN module to process sequential EHR records, though it lacks a hierarchical approach for handling code-visit level data. In contrast, HiTANet [Luo *et al.*, 2020] melds the Transformer architecture with hierarchical modeling. It includes a local evaluation stage, employing a time-aware Transformer for embedding temporal aspects into visit-level data, and a global synthesis stage, leveraging a time-aware key-query attention mechanism for a comprehensive analysis. Adopting Transformer-based models in EHR predictive modeling marks a significant paradigm shift from traditional RNNs. These models offer a more sophisticated approach to managing the intricate and sequential nature of EHR data, demonstrating their immense potential in enhancing healthcare predictive analytics.

3.2 Time-Aware Predictive Modeling

A critical aspect of predictive modeling with EHR data is the integration of temporal information. Contrasting with textual data, the sequence of EHR data hinges on time information, and the intervals between recordings often vary. Accurately modeling this time aspect is essential for evaluating the impact of each patient visit. Addressing this need, T-LSTM [Baytas *et al.*, 2017] was introduced, incorporating an information decay function that acknowledges potential decay in patient information over time gaps between visits. This approach modifies the gates of LSTM to enhance risk prediction accuracy. Building on this concept, several models have emerged. RetainEX [Kwon *et al.*, 2018], an extension of the Retain model, considers information decay and employs traditional attention mechanisms to assign weights to visits and diagnosis codes. TimeLine [Bai *et al.*, 2018] implements self-attention mechanisms to boost performance and integrate the information decay function into patient representation learning. Further advancing this field, HiTANet [Luo *et al.*, 2020] introduces a non-monotonic time attention mechanism tailored to disease-specific time preferences using a learnable time preference embedding. Concare [Ma *et al.*, 2020b] enhances time modeling through a multi-channel time series

Model Name	Venue	Model Focus							Tasks	Datasets
		F1	F2	F3	F4	F5	F6	F7		
NNR [Chen and Cai, 2015]	JMBE							✓	Diabetes	Private
Retain [Choi *et al.*, 2016b]	NeurIPS	✓			✓				Heart Failure	Private
Dipole [Ma *et al.*, 2017]	KDD	✓			✓				Diagnosis Prediction	Private
T-LSTM [Baytas *et al.*, 2017]	KDD	✓							Diabetes Mellitus Parkinson's Progression	EMRBots, PPMI
GRAM [Choi *et al.*, 2017]	KDD	✓		✓	✓				Dignosis Prediction	MIMIC, Private
Timeline [Bai *et al.*, 2018]	KDD	✓			✓				Admission	SEER
Health-ATM [Ma *et al.*, 2018c]	SDM	✓	✓						Congestive Heart Failure	Private, EMRbots
RetainEX [Kwon *et al.*, 2018]	TVCG	✓			✓				Heart Failure Cataract	HIRA
KAME [Ma *et al.*, 2018b]	CIKM	✓		✓	✓				Dignosis Prediction	MIMIC, Private
PRIME [Ma *et al.*, 2018a]	KDD	✓		✓					Heart Failure, COPD kidney Disease	Private
RAIM [Xu *et al.*, 2018]	KDD	✓				✓			Decompensation Length of Stay	MIMIC
DCMN [Feng *et al.*, 2019]	ICDM	✓				✓			Mortality, Cost	MIMIC, Private
K-Boosted C5.0 [Zhang *et al.*, 2019]	JHE							✓	Breast Cancer	Private
Stage-Net [Gao *et al.*, 2020]	WWW	✓							Decompensation Mortality	MIMIC, ESRD
BEHRT [Li *et al.*, 2020]	Scientific Reports	✓			✓				Disease Prediction Diesease Embedding	CPRD
HiTANet [Luo *et al.*, 2020]	KDD	✓	✓		✓				Disease Prediction	Private
Concare [Ma *et al.*, 2020b]	AAAI	✓	✓		✓				Mortality	MIMIC, Private
F-LSTM & F-CNN [Tang *et al.*, 2020]	JAMIA	✓				✓			Acute Renal Failure Shock, Mortality	MIMIC
MaskEHR [Ma *et al.*, 2020a]	SDM	✓						✓	Pastic Quadriplegia CerebralPalsy, Quadriplegia, Diplegic Cerebral Palsy	Private
medGAN [Armanious *et al.*, 2020]	MLHC							✓	Heart Failure	Private, MIMIC
LSAN [Ye *et al.*, 2020]	CIKM	✓			✓				Heart Failure COPD, Kidney Disease	Private
MedPath [Ye *et al.*, 2021a]	WWW	✓		✓	✓				Heart Failure, COPD Kidney Disease	Private
MedRetriever [Ye *et al.*, 2021b]	CIKM	✓		✓	✓				Heart Failure, COPD Kidney Disease	Private
LstmBert [Yang and Wu, 2021]	EMNLP	✓				✓			Acute Renal Failure Diagnoses	MIMIC, Private
MUFASA [Xu *et al.*, 2021]	AAAI	✓				✓	✓		Diagnoses	MIMIC
synTEG [Zhang *et al.*, 2021]	JAMIA	✓						✓	Type-2 diabetes, Heart Failure, Hypertension, COPD	Private
EVA [Biswal *et al.*, 2021]	MLHC	✓						✓	Heart Failure	Private
ConCAD [Huang and Ma, 2021]	ECMLPKDD	✓		✓					Sleep Apnea	Apnea-ECG, MIT-BIH PSG
TContextGGAN [Xu *et al.*, 2022]	TKDE	✓	✓						Heart Failure Sepsis Risk	MIMIC
FedCovid [Wang *et al.*, 2022]	ECMLPKDD	✓				✓		✓	Covid-19 Vaccine Side Effects	Private
AutoMed [Cui *et al.*, 2022]	BIBM	✓					✓		COPD, Amnesia Kidney Disease	Private
promptEHR [Wang and Sun, 2022]	EMNLP	✓				✓		✓	Heart Failure	MIMIC
AccSleepNet [Huang *et al.*, 2022]	BIBM	✓							Sleep Stage Scoring	Sleep-Accel, Newcastle-Accel
TrustSleepNet [Huang *et al.*, 2022]	BHI	✓				✓			Sleep Stage Scoring	SHHS
KerPrint [Yang *et al.*, 2023a]	AAAI	✓		✓	✓				Disease Prediction	MIMIC, Private
MedHMP [Wang *et al.*, 2023]	EMNLP	✓				✓			Acute Renal Failure, Shock, Mortality, Readmission, Heart Falire, COPD, Amnesia	MIMIC, Private
SASMOTE [Kosolwattana *et al.*, 2023]	BioData Mining							✓	Autism Spectrum Disorder, Congenital Heart Disease	Private
TWIN [Das *et al.*, 2023]	KDD	✓				✓		✓	Breast Cancer, Lung Cancer	Project Data Sphere
AutoFM [Cui *et al.*, 2024]	SDM	✓				✓	✓		Acute Renal Failure Diagnoses	MIMIC
MedDiffusion [Zhong *et al.*, 2024]	SDM	✓						✓	COPD,Heart Failure, Amnesia,Kidney	MIMIC, Private

Table 1: Summarization of recent work on predictive modeling in healthcare (in publication years). **F1**: modeling temporality; **F2**: modeling time irregularity; **F3**: incorporating extra knowledge; **F4**: interpretability; **F5**: modeling multimodal EHR data; **F6**: using automated machine learning (AutoML) techniques; and **F7**: addressing class imbalance issues.

embedding, capturing complex temporal relationships. Additionally, T-ContextGGAN [Xu *et al.*, 2022] employs graph neural networks to model intricate time paths on graphs. The evolution of time-aware predictive models in EHR data has significantly progressed from simple time information integration to complex, disease-specific time preference modeling. These developments underscore the importance of temporal dynamics in predictive healthcare analytics, leading to more accurate and personalized patient care predictions.

3.3 Knowledge-Enhanced Predictive Modeling

A critical avenue for enhancing predictive models in the EHR domain lies in integrating additional external knowledge. Below, we categorize the methods for leveraging external knowledge based on the type of knowledge being utilized.

Structured Knowledge. GRAM [Choi *et al.*, 2017] utilizes a graph attention network to integrate hierarchical medical ontologies into the learning process of code representations. Building upon GRAM, KAME [Ma *et al.*, 2018b] advances this concept by embedding ontology information throughout the entire prediction process, thereby enriching the model's contextual understanding. Beyond merely serving as an additional feature, external knowledge can also function as a regularization component. For instance, PRIME [Ma *et al.*, 2018a] generates a posterior regularization term from external knowledge, aligning the model's predictions with established medical insights. This strategy simplifies incorporating external knowledge, making it broadly applicable across various EHR risk prediction frameworks. Besides graph attention, ConCAD [Huang and Ma, 2021] develops a cross-attention mechanism to combine deep representations with expert knowledge.

MedPath [Ye *et al.*, 2021a] introduces a novel approach by retrieving personalized external information from knowledge graphs, thereby enhancing the relevance and utility of the external knowledge. It constructs a personalized ontology map based on patient-specific data, which is then encoded via a graph network for application in predictive tasks. Similarly, KerPrint [Yang *et al.*, 2023a] employs both a personalized local knowledge graph based on temporal data and a global-level knowledge graph, optimizing the usage of external information. Subsequent methodologies extend the application of this external knowledge to diverse scenarios, such as the cold-start setting [Tan *et al.*, 2022] and treatment recommendation systems [Yao *et al.*, 2023].

Unstructured Knowledge. In contrast, some researchers focus on harnessing unstructured data, such as textual information, which presents unique challenges due to its unorganized nature. MedRetriever [Ye *et al.*, 2021b] innovatively addresses this by creating a pool of medical text segments from various online sources. It leverages EHR data and the target disease document embeddings to dynamically retrieve relevant text segments, offering insights into the progression of a disease from symptoms to diagnosis.

3.4 Preditive Modeling with Multimodal Data

Compared to conventional methods that rely solely on a single modality, multimodal approaches leverage diverse clinical modalities as input. This enables them to gather comprehensive information, leading to more promising performance outcomes.

Centralized Learning. The prior study by Tang et al. [Tang *et al.*, 2020] suggests the fusion of demographics and temporal clinical features through concatenation in the early stages, followed by conventional machine learning models for the ultimate prediction. In contrast, DCMN [Feng *et al.*, 2019] employs a dual memory network to concurrently process two input modalities, namely waveform and clinical sequence. The outputs of both memory networks are then summed to facilitate further predictions.

RAIM [Xu *et al.*, 2018] integrates waveform and vital signs as inputs into a multi-channel attention mechanism module. The module subsequently aggregates discrete clinical features into the generated embeddings for subsequent prediction tasks. Another approach, proposed by Yang et al. [Yang and Wu, 2021], focuses on fusing temporal clinical features, time-invariant data, and clinical notes through summation in the output stage. The researchers explore various combinations of modality-specific encoders to achieve competitive prediction performance.

A more recent contribution by Wang et al. [Wang *et al.*, 2023] leverages an attention mechanism to obtain a weighted summation of multiple clinical modalities, including temporal clinical features, diagnosis, medication, clinical notes, and demographics. This approach aims to generate a meaningful multimodal Electronic Health Record (EHR) representation for both pretraining and task-specific fine-tuning.

Federated Learning. To protect EHR data privacy, the multimodal EHR data is usually stored in local servers and not shared with others. Thus, collaborative training of predictive models without sharing EHR data is a practical and challenging task. FedCovid [Wang *et al.*, 2022], a federated learning framework, is developed to adaptively fuse heterogeneous EHR data at local client sites for predicting COVID-19 vaccine side effects. Besides, [Sachin *et al.*, 2023] introduces a multimodal contrastive federated learning framework for digital healthcare, employing a geometric multimodal contrastive representation method to enhance the representation of various modalities in a shared space, thus improving intermodal relationship capture and overall model performance.

3.5 AutoML-Based Predictive Modeling

EHR typically encompasses structured and unstructured data with sparse and irregular longitudinal features. In multimodal fusion, the challenge lies in determining how to effectively fuse different modalities, a problem that often relies on manual modeling and intuition. We have encountered difficulties choosing between early, hybrid, and late fusion approaches. Recently, automated machine learning (AutoML)-based approaches have been proposed to search for optimal fusion strategies automatically.

MUFASA [Xu *et al.*, 2021] is the first work that applied the neural architecture search (NAS) technique to medical multimodal fusion. It jointly optimizes multimodal fusion strategies and modality-specific architectures. MUFASA delineates two types of blocks: modality-specific block, which only allows inputs from the same modality, and fusion block, which accepts inputs from both fusion architecture hidden states and modality-specific states. AutoMed [Cui *et al.*, 2022] employs a general Directed Acyclic Graph (DAG) framework, opting for a cell-based search where each cell consists of a sequentially ordered set of computation nodes. It uses the same search space design for modality encoding cells and fusion cells. Another noteworthy work, AutoFM [Cui *et al.*, 2024], also focuses on modality-specific search and multimodal fusion search. This approach designs separate search spaces for static and sequential modalities. Furthermore, it incorporates modality interaction operations within these modality-specific search spaces to facilitate early interactions. During the fusion search phase, a specialized set of simple fusion operations is devised to enable effective feature fusion.

3.6 Predictive Modeling with Imbalanced Classes

Addressing the issue of class imbalance is another critical challenge in healthcare predictive modeling. Such imbalance can significantly degrade the performance of predictive models. To tackle this issue, a variety of strategies have been developed and extensively researched. In this section, we systematically categorize and review recent methodologies, providing a structured overview of the techniques employed to effectively manage class imbalance in healthcare datasets.

Oversampling and Undersampling Techniques. Oversampling techniques are crucial for balancing datasets, effectively increasing the size of the minority class. SASMOTE [Kosolwattana *et al.*, 2023] that builds on the Synthetic Minority Oversampling Technique (SMOTE) proves effective in enhancing model accuracy for imbalanced gene risk and heart disease datasets. Furthermore, deep learning-based resampling NNR [Chen and Cai, 2015] shows great success in imbalanced diabetes prediction. Compared to oversampling, undersampling works by selecting informative samples near the boundary of the class. For instance, K-Boosted C5.0 [Zhang *et al.*, 2019] utilizes K-means to segregate the majority and minority classes and select an equal amount of patients per cluster for Breast Cancer prediction.

Generative Techniques. Generative models, primarily developed for enhancing entire datasets, can be specifically tailored to augment minority data. MaskEHR [Ma *et al.*, 2020a], medGAN [Armanious *et al.*, 2020], and synTEG [Zhang *et al.*, 2021] employ Generative Adversarial Networks (GANs) for generating synthetic data. EVA [Biswal *et al.*, 2021] and TWIN [Das *et al.*, 2023] use Variational Autoencoders (VAEs) in their approach. MedDiffusion [Zhong *et al.*, 2024] leverages the capabilities of Denoising Diffusion Probabilistic Models (DDPMs). Additionally, promptEHR [Wang and Sun, 2022] utilizes a language model for similar purposes. These models use sequential learning techniques like RNNs and Transformers to add historical context, which maintains the temporal coherence of generated data. By incorporating this augmented minority class data into the training sets, predictive models benefit significantly, exhibiting enhanced performance and more balanced outcomes.

3.7 Interpretable Predictive Modeling

Interpretability is an important property of health risk prediction models because it is related to the life and death of patients when they are applied in real life. Interpretability helps doctors and patients understand the reasoning behind the predictive model and trust its risk prediction or treatment recommendation. Existing interpretability mechanisms designed in the health risk prediction models can be divided into three categories.

Attention-Based Interpretation. The first type of interpretability method utilizes attention weights to explain the importance of different diagnosis codes or visits. This method has been widely used for RNN-based methods [Choi *et al.*, 2016b; Ma *et al.*, 2017; Bai *et al.*, 2018; Ma *et al.*, 2018c] as we have mentioned above. Attention mechanisms can also be incorporated into Transformer-based predictive models such as HiTANet [Luo *et al.*, 2020] and LSAN [Ye *et al.*, 2020]. Take LSAN as an example for illustration. It introduces a hierarchical attention mechanism that assigns flexible attention weights to different diagnosis codes by their relevance to corresponding diseases in the diagnosis code level and pays greater attention to visits with higher relevance in the visit level. Such design provides a fine-grained interpretability. The attention weights in the hierarchy of diagnosis codes and visits can be interpreted into which symptoms and visits are paid more attention to for health risk prediction.

Personalized Knowledge Graph-Based Interpretation. A weakness of attention-based interpretation is that they cannot explicitly express the reasoning path of the health risk prediction models. To tackle that, later research proposes the idea of using a medical knowledge graph to provide explicit reasoning for interpretation. A representative method for interpretability in this type is named MedPath [Ye *et al.*, 2021a]. MedPath uses SemMed, a large-scale online medical knowledge graph, to extract personalized knowledge graphs (PKGs) containing all possible disease progression paths from observed symptoms to target diseases. The extracted PKGs can show how the existing observed symptoms from a patient can gradually lead to the target disease, which is more explicit compared to attention weights. Other health risk prediction models using knowledge graphs include the ones designed by [Yang *et al.*, 2023a; Lyu *et al.*, 2023; Xu *et al.*, 2022]. These methods have shown that knowledge graphs are good candidates for improving performance and interpretability.

Medical Text-Based Explicit Interpretation. Another way to provide explicit interpretability is to use unstructured medical text, first proposed in MedRetriever [Ye *et al.*, 2021b]. MedRetriever creates a pool of candidate medical text segments from online medical text sources and uses the EHR and target disease document embeddings to dynamically retrieve the related ones that can explain the disease progression from the symptoms to the target disease. By reading the retrieved medical text in the end, both patients and doctors can obtain an easy-to-understand interpretation.

Uncertainty-Based Interpretation. Although the aforementioned interpretation methods can provide a decent ex-

Name	Data Type	# of Data	Modalities	Link
MIMIC-III	Real	38,597 patients	Demographics, vital signs, medications, laboratory measurements, observations and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data	https://physionet.org/content/mimiciii/1.4/
MIMIC-IV	Real	40,000+ patients	Demographics, vital signs, medications, laboratory measurements, observations and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data	https://physionet.org/content/mimiciv/2.2/
MIMIC-CXR	Real	377,110 images 227,835 reports	Electronic health record data, images (chest radiographs), and natural language (free-text reports)	https://physionet.org/content/mimic-cxr/2.0.0/
eICU	Real	200,000+ admissions	Vital sign measurements, care plan documentation, severity of illness measures, diagnosis information, treatment information, and more	https://physionet.org/content/mimic-cxr/2.0.0/
PPMI	Real	2,230 patients	Subject characteristics, biospecimen, images, medical history, etc.	https://www.ppmi-info.org/
ADNI	Real	2,775 patients	Subject characteristics, genetic data, images, medical history, neuropathology, etc.	https://adni.loni.usc.edu/
Apnea-ECG	Real	70 recordings	Subject characteristics, electrocardiogram	https://physionet.org/content/apnea-ecg/1.0.0/
MIT-BIH PSG	Real	18 recordings	Subject characteristics, electrocardiogram, electroencephalography, electrooculography, electromyography, etc.	https://physionet.org/content/slpdb/1.0.0/
SHHS	Real	6,441 patients	Subject characteristics, electrocardiogram, electroencephalography, electrooculography, electromyography, airflow, etc.	https://sleepdata.org/datasets/shhs
Newcastle-Accel	Real	28 patients	Subject characteristics, acceleration, polysomnography.	https://zenodo.org/records/1160410#.YLqiSC1h1qt
Sleep-Accel	Real	31 patients	Acceleration, heart rate, steps.	https://physionet.org/content/sleep-accel/1.0.0/
EMRBOTS	Synthetic	100,000 patients	Patients' admissions, demographics, socioeconomics, labs, medications, etc.	http://www.emrbots.org/
Project Data Sphere	Real	242 studies	Data provider, sponsor, study phase, linked data, tumor type, access, etc.	https://www.projectdatasphere.org

Table 2: Commonly used publically available health datasets.

planation about how the prediction models work, they still cannot simply say "I do not know," when they are uncertain in their predictions. However, it is particularly important in clinically relevant tasks. Knowing how confident the prediction is, the clinical experts can trust the results and interpretations with high confidence and set the ones with low confidence aside for another manual inspection. To address this issue, [Huang and Ma, 2022] proposes to use evidential deep learning to quantify the uncertainty. Besides predicting the target class, it also predicts the density of the probability of each class, which can derive associated uncertainty to express how confident the prediction is.

4 Benchmarks and Toolkits

Publicly Available Healthcare Databases. We present a summary of commonly used publicly available health datasets in Table 2. In addition to these widely utilized datasets, certain country-specific datasets play a crucial role in predictive modeling within the healthcare domain. For instance, the Clinical Practice Research Datalink (CPRD, https://cprd.com/) dataset is a comprehensive, anonymized database of primary care records from the UK. The Health Insurance Review & Assessment (HIRA) Database (https://www.hira.or.kr/eng/main.do) is a comprehensive medical database in South Korea, which holds extensive healthcare information for the entire South Korean population.

Private Healthcare Databases. Most extensive healthcare databases operate on a private basis, accessible only to select researchers who can extract pertinent data, as outlined in Table 1. Notably, TriNetX (https://trinetx.com/) stands out as a notable example, representing a comprehensive dataset that seamlessly integrates real-time access to longitudinal clinical data with cutting-edge analytics. This integration serves to optimize various aspects, including protocol design, feasibility assessments, site selection, patient recruitment, and the generation of real-world evidence. It is noteworthy that the Diamond Network within TriNetX encompasses data from 92 global institutions, capturing information from an expansive cohort of 213,167,071 patients. Similarly, the IBM MarketScan Research Database mirrors TriNetX in functionality, offering de-identified, longitudinal patient-level claims, and specialty data. This resource boasts coverage for over 273 million unique patients, providing a rich source for diverse healthcare analyses.

Toolkits. PyHealth (https://pyhealth.readthedocs.io/en/latest/) is a comprehensive deep-learning toolkit for predictive modeling, which integrates diverse EHR datasets, such as MIMC, eICU, and all OMOP-CDM-based databases, and several deep learning algorithms for multiple health-related tasks, e.g., patient hospitalization prediction, mortality prediction, and ICU length stay forecasting. This toolkit allows the users to customize their own pipelines by following the 5 stages: load dataset, define task function, build deep learning models, model training, and inference. There are also several other toolkits or open-source codes for specific uses such as MedCAT (https://medcat.readthedocs.io/en/latest/), Fasten (https://github.com/fastenhealth), MONAI (https://monai.io/) and NiftyNet (https://niftynet.io/).

5 Open Challenges and Future Directions

In this section, we delve into various open challenges and propose future research directions across key aspects, including model trustworthiness, data characteristics, model training, collaborative learning paradigms.

5.1 Trustworthy Predictive Modeling

Contemporary predictive modeling techniques predominantly prioritize enhancing model performance. While performance is undoubtedly pivotal, equal emphasis must be placed on trustworthiness, encompassing accuracy, reliability, ethics, and transparency. Trustworthy predictive models play a crucial role in supporting healthcare professionals in making informed decisions, improving patient outcomes, and optimizing healthcare processes. Despite its paramount importance, the exploration of trustworthy predictive models in healthcare remains an underexplored domain.

In light of this, we outline the following research directions: (1) *LLM-driven interpretable model design*. Our prior work [Wang and Sun, 2022] demonstrates that LLMs encapsulate medical knowledge and possess the potential to interpret model outputs in human-understandable lay language, thereby enhancing model reliability. (2) *Ethical model design*. Given the diversity within EHR databases, designing fair and robust models while ensuring patient privacy is imperative. A promising research direction involves the development of unbiased and robust predictive models. (3) *Human-in-the-loop learning*. Existing models primarily rely on data, often neglecting the vital input of healthcare professionals. A human-in-the-loop approach advocates for the inclusion of domain expertise, enabling clinicians to provide feedback and contributing to the trustworthiness of the model.

These research directions collectively pave the way for the development of predictive models that not only excel in performance but also uphold the principles of interpretability, ethics, and collaboration with healthcare professionals, fostering trust and reliability in healthcare applications.

5.2 Data Scarcity/Sparsity

Training deep learning models in the healthcare domain often demands a substantial volume of data. However, as indicated in Table 2, publicly available healthcare data is often limited, especially for rare diseases or specific patient populations. Compounding this issue, EHR data frequently features missing values, leading to substantial data loss during preprocessing. Furthermore, inherent quality challenges in EHR data, such as inaccuracies, duplications, or inconsistencies, further exacerbate the difficulties associated with working with sparse and scarce data. Consequently, acquiring an ample dataset for model training becomes a formidable challenge. To address this fundamental challenge, one potential solution involves the generation of synthetic EHR data using innovative techniques. Despite recent proposals of models utilizing VAE [Das *et al.*, 2023; Biswal *et al.*, 2021], GAN [Armanious *et al.*, 2020; Zhang *et al.*, 2021; Ma *et al.*, 2020a], diffusion-based models [Zhong *et al.*, 2024], and LLM [Wang and Sun, 2022], these approaches still grapple with the limitation of producing realistic EHR data, specifically in simulating the unique characteristics inherent in EHR datasets.

5.3 Pretraining across Multiple Data Sources

Representation learning is a fundamental task in healthcare, particularly in scenarios where labeled data is scarce. Unsupervised or self-supervised learning techniques become essential for accurately capturing health features. While models like GRAM [Choi *et al.*, 2017] and KAME [Ma *et al.*, 2018b] have been proposed to enhance code representations, they often overlook the multimodal and hierarchical characteristics inherent in EHR data. In our ongoing work, we are engaged in training a multimodal model named MedHMP [Wang *et al.*, 2023] on the MIMIC-III dataset. However, this model currently covers only a subset of modalities, leaving others unexplored. To address this limitation and enhance representation learning, a promising avenue for future research involves collaborative pre-training with multiple healthcare datasets from distinct sources. By leveraging shared modalities across these datasets, we can amass sufficient training data to establish a health-specific pre-trained model. This approach holds the potential to comprehensively address the challenge of representation learning across diverse healthcare data sources, accounting for their multimodal and

5.4 Federated Training for Foundation Models

While the transfer of knowledge from a large pre-trained foundation model significantly enhances performance, this approach remains largely unexplored in the context of training foundation models for healthcare. The challenges arise due to the sensitive and private nature of healthcare data, making it infeasible to centrally train such a model. To address these challenges, a potential solution involves the utilization of advanced federated learning techniques. This approach allows for the collaborative training of a foundation model without necessitating the sharing of stakeholders' private data. Federated learning thus presents a promising avenue for efficiently training healthcare-specific foundation models while respecting privacy constraints.

Nevertheless, the direct application of federated learning in healthcare encounters challenges, primarily stemming from the significant heterogeneity in the distribution of healthcare data, as illustrated in our prior work [Wang *et al.*, 2022]. Addressing the substantial imbalance and heterogeneity in EHR data across clients represents an open challenge, thereby establishing a promising research direction for the application of federated learning in healthcare.

Acknowledgements

This work is partially supported by the National Science Foundation under Grant No. 2238275 and the National Institutes of Health under Grant No. R01AG077016.

References

[Armanious *et al.*, 2020] Karim Armanious, Chenming Jiang, Marc Fischer, Thomas Küstner, Tobias Hepp, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. Medgan: Medical image translation using gans. *Computerized medical imaging and graphics*, 79:101684, 2020.

[Bai *et al.*, 2018] Tian Bai, Shanshan Zhang, Brian L Egleston, and Slobodan Vucetic. Interpretable representation learning for healthcare via capturing disease progression through time. In *KDD*, pages 43–51, 2018.

[Baytas *et al.*, 2017] Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. Patient subtyping via time-aware lstm networks. In *KDD*, pages 65–74, 2017.

[Biswal *et al.*, 2021] Siddharth Biswal, Soumya Ghosh, Jon Duke, Bradley Malin, Walter Stewart, Cao Xiao, and Jimeng Sun. Eva: Generating longitudinal electronic health records using conditional variational autoencoders. In *MLHC*, pages 260–282, 2021.

[Chen and Cai, 2015] Long-Sheng Chen and Sheng-Jhe Cai. Neural-network-based resampling method for detecting diabetes mellitus. *JMBE*, 35:824–832, 2015.

[Choi *et al.*, 2016a] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *MLHC*, pages 301–318, 2016.

[Choi *et al.*, 2016b] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *NeurIPS*, pages 3504–3512, 2016.

[Choi *et al.*, 2017] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. Gram: graph-based attention model for healthcare representation learning. In *KDD*, pages 787–795, 2017.

[Cui *et al.*, 2022] Suhan Cui, Jiaqi Wang, Xinning Gui, Ting Wang, and Fenglong Ma. Automed: Automated medical risk predictive modeling on electronic health records. In *BIBM*, pages 948–953, 2022.

[Cui *et al.*, 2024] Suhan Cui, Jiaqi Wang, Yuan Zhong, Han Liu, Ting Wang, and Fenglong Ma. Automated fusion of multimodal electronic health records for better medical predictions. In *SDM*, pages 361–369, 2024.

[Das *et al.*, 2023] Trisha Das, Zifeng Wang, and Jimeng Sun. Twin: Personalized clinical trial digital twin generation. In *KDD*, pages 402–413, 2023.

[Feng *et al.*, 2019] Yujuan Feng, Zhenxing Xu, Lin Gan, Ning Chen, Bin Yu, Ting Chen, and Fei Wang. Dcmn: Double core memory network for patient outcome prediction with multimodal data. In *ICDM*, pages 200–209, 2019.

[Gao *et al.*, 2020] Junyi Gao, Cao Xiao, Yasha Wang, Wen Tang, Lucas M Glass, and Jimeng Sun. Stagenet: Stage-aware neural networks for health risk prediction. In *WWW*, pages 530–540, 2020.

[Huang and Ma, 2021] Guanjie Huang and Fenglong Ma. Concad: Contrastive learning-based cross attention for sleep apnea detection. In *ECML-PKDD*, pages 68–84, 2021.

[Huang and Ma, 2022] Guanjie Huang and Fenglong Ma. Trustsleepnet: A trustable deep multimodal network for sleep stage classification. In *BHI*, pages 01–04, 2022.

[Huang *et al.*, 2022] Guanjie Huang, Ye Yuan, Guohong Cao, and Fenglong Ma. Accsleepnet: An axis-aware hybrid deep fusion model for sleep stage classification using wrist-worn accelerometer data. In *BIBM*, pages 1005–1012, 2022.

[Kosolwattana *et al.*, 2023] Tanapol Kosolwattana, Chenang Liu, Renjie Hu, Shizhong Han, Hua Chen, and Ying Lin. A self-inspected adaptive smote algorithm (sasmote) for highly imbalanced data classification in healthcare. *BioData Mining*, 16(1):15, 2023.

[Kwon *et al.*, 2018] Bum Chul Kwon, Min-Je Choi, Joanne Taery Kim, Edward Choi, Young Bin Kim, Soonwook Kwon, Jimeng Sun, and Jaegul Choo. Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE TVCG*, 25(1):299–309, 2018.

[Li *et al.*, 2020] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020.

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

[Luo *et al.*, 2020] Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In *KDD*, pages 647–656, 2020.

[Lyu *et al.*, 2023] Kewei Lyu, Yu Tian, Yong Shang, Tianshu Zhou, Ziyue Yang, Qianghua Liu, Xi Yao, Ping Zhang, Jianghua Chen, and Jingsong Li. Causal knowledge graph construction and evaluation for clinical decision support of diabetic nephropathy. *JBI*, 139:104298, 2023.

[Ma *et al.*, 2017] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *KDD*, pages 1903–1911, 2017.

[Ma *et al.*, 2018a] Fenglong Ma, Jing Gao, Qiuling Suo, Quanzeng You, Jing Zhou, and Aidong Zhang. Risk prediction on electronic health records with prior medical knowledge. In *KDD*, pages 1910–1919. ACM, 2018.

[Ma *et al.*, 2018b] Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. Kame:

Knowledge-based attention model for diagnosis prediction in healthcare. In *CIKM*, pages 743–752, 2018.

[Ma *et al.*, 2018c] Tengfei Ma, Cao Xiao, and Fei Wang. Health-atm: A deep architecture for multifaceted patient health record representation and risk prediction. In *SDM*, pages 261–269, 2018.

[Ma *et al.*, 2020a] Fenglong Ma, Yaqing Wang, Jing Gao, Houping Xiao, and Jing Zhou. Rare disease prediction by generating quality-assured electronic health records. In *SDM*, pages 514–522, 2020.

[Ma *et al.*, 2020b] Liantao Ma, Chaohe Zhang, Yasha Wang, Wenjie Ruan, Jiantao Wang, Wen Tang, Xinyu Ma, Xin Gao, and Junyi Gao. Concare: Personalized clinical feature embedding via capturing the healthcare context. In *AAAI*, volume 34, pages 833–840, 2020.

[Rasmy *et al.*, 2021] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.

[Sachin *et al.*, 2023] DN Sachin, B Annappa, Sateesh Ambasange, and Alan E Tony. A multimodal contrastive federated learning for digital healthcare. *SN Computer Science*, 4(5):674, 2023.

[Tan *et al.*, 2022] Yanchao Tan, Carl Yang, Xiangyu Wei, Chaochao Chen, Weiming Liu, Longfei Li, Jun Zhou, and Xiaolin Zheng. Metacare++: Meta-learning with hierarchical subtyping for cold-start diagnosis prediction in healthcare data. In *SIGIR*, pages 449–459, 2022.

[Tang *et al.*, 2020] Shengpu Tang, Parmida Davarmanesh, Yanmeng Song, Danai Koutra, Michael W Sjoding, and Jenna Wiens. Democratizing ehr analyses with fiddle: a flexible data-driven preprocessing pipeline for structured clinical data. *JAMIA*, 27(12):1921–1934, 2020.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[Wang and Sun, 2022] Zifeng Wang and Jimeng Sun. Promptehr: Conditional electronic healthcare records generation with prompt learning. In *EMNLP*, 2022.

[Wang *et al.*, 2022] Jiaqi Wang, Cheng Qian, Suhan Cui, Lucas Glass, and Fenglong Ma. Towards federated covid-19 vaccine side effect prediction. In *ECML-PKDD*, pages 437–452. Springer, 2022.

[Wang *et al.*, 2023] Xiaochen Wang, Junyu Luo, Jiaqi Wang, Ziyi Yin, Suhan Cui, Yuan Zhong, Yaqing Wang, and Fenglong Ma. Hierarchical pretraining on multimodal electronic health records. In *EMNLP*, pages 2839–2852, 2023.

[Xu *et al.*, 2018] Yanbo Xu, Siddharth Biswal, Shriprasad R Deshpande, Kevin O Maher, and Jimeng Sun. Raim: Recurrent attentive and intensive model of multimodal patient monitoring data. In *KDD*, pages 2565–2573, 2018.

[Xu *et al.*, 2021] Zhen Xu, David R So, and Andrew M Dai. Mufasa: Multimodal fusion architecture search for electronic health records. In *AAAI*, volume 35, pages 10532–10540, 2021.

[Xu *et al.*, 2022] Yuyang Xu, Haochao Ying, Siyi Qian, Fuzhen Zhuang, Xiao Zhang, Deqing Wang, Jian Wu, and Hui Xiong. Time-aware context-gated graph attention network for clinical risk prediction. *IEEE TKDE*, 2022.

[Yang and Wu, 2021] Bo Yang and Lijun Wu. How to leverage multimodal ehr data for better medical predictions? In *EMNLP*, pages 4029–4038, 2021.

[Yang *et al.*, 2023a] Kai Yang, Yongxin Xu, Peinie Zou, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. Kerprint: local-global knowledge graph enhanced diagnosis prediction for retrospective and prospective interpretations. In *AAAI*, volume 37, pages 5357–5365, 2023.

[Yang *et al.*, 2023b] Zhichao Yang, Avijit Mitra, Weisong Liu, Dan Berlowitz, and Hong Yu. Transformehr: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nature Communications*, 2023.

[Yao *et al.*, 2023] Zijun Yao, Bin Liu, Fei Wang, Daby Sow, and Ying Li. Ontology-aware prescription recommendation in treatment pathways using multi-evidence healthcare data. *ACM TIS*, 41(4):1–29, 2023.

[Ye *et al.*, 2020] Muchao Ye, Junyu Luo, Cao Xiao, and Fenglong Ma. Lsan: Modeling long-term dependencies and short-term correlations with hierarchical attention for risk prediction. In *CIKM*, pages 1753–1762, 2020.

[Ye *et al.*, 2021a] Muchao Ye, Suhan Cui, Yaqing Wang, Junyu Luo, Cao Xiao, and Fenglong Ma. Medpath: Augmenting health risk prediction via medical knowledge paths. In *WWW*, pages 1397–1409, 2021.

[Ye *et al.*, 2021b] Muchao Ye, Suhan Cui, Yaqing Wang, Junyu Luo, Cao Xiao, and Fenglong Ma. Medretriever: Target-driven interpretable health risk prediction via retrieving unstructured medical text. In *CIKM*, pages 2414–2423, 2021.

[Zhang *et al.*, 2019] Jue Zhang, Li Chen, Fazeel Abid, et al. Prediction of breast cancer from imbalance respect using cluster-based undersampling method. *Journal of healthcare engineering*, 2019, 2019.

[Zhang *et al.*, 2021] Ziqi Zhang, Chao Yan, Thomas A Lasko, Jimeng Sun, and Bradley A Malin. Synteg: a framework for temporal structured electronic health data simulation. *JAMIA*, 28(3):596–604, 2021.

[Zhong *et al.*, 2024] Yuan Zhong, Suhan Cui, Jiaqi Wang, Xiaochen Wang, Ziyi Yin, Yaqing Wang, Houping Xiao, Mengdi Huai, Ting Wang, and Fenglong Ma. Meddiffusion: Boosting health risk prediction via diffusion-based data augmentation. In *SDM*, pages 499–507, 2024.