Self-Explaining Hypergraph Neural Networks for Diagnosis Prediction

Leisheng Yu   LEISHENG.YU@RICE.EDU
Rice University, United States

Yanxiao Cai   YC139@RICE.EDU
Rice University, United States

Minxing Zhang   MINXING.ZHANG@DUKE.EDU
Duke University, United States

Xia Hu   XIA.HU@RICE.EDU
Rice University, United States

Abstract

The burgeoning volume of electronic health records (EHRs) has enabled deep learning models to excel in predictive healthcare. However, for high-stakes applications such as diagnosis prediction, model interpretability remains paramount. Existing deep learning diagnosis prediction models with intrinsic interpretability often assign attention weights to every past diagnosis or hospital visit, providing explanations lacking flexibility and succinctness. In this paper, we introduce SHy, a self-explaining hypergraph neural network model, designed to offer personalized, concise and faithful explanations that allow for interventions from clinical experts. By modeling each patient as a unique hypergraph and employing a message-passing mechanism, SHy captures higher-order disease interactions and extracts distinct temporal phenotypes as personalized explanations. It also addresses the incompleteness of the EHR data by accounting for essential false negatives in the original diagnosis record. A qualitative case study and extensive quantitative evaluations on two real-world EHR datasets demonstrate the superior predictive performance and interpretability of SHy over existing state-of-the-art models. The code is available at https://github.com/ThunderbornSakana/SHy.

Data and Code Availability This study employs the MIMIC-III (Johnson et al., 2016b) and MIMIC-IV (Johnson et al., 2023) datasets, both available on the PhysioNet repository (Johnson et al., 2016a).

Institutional Review Board (IRB) This research does not require IRB approval.

1. Introduction

Electronic health records (EHRs) are large-scale chronologies of patients' hospital visits, encapsulating their longitudinal healthcare experiences (Si et al., 2021). The surge in EHR data has fostered the development of deep learning models for tasks like diagnosis prediction, mortality prediction, and drug recommendation (Xu et al., 2023; Song et al., 2018; Tan et al., 2022). Among these, diagnosis prediction based on longitudinal EHRs is of particular significance, as it directly correlates with health risk identification and the quality of personalized healthcare (Zhang et al., 2019; Van Smeden et al., 2021). As shown in Figure 1, longitudinal EHR data comprise sequentially time-stamped hospital visits, each being an unordered set of diagnoses (Zhang et al., 2020a). The aim of diagnosis prediction is to forecast potential diagnoses for a patient's subsequent visit based on historical records.
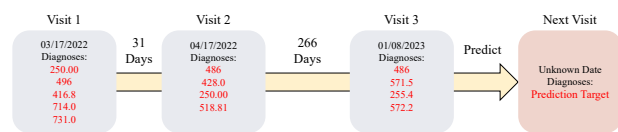

Figure 1: An illustration of diagnosis prediction using the longitudinal EHR of a patient. Diagnoses are denoted by ICD-9 codes.

Because the outcomes of diagnosis prediction are communicated to both clinicians and patients, model

interpretability is crucial (Tonekaboni et al., 2019; Meng et al., 2022). Concerns that post-hoc explanations for black-box deep learning models may not faithfully reflect their actual reasoning processes (Rudin, 2019) have spurred research into inherently interpretable models for high-stakes EHR-based applications. These models provide explanations primarily by assigning an attention weight to each historical diagnosis or visit (Zhang et al., 2020b). This method of explanation has four major limitations:

- **Insufficient personalization:** patients receive a uniform explanation format with importance weights attached to every diagnosis or visit, making it harder for clinicians to discern individual health conditions at first glance compared to customized formats that directly visualize individual comorbidity patterns.

- **Lack of conciseness:** with each diagnosis receiving a weight, size of the explanation equals the input size, failing to filter out redundant information.

- **Not robust against false negative diagnoses:** since patients may not exhibit symptoms for certain chronic conditions, a visit record may be incomplete. Limiting explanations to documented diagnoses can miss false negatives critical for the final prediction.

- **Difficulty in intervention:** models that do not base predictions solely on learned attention weights hinder domain experts from directly influencing the final prediction by editing explanations.

To address these challenges, we introduce SHy, utilizing temporal phenotypes extracted from the diagnosis history as explanations. Temporal phenotypes, representing observable, informative, and interpretable patterns, can depict the progression of patients' health conditions (Che et al., 2015). Different from Liu et al. (2015), which treats each patient as a temporal graph, SHy represents each patient as a hypergraph and extracts sub-hypergraphs as temporal phenotypes, greatly reducing the space complexity. Personalization is enhanced through employing hypergraph neural networks to model higher-order disease interactions within and across visits, leading to tailored disease embeddings that capture individual comorbidity patterns. Moreover, inspired by existing work on hypergraph structure learning (Cai et al., 2022a), SHy adds additional diagnosis-visit pairs to

the original hypergraph based on embedding similarity, before extracting a collection of phenotypes using the Gumbel-Softmax trick (Jang et al., 2017; Maddison et al., 2017). By enforcing novel regularization, SHy produces succinct, non-overlapping, and faithful temporal phenotypes, each reflecting an essential aspect of a patient's evolving health status. Instead of receiving the identical explanation format, different patients now have distinct sets of temporal phenotypes as customized explanations. Since the extracted phenotypes are the basis for the subsequent prediction process, SHy is a concept bottleneck model (Koh et al., 2020), allowing domain experts to intervene in the generated explanations to optimize performance or study model behaviors.

We quantitatively assess both predictive performance and explanation quality, demonstrating that SHy generates accurate predictions while providing superior explanations. The main contributions of our work are as follows:

- We propose SHy, a self-explaining diagnosis prediction model that extracts personalized temporal phenotypes as explanations. This concept bottleneck model, representing patients as hypergraphs, enhances robustness against false negatives and allows domain experts to edit explanations.

- SHy introduces a novel combination of objectives that ensures the extracted phenotypes are concise, non-overlapping, and faithful.

- We perform extensive experiments on two real-world EHR datasets, validating SHy's superior predictive performance and interpretability. A case study with clinical experts shows how the explanations generated by SHy can be edited.

2. Related Work

2.1. Deep Learning on Longitudinal EHRs

Applying deep learning models to longitudinal EHRs for predictive tasks centers around learning adequate patient representation (Choi et al., 2018, 2016b; Tan et al., 2024). To model temporal disease progression patterns, Doctor AI (Choi et al., 2016a) leveraged recurrent neural networks (RNNs), StageNet (Gao et al., 2020) employed a stage-aware long short-term memory (LSTM) module, and Hi-BEHRT (Li et al., 2023) adopted a hierarchical Transformer effective for long visit histories. To account for irregular time gaps

between visits, variants of LSTM and self-attention mechanisms that consider timestamps have been introduced (Ma et al., 2020b; Baytas et al., 2017; Ren et al., 2021). To utilize external medical knowledge, methods such as GRAM (Choi et al., 2017; Peng et al., 2021a,b; Ma et al., 2018b; Song et al., 2019) infused information from medical ontologies into representation learning via attention mechanisms; PRIME (Ma et al., 2018a) incorporated rule-based prior medical knowledge through posterior regularization; and SeqCare (Xu et al., 2023) employed online medical knowledge graphs with adaptive graph structure learning. To address low-quality data, methods such as GRASP (Zhang et al., 2021), which utilized knowledge from similar patients through graph neural networks, and MedSkim (Cui et al., 2022), which filtered out noisy diagnoses using the Gumbel-Softmax trick, were developed. Existing approaches tried to enhance model interpretability by giving weights to every past diagnosis or visit: methods such as RETAIN achieved this with attention mechanisms (Choi et al., 2016c; Ma et al., 2017; Luo et al., 2020; Bai et al., 2018; Zhang et al., 2018; Qiao et al., 2018; Lu et al., 2021b; Cai et al., 2022c); AdaCare (Ma et al., 2020a) employed a scale-adaptive feature recalibration module. To mitigate data insufficiency in certain scenarios, Rasmy et al. (2021) pre-trained BERT on a large EHR corpus and fine-tuned it on smaller datasets.

2.2. Intrinsically Interpretable Models

Intrinsic interpretability, which means that predictive models provide their own explanations, is favored over post-hoc interpretability that necessitates a separate model for explaining a black-box model, because the explanations provided by intrinsically interpretable models are exploited in the decision-making process and faithful to model predictions (Du et al., 2019; Rudin, 2019). SENN (Alvarez Melis and Jaakkola, 2018) is a class of self-explaining neural networks whose interpretability is enforced via regularization. A self-explaining deep learning model proposed by Li et al. (2018) utilizes an autoencoder and a prototype classifier network to provide case-based rationales. The attention mechanism has been employed to achieve intrinsic interpretability by using attention weights as explanations (Mohankumar et al., 2020), but its reliability is arguable (Serrano and Smith, 2019; Wiegreffe and Pinter, 2019; Jain and Wallace, 2019). SITE (Wang and Wang, 2021) emphasizes robust interpretations equivariant

to geometric transformations. ProtoVAE (Gautam et al., 2022) leverages a variational autoencoder to learn class-specific prototypes. The Bort optimizer (Zhang et al., 2023) enhances model explainability by imposing boundedness and orthogonality constraints on model parameters. Intrinsic interpretability has been studied in healthcare (Payrovnaziri et al., 2020). However, unlike SHy, existing self-explaining deep learning models for EHR-based diagnosis prediction fail to provide temporal explanations that reflect the distinct comorbidities of individual patients.

The related work on deep learning on hypergraphs is discussed in Appendix A.

3. Methodology

An overview of the proposed model, SHy, is illustrated in Figure 2. Before delving into the details of the method, it is essential to formulate the problem of diagnosis prediction using longitudinal EHR data.
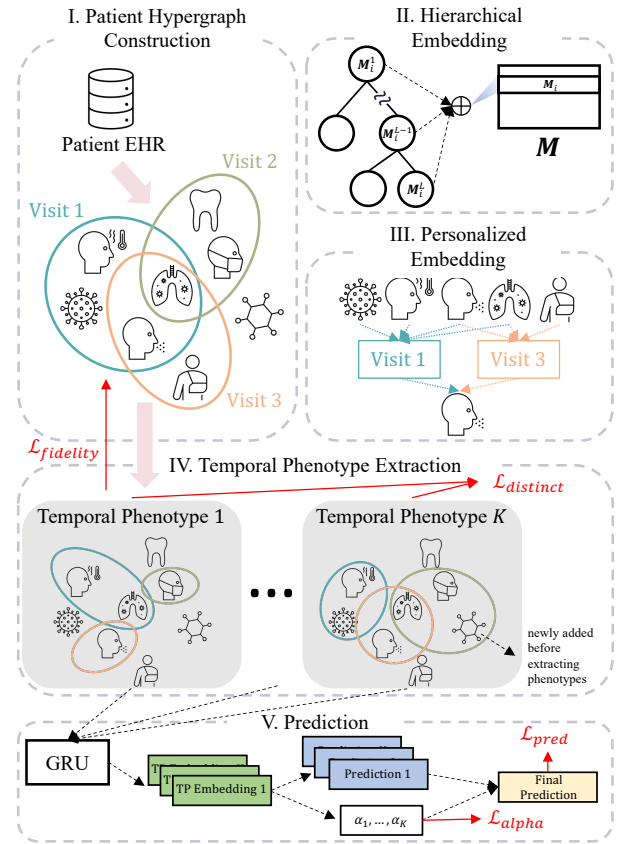

Figure 2: Overview of SHy.

3.1. Problem Formulation

An EHR dataset contains various diseases, each assigned a diagnosis code according to the International Classification of Diseases, Ninth Revision (ICD-9). ICD-9 codes have a hierarchical structure. For example, both *acute respiratory failure* (ICD-9 code 518.81), and *chronic respiratory failure* (518.83) are children of *other diseases of lung* (518.8). The set of all unique diagnosis codes within an EHR dataset is denoted as $c_1, c_2, ..., c_{|\mathcal{C}|} \in \mathcal{C}$, and $|\mathcal{C}|$ indicates the size of this set. Assuming that we have N patients in a longitudinal EHR dataset, we can represent the n-th patient as a sequence, $[(\boldsymbol{e_1^n}, t_1^n), (\boldsymbol{e_2^n}, t_2^n), ..., (\boldsymbol{e_{T^{(n)}}^n}, t_{T^{(n)}}^n)]$, where $T^{(n)}$ is the number of past visits, $\boldsymbol{e_j^n}$ denotes the diagnosis codes recorded in the j-th visit, and t_j^n signifies the timestamp. $\boldsymbol{e_j^n} \in \{0,1\}^{|\mathcal{C}|}$ is a multi-hot vector, where the i-th entry equals 1 if visit j includes c_i. The task of diagnosis prediction is to predict the diagnoses in the next visit, $\boldsymbol{e_{T^{(n)}+1}^n}$, based on past visits. To simplify notation, the superscript n is dropped in later sections, where the context allows.

3.2. Personalized Disease Representation Learning

Before detailing how SHy learns personalized diagnosis code embeddings to reflect different comorbidities, we outline how it models hierarchical disease relationships, a common practice in related work (Ma et al., 2018b; Qiao et al., 2020; Peng et al., 2021a,b).

3.2.1. LEARNING HIERARCHICAL DISEASE EMBEDDINGS

ICD-9 organizes diseases into a tree structure with each child node having a single parent. This medical ontology can be used to enhance the representation learning process for diagnosis codes, defining relative disease distances. Thus, previous work has explored effective modeling of this hierarchical relationship, a type of external medical knowledge (Choi et al., 2017; Lu et al., 2021b; Song et al., 2019). As the primary focus of our work is not developing new strategies to utilize ICD-9, we adopt the hierarchical embedding module from CGL (Lu et al., 2021a). Specifically, the ICD-9 tree comprises L levels (we set $L = 4$ following CGL), and most diagnosis codes in the EHR dataset are leaf nodes. We accommodate non-leaf nodes in the dataset by creating virtual child nodes, and padding them to level L. Each node

at every level gets an initialized embedding, resulting in L embedding matrices, $\{\mathbf{M}^1, \mathbf{M}^2, ..., \mathbf{M}^L\}$. Here, $\mathbf{M}^l \in \mathbb{R}^{n_l \times d_c}$ is the embedding matrix for the l-th level, with n_l and d_c denoting the number of nodes at the level l and the embedding dimension, respectively. Then, for example, the ontology-aware embedding, $\mathbf{M}_i \in \mathbb{R}^{L d_c}$, for c_i is obtained via concatenating its embeddings with those of its ancestors:

$$\mathbf{M}_i = \mathbf{M}_i^1 \parallel \mathbf{M}_i^2 \parallel ... \parallel \mathbf{M}_i^L. \quad (1)$$

Simply put, after this stage, an embedding table, $\mathbf{M} \in \mathbb{R}^{|\mathcal{C}| \times L d_c}$, is initialized as learnable parameters for all diagnosis codes in the EHR dataset.

3.2.2. CONSTRUCTING PATIENT HYPERGRAPHS

To model individual comorbidities for personalized patient representation, we should represent each patient as an independent entity and effectively capture disease interactions. One intuitive method is to represent patients as ordinary graphs where diseases form nodes and co-occurrences create edges. However, this strategy requires the construction of an adjacency matrix of size $|\mathcal{C}| \times |\mathcal{C}|$ for every patient. Since an EHR dataset typically contains thousands of unique diagnosis codes and tens of thousands of patients, this leads to prohibitively high space complexity. Moreover, an ordinary graph only models pairwise interactions, but the co-occurrence of multiple diseases in a single visit suggests non-pairwise relationships. Hypergraphs, as an alternative to ordinary graphs, can address these issues. They can represent higher-order relations because hyperedges in hypergraphs can connect any number of nodes. Therefore, we model each patient as a hypergraph, with diseases as nodes and hospital visits as hyperedges. In this way, higher-order interactions among diseases can be captured. Each patient hypergraph, $\mathcal{G} = (\mathcal{C}, \mathcal{E})$, can be represented as an incidence matrix $\mathbf{P} \in \{0,1\}^{|\mathcal{C}| \times T}$, where $\mathbf{P}_{ij} = 1$ if the j-th visit contains c_i, and each hyperedge $\boldsymbol{e_j} \in \mathcal{E}$ is a subset of $\mathcal{C}$. Since the average number of hospital visits of patients in typical EHRs is below 10, modeling patients as hypergraphs greatly reduces space complexity.

3.2.3. MODELING INDIVIDUAL COMORBIDITIES

To model distinctive comorbidities and thus learn personalized diagnosis code embeddings, SHy conducts message passing on the constructed patient hypergraphs. With numerous hypergraph neural net-

work architectures in the existing literature, we experimented with several state-of-the-art models, the results of which are discussed in Appendix B. We empirically selected UniGIN (Huang and Yang, 2021) as SHy's message passing mechanism. UniGIN generalizes Graph Isomorphism Networks (GIN) (Xu et al., 2018) to hypergraphs by formulating it as a two-stage aggregation process. Specifically, SHy first obtains visit embeddings by aggregating the embeddings of diagnosis codes within the respective visit:

$$\mathbf{V}_j^{(z)} = \frac{1}{|\boldsymbol{e_j}|} \sum_{i \in \boldsymbol{e_j}} \mathbf{M}_i^{(z)}, \qquad (2)$$

where $\mathbf{V}_j^{(z)}$, $|\boldsymbol{e_j}|$, and z denote the embedding of the j-th visit, number of diseases within the j-th visit, and the index of UniGIN layers, respectively. We set $\mathbf{M}_i^{(0)} = \mathbf{M}_i$. Next, SHy updates diagnosis embeddings by aggregating the embeddings of visits containing the corresponding diagnosis:

$$\mathbf{M}_i^{(z+1)} = \sigma(\mathbf{W}_{\text{UniGIN}}^{(z)}((1+\varepsilon)\mathbf{M}_i^{(z)} + \sum_{j \in \mathcal{E}_i} \mathbf{V}_j^{(z)})), \quad (3)$$

where σ is Leaky ReLU, $\mathbf{W}_{\text{UniGIN}}^{(z)} \in \mathbb{R}^{d_c^{(z+1)} \times d_c^{(z)}}$ denotes learnable weights of the z-th UniGIN layer, ε is a learnable parameter, and $\mathcal{E}_i$ represents the set of indices of visits including c_i. By stacking Z layers of message passing mechanisms on individual patient hypergraphs, SHy adeptly models complex interactions among diseases within the Z-hop neighborhood. Thus, disease embeddings for a specific patient are strongly influenced by the combinations of diseases that this patient had within and across visits.

In summary, after this stage, an updated personalized embedding table of diagnosis codes, $\mathbf{M}^{(Z)} \in \mathbb{R}^{|\mathcal{C}| \times d_c^{(Z)}}$, is obtained for each patient. The superscript (Z) will be omitted in subsequent sections.

3.3. Temporal Phenotype Extraction & Modeling

Since the diagnosis history of a patient possibly suffers from incompleteness, SHy introduces false negative disease-visit pairs into the constructed patient hypergraph prior to phenotype extraction. These additions are based on the multi-head weighted cosine similarity between nodes and hyperedges:

$$\mathbf{S}_{ij} = \frac{1}{n_s} \sum_{k=1}^{n_s} \frac{(\boldsymbol{\Phi}_k \odot \mathbf{M}_i) \cdot (\boldsymbol{\Phi}_k \odot \mathbf{V}_j)}{\|\boldsymbol{\Phi}_k \odot \mathbf{M}_i\| \|\boldsymbol{\Phi}_k \odot \mathbf{V}_j\|}, \qquad (4)$$

where $\mathbf{S}_{ij}$ denotes the similarity score between the embeddings of c_i and the j-th visit, $\boldsymbol{\Phi} \in \mathbb{R}^{n_s \times d_c^{(Z)}}$ represents a stack of n_s independent learnable weight vectors, and $\mathbf{V}_j$ is obtained through Equation (2). To avoid adding redundant connections, SHy enforces $\mathbf{S}_{ij} = 0$ if c_i was originally included in the j-th visit. Next, SHy derives a supplementary hypergraph, $\Delta\mathbf{P} \in \{0,1\}^{|\mathcal{C}| \times T}$, based on the similarity scores:

$$\Delta\mathbf{P}_{ij} = \begin{cases} 1, & \mathbf{S}_{ij} \in \texttt{topk}(\mathbf{S}, p| \sum_{j=1}^{T} \boldsymbol{e_j}|) \\ 0, & \mathbf{S}_{ij} \notin \texttt{topk}(\mathbf{S}, p| \sum_{j=1}^{T} \boldsymbol{e_j}|), \end{cases} \quad (5)$$

where p is the ratio of the number of connections in the supplementary hypergraph to those in the initial patient hypergraph. Then, an updated patient hypergraph, $\tilde{\mathbf{P}} = \mathbf{P} + \Delta\mathbf{P}$, is produced. SHy essentially augments the original patient records with additional disease-visit pairs possessing the highest similarity scores.

SHy extracts K temporal phenotypes, each being a unique subgraph of the updated patient hypergraph. Specifically, to extract Phenotype k, SHy learns a binary matrix denoted by $\boldsymbol{\Gamma}^k \in \{0,1\}^{|\mathcal{C}| \times T}$. Each entry, $\boldsymbol{\Gamma}_{ij}^k$, serves as a masking factor for $\tilde{\mathbf{P}}_{ij}$ and is a random variable following a Bernoulli distribution parameterized by a probability weight, $\mathbf{O}_{ij}^k$. This weight is derived from the embeddings of the corresponding disease and visit:

$$\mathbf{O}_{ij}^k = \text{MLP}([\mathbf{M}_i \,\|\, \mathbf{V}_j]). \qquad (6)$$

To allow backpropagation while producing the discrete binary matrix $\boldsymbol{\Gamma}^k$, SHy employs the Gumbel-Softmax trick:

$$\boldsymbol{\Gamma}_{ij}^k = \sigma\left(\frac{\log(\frac{\mathbf{O}_{ij}^k}{1-\mathbf{O}_{ij}^k}) + (\delta^0 - \delta^1)}{\tau}\right), \qquad (7)$$

where $\delta^0, \delta^1 \sim \text{Gumbel}(0,1)$, σ is the sigmoid function, and τ is the temperature. With the generated binary matrix, SHy extracts Phenotype k through

$$\boldsymbol{\Psi}^k = \tilde{\mathbf{P}} \odot \boldsymbol{\Gamma}^k, \qquad (8)$$

where $\boldsymbol{\Psi}^k \in \{0,1\}^{|\mathcal{C}| \times T}$ is the incidence matrix of Phenotype k. The process, from Equation (6) to Equation (8), constitutes a temporal phenotype extractor. SHy utilizes K independent temporal phenotype extractors, operating concurrently to generate $\{\boldsymbol{\Psi}^1, \boldsymbol{\Psi}^2, ..., \boldsymbol{\Psi}^K\}$.

3.4. Prediction and Objectives

To predict subsequent diagnoses, SHy embeds K temporal phenotypes using a Gated Recurrent Unit (GRU) and location-based attention mechanism:

$$\mathbf{V}^k = \mathbf{\Psi}^{k\top}\mathbf{M} \tag{9}$$

$$\mathbf{H}_1^k, \mathbf{H}_2^k, ..., \mathbf{H}_T^k = \text{GRU}(\mathbf{V}_1^k, \mathbf{V}_2^k, ..., \mathbf{V}_T^k) \tag{10}$$

$$\boldsymbol{\alpha}^k = \text{Softmax}(\text{MLP}(\mathbf{H}^k)) \tag{11}$$

$$\mathbf{U}_k = \boldsymbol{\alpha}^k \mathbf{H}^k, \tag{12}$$

where $\mathbf{V}^k \in \mathbb{R}^{T \times d_c^{(Z)}}$ is a stack of embeddings of visits in Phenotype k, $\mathbf{H}_t^k \in \mathbb{R}^{d_{hid}}$ denotes the hidden state corresponding to the t-th visit, $\boldsymbol{\alpha}^k \in \mathbb{R}^T$ represents weights of hidden states, and $\mathbf{U}_k \in \mathbb{R}^{d_{hid}}$ is the embedding of Phenotype k. Next, SHy derives importance weights, $\boldsymbol{\alpha} \in \mathbb{R}^K$, for all phenotypes with a self-attention mechanism:

$$head_i = \text{Attention}(\mathbf{U}\mathbf{W}_i^Q, \mathbf{U}\mathbf{W}_i^K, \mathbf{U}\mathbf{W}_i^V) \tag{13}$$

$$\boldsymbol{\alpha} = \text{Softmax}([head_1 \parallel head_2 \parallel ... \parallel head_{n_h}]\mathbf{w}^O). \tag{14}$$

Here, $\mathbf{W}_i^Q \in \mathbb{R}^{d_{hid} \times d_Q}$, $\mathbf{W}_i^K \in \mathbb{R}^{d_{hid} \times d_Q}$, $\mathbf{W}_i^V \in \mathbb{R}^{d_{hid} \times d_V}$, and $\mathbf{w}^O \in \mathbb{R}^{d_V}$ are learnable parameters, and $\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_K}})\mathbf{V}$, where d_K denotes the dimensionality of $\mathbf{K}$. The final prediction, $\hat{\boldsymbol{y}} \in \mathbb{R}^{|\mathcal{C}|}$, is a weighted sum of predictions from K phenotypes:

$$\hat{\boldsymbol{y}} = \boldsymbol{\alpha}(\text{Softmax}(\mathbf{U}\mathbf{W} + \boldsymbol{b})), \tag{15}$$

where $\mathbf{W} \in \mathbb{R}^{d_{hid} \times |\mathcal{C}|}$ and $\mathbf{b} \in \mathbb{R}^{|\mathcal{C}|}$ denote learnable parameters. The prediction loss for all patients is calculated using cross-entropy:

$$\mathcal{L}_{pred} = -\frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{y_n}^\top \log(\hat{\boldsymbol{y}_n}) + (1 - \boldsymbol{y_n})^\top \log(1 - \hat{\boldsymbol{y}_n})),$$

where $\boldsymbol{y_n} = \boldsymbol{e_{T^{(n)}+1}^n}$. The K temporal phenotypes, along with their respective importance weights, act as model explanations. To ensure that the extracted phenotypes faithfully preserve relevant input information, SHy employs a GRU decoder, the details of which are introduced in Appendix C, to reconstruct the original patient hypergraph—a sequence of multi-hot vectors—from the concatenated phenotype embeddings. The fidelity of explanations is enhanced by penalizing the reconstruction error:

$$\mathcal{L}_{fidelity} = -\frac{1}{N}\sum_{n=1}^{N}\frac{1}{T^{(n)}|\mathcal{C}|}\sum_{j=1}^{T^{(n)}}\sum_{i=1}^{|\mathcal{C}|}(\mathbf{P}_{ij}^n \cdot \log \hat{\mathbf{P}}_{ij}^n$$
$$+ (1 - \mathbf{P}_{ij}^n) \cdot \log(1 - \hat{\mathbf{P}}_{ij}^n)),$$

where $\hat{\mathbf{P}}^n$ denotes the reconstructed hypergraph for the n-th patient. Moreover, the K temporal phenotypes need to be non-overlapping in order to be meaningful, i.e., different phenotypes include different diagnoses for a given visit. SHy achieves distinctiveness by penalizing common diagnoses across K binary masks for one visit:

$$\mathbf{B}_j = [\mathbf{\Gamma}_{\cdot j}^1, \mathbf{\Gamma}_{\cdot j}^2, ..., \mathbf{\Gamma}_{\cdot j}^K] \in \mathbb{R}^{|\mathcal{C}| \times K}$$

$$\mathcal{L}_{distinct} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{T^{(n)}}\sum_{j=1}^{T^{(n)}}\|\mathbf{I}_{T^{(n)}} - \mathbf{B}_j^{n\top}\mathbf{B}_j^n\|_2,$$

where $\mathbf{I}_{T^{(n)}}$ denotes an identity matrix, $\mathbf{\Gamma}_{\cdot j}^k$ represents the j-th column of the binary mask $\mathbf{\Gamma}^k$, and $\mathbf{B}_j$ specifies the diagnoses each masking matrix retains for the j-th visit. The aim is to encourage columns of $\mathbf{B}_j$ to form an orthonormal set. Since $\mathbf{B}_j$ is a binary matrix, columns being orthonormal vectors implies that there is no overlap across the K phenotypes for the j-th visit and each phenotype only includes one diagnosis for this visit. Thus, $\mathcal{L}_{distinct}$ also pushes the phenotypes to be concise. Lastly, SHy enforces regularization to prevent the distribution of the K attention weights from being uniform or too extreme, ensuring clear relative importance among phenotypes and no extracted phenotype receiving 0 as the weight:

$$\mathcal{L}_{alpha} = -\frac{1}{N}\sum_{n=1}^{N}(\sqrt{\frac{\sum_k(\boldsymbol{\alpha}_k^n - \bar{\boldsymbol{\alpha}^n})^2}{K}} - \|\boldsymbol{\alpha}^n\|_2),$$

where $\boldsymbol{\alpha}^n$ denotes attention weights for the phenotypes of the n-th patient. The final loss function is as follows:

$$\mathcal{L} = \mathcal{L}_{pred} + \epsilon \cdot \mathcal{L}_{fidelity} + \eta \cdot \mathcal{L}_{distinct} + \omega \cdot \mathcal{L}_{alpha}, \tag{16}$$

where ϵ, η, and ω are hyperparameters. This weighted combination enables SHy to provide distinct and concise temporal phenotypes, with each elucidating a unique aspect of the final prediction faithfully. A salient feature of SHy is that it is a concept bottleneck model, meaning that it derives a set of phenotypes and utilizes them to predict the target. This attribute allows clinical practitioners to edit the extracted temporal phenotypes, with changes subsequently propagated to the final prediction via Equation (9) through Equation (15).

4. Experiments

In this section, we evaluate SHy against state-of-the-art diagnosis prediction models in terms of predictive performance and interpretability, and assess the effectiveness of its novel components. We further demonstrate SHy's robustness against false negative diagnoses and its explanation capabilities, highlighting how domain experts can seamlessly intervene in the prediction process.

4.1. Data Description & Experimental Setup

We utilize the MIMIC-III and MIMIC-IV datasets for our experiments, both of which are de-identified and publicly-available collections of electronic health records associated with patients admitted to the Beth Israel Deaconess Medical Center (Goldberger et al., 2000; Johnson et al., 2016b). As observed in Table 5, MIMIC-IV notably possesses a greater number of diagnosis codes and patients.

Our evaluation metrics for diagnosis prediction are Recall@k and nDCG@k. We employ three metrics for a quantitative evaluation of explanation quality. Faithfulness measures the Pearson correlation between importance weights and changes in prediction upon removal of the corresponding explanation units from the input. Ranging from -1 to 1, a high value indicates that the provided explanations can faithfully reflect the model's reasoning process (Alvarez Melis and Jaakkola, 2018). Complexity counts the number of diagnoses in the given explanation: a smaller number implies greater conciseness. Distinctness, exclusive to evaluating SHy's explanations, gauges the overlap among temporal phenotypes:

$$\text{Distinctness} = \frac{1}{N} \sum_{n=1}^{N} \frac{\|\boldsymbol{\Psi}^{n,1} + \boldsymbol{\Psi}^{n,2} + ... + \boldsymbol{\Psi}^{n,K}\|_0}{\|\boldsymbol{\Psi}^{n,1} + \boldsymbol{\Psi}^{n,2} + ... + \boldsymbol{\Psi}^{n,K}\|},$$

where $\boldsymbol{\Psi}^{n,k}$ denotes the k-th phenotype for the n-th patient. A higher Distinctness value implies that SHy produces more varied phenotypes for an individual patient. Details on baselines and hyperparameters are introduced in Appendix D.

4.2. Predictive Performance Comparison

For both MIMIC-III and MIMIC-IV, where the average number of diagnoses per visit ranges between 10 and 20, we set $k = \{10, 20\}$ for Recall@k and nDCG@k. Table 1 compares the performance of SHy with baseline models for diagnosis prediction,

along with the count of learnable parameters for each model. Overall, SHy delivers strong predictive performance, surpassing most baselines across all metrics on both datasets. In particular, SHy outperforms ConCare and Doctor AI, two black-box models, by up to 37.44% in Recall@10 (MIMIC-III) and 64.43% in nDCG@20 (MIMIC-IV), respectively. In contrast, SHy's edge over T-LSTM isn't as pronounced, especially on MIMIC-IV, which can be attributed to T-LSTM's much larger number of parameters. The large model size of T-LSTM can bring high expressiveness given ample training data. When juxtaposed with interpretable models, SHy's superiority is even more pronounced: outperforming RETAIN, Dipole, and Timeline by up to 70.53% in Recall@20 on MIMIC-IV, 66.84% in nDCG@10 on MIMIC-IV, and 28.50% in nDCG@10 on MIMIC-III, respectively. It is intriguing to note that SHy's margin over AdaCare is much narrower (4.39% in Recall@20 on MIMIC-III), with the difference reducing further on MIMIC-IV. This is potentially due to AdaCare's larger size.

Although SHy adopts a hierarchical disease embedding approach from CGL, it outclasses CGL by 5.86% in Recall@10 on MIMIC-IV. This can be ascribed to CGL's reliance on a disease co-occurrence graph, which lacks the capability to capture higher-order disease interactions in the manner hypergraphs can. However, modeling patients as hypergraphs is insufficient. The essence lies in performing message passing to learn personalized disease embeddings that reflect individual comorbidities. This claim can be supported by CGL's superior performance over SHy w/o. MP across all metrics.

The other interesting observation is that on MIMIC-III, SHy outperforms SHy ($K = 1$) by up to 4.61% in Recall@20, but on MIMIC-IV, SHy ($K = 1$) surpasses the standard SHy in certain metrics. To unravel this discrepancy, we note that SHy ($K = 1$) can only extract one temporal phenotype, rendering the objectives $\mathcal{L}_{distinct}$ and $\mathcal{L}_{alpha}$ inapplicable. Therefore, SHy ($K = 1$) is trained with only two objectives, $\mathcal{L}_{pred}$ and $\mathcal{L}_{fidelity}$. Similarly, SHy outperforms SHy w/o. $\mathcal{L}_{distinct,alpha}$ on MIMIC-III across all metrics, but is surpassed by SHy w/o. $\mathcal{L}_{distinct,alpha}$ on MIMIC-IV. These findings suggest that the regularization we enforce for enhanced interpretability can be particularly beneficial with smaller training datasets, helping the model achieve better performance by mitigating the risk of overfitting. In contrast, with voluminous training data, these added objectives can actually act as a constraint, limit-

ing the predictive power of the model. Supporting this observation, SHy (only $\mathcal{L}_{pred}$) consistently, albeit marginally, outshines the standard SHy on MIMIC-IV, demonstrating the inherent trade-off between model interpretability and prediction accuracy, while on MIMIC-III their performances are largely indistinguishable.

4.3. Robustness Against False Negatives

To quantitatively evaluate whether SHy can effectively deal with incomplete patient EHRs, we randomly mask out 25% and 75% of input diagnoses in the test data to simulate varying levels of data incompleteness and assess the impact on SHy's performance relative to three top-performing baselines. From Table 2, we observe that SHy consistently surpasses all baselines across all metrics, affirming its robustness against false negative diagnoses. Notably, SHy exhibits the least performance degradation, verifying its efficacy in contexts with incomplete data. Additionally, SHy's phenotypes, despite their conciseness, recovered an average of 4.83%, 0.74%, 10.14%, and 1.67% of the masked diagnoses across the four different settings, demonstrating SHy's adeptness at handling data incompleteness. The inclusion of $\mathcal{L}_{distinct}$, which encourages the explanations to be concise, results in false negatives that are not critical in predicting future diagnoses being excluded from the extracted phenotypes, leading to a modest recovery rate. This also shows that with $\mathcal{L}_{distinct}$, the likelihood of SHy adding true negatives is low.

4.4. Evaluation of Model Explanations

Table 3 highlights the superior quality of explanations provided by SHy compared to other self-explaining diagnosis prediction models. Notably, while all baseline models have Faithfulness scores below 0.5, suggesting a weak to medium correlation between generated weights and prediction changes upon removal of the explanation units, SHy showcases a much stronger correlation. Since RETAIN, AdaCare, and Timeline offer explanations by assigning a weight to every diagnosis, their Complexity score equals to the average number of historical diagnoses per patient in MIMIC-III. Dipole assigns visit-level attention scores, making the generated explanations too coarse to be eligible for calculating Complexity. Although SHy provides explanations by extracting multiple phenotypes (subgraphs of the patient hypergraph) and introduces false negative disease-visit pairs before extraction, the

total number of diagnoses in all phenotypes is, on average, lower than number of historical diagnoses the patient has in the record. This indicates that SHy's phenotype extraction module effectively filters out irrelevant noise, yielding concise explanations.

A peculiar observation is the superior quality of explanations by SHy w/o. $\mathcal{L}_{alpha}$, based on the three metrics. We looked into the explanations it offered, and found that for most of the patients, the generated $\boldsymbol{\alpha}$ were $\{1.0, 0.0, 0.0, 0.0, 0.0\}$, and only one of the five temporal phenotypes was non-empty. Based on how Faithfulness and Distinctness are calculated, it is reasonable for SHy w/o. $\mathcal{L}_{alpha}$ to achieve outstanding results in these two metrics. SHy w/o. $\mathcal{L}_{fidelity,alpha}$ shows a similar behavior. Thus, we can understand the importance of $\mathcal{L}_{alpha}$ in preventing the attention scores from being too extreme.

Comparing SHy with SHy w/o. $\mathcal{L}_{fidelity}$, SHy w/o. $\mathcal{L}_{distinct}$ with SHy w/o. $\mathcal{L}_{fidelity,distinct}$, SHy w/o. $\mathcal{L}_{alpha}$ with SHy w/o. $\mathcal{L}_{fidelity,alpha}$, and SHy w/o. $\mathcal{L}_{distinct,alpha}$ with SHy (only $\mathcal{L}_{pred}$) reveals that the inclusion of $\mathcal{L}_{fidelity}$ consistently enhances Faithfulness, albeit modestly. Moreover, removing $\mathcal{L}_{distinct}$ leads to substantial degradation across all metrics. The uptick in Complexity and downturn in Distinctness underscore its significance in generating concise, non-overlapping explanations. The pronounced decline in Faithfulness when excluding $\mathcal{L}_{distinct}$ stems from near-identical phenotypes having identical weights, making the calculation of correlation coefficients highly unstable. In conclusion, results from Table 3 affirm the importance of all three interpretability objectives, with $\mathcal{L}_{distinct}$ and $\mathcal{L}_{alpha}$ being particularly crucial.

4.5. Case Study

To showcase SHy's capabilities in accommodating interventions from domain experts, we conduct a case study on a patient with three historical visits. Figure 3 illustrates this case study. SHy accurately predicted 7 out of 9 diagnoses for the next visit. In particular, SHy anticipated the onset of a urinary tract infection, despite the lack of information on the timing of the next visit, and in the absence of any previous records explicitly indicating urinary tract problems for the patient. Additionally, SHy identified acidosis, a condition that was not recorded in the patient's last three visits. Thus, SHy demonstrated capabilities beyond merely replicating past diagnoses. Interestingly, SHy's prediction listed diabetes, which

Table 1: Results of Predictive Performance Evaluation across Two EHR Datasets

Model	Recall@10	Recall@20	nDCG@10	nDCG@20	# Params	Recall@10	Recall@20	nDCG@10	nDCG@20	# Params
			MIMIC-III					MIMIC-IV		
Doctor AI	0.2154	0.3021	0.3353	0.3282	2.90M	0.1811	0.2490	0.2591	0.2578	4.71M
RETAIN	0.2024	0.2958	0.3136	0.3167	1.47M	0.1713	0.2504	0.2529	0.2605	2.37M
Dipole	0.2022	0.2965	0.3130	0.3170	2.80M	0.1721	0.2514	0.2539	0.2615	4.16M
T-LSTM	0.2758	0.3735	0.4132	0.4055	30.38M	0.3328	0.4268	0.4208	0.4232	48.49M
GRAM	0.2219	0.3308	0.3489	0.3551	1.62M	0.1796	0.2515	0.2542	0.2611	2.08M
CGL	0.2698	0.3679	0.4125	0.4022	1.52M	0.3159	0.4198	0.4100	0.4180	2.90M
Timeline	0.2153	0.3090	0.3218	0.3239	1.20M	0.2879	0.3898	0.3651	0.3748	1.84M
AdaCare	0.2662	0.3670	0.3970	0.3935	16.43M	0.3323	0.4259	0.4197	0.4221	42.98M
StageNet	0.2174	0.3113	0.3245	0.3270	4.82M	0.3013	0.4003	0.3760	0.3850	6.63M
ConCare	0.2019	0.2937	0.3066	0.3103	2.63M	0.2715	0.3599	0.3372	0.3456	3.19M
Chet	0.2155	0.3102	0.3259	0.3296	2.12M	0.1814	0.2566	0.2609	0.2671	3.48M
SETOR	0.2266	0.3317	0.3503	0.3568	10.02M	0.1810	0.2532	0.2556	0.2648	12.78M
MIPO	0.2251	0.3321	0.3499	0.3589	12.13M	0.1819	0.2545	0.2555	0.2651	14.92M
SHy ($K = 1$)	0.2660	0.3662	0.4006	0.3965	2.00M	0.3340	0.4272	0.4238	<u>0.4257</u>	2.89M
SHy (only $\mathcal{L}_{pred}$)	**0.2780**	<u>0.3793</u>	**0.4139**	<u>0.4079</u>	2.69M	**0.3401**	**0.4344**	**0.4304**	**0.4318**	3.59M
SHy w/o. $\mathcal{L}_{distinct,alpha}$	0.2693	0.3684	0.3998	0.3970	2.69M	<u>0.3353</u>	<u>0.4272</u>	<u>0.4252</u>	0.4255	3.59M
SHy w/o. MP	0.2648	0.3650	0.4021	0.3960	2.45M	0.3221	0.4172	0.4093	0.4131	3.48M
SHy	<u>0.2775</u>*	**0.3831***	<u>0.4135</u>	**0.4088***	2.69M	0.3344*	0.4270	0.4236*	0.4239	3.59M

a. Bolded values highlight the best performance, while underlined values denote the second-best.
b. SHy ($K = 1$) refers to the SHy variant extracting only one temporal phenotype per patient.
c. SHy (only $\mathcal{L}_{pred}$) indicates SHy trained without objectives related to interpretability.
d. SHy w/o. MP represents the SHy variant without message passing on patient hypergraphs.
e. An * indicates that the superiority of SHy over the best-performing baseline is statistically significant.

Table 2: Robustness against False Negatives under Four Settings

Model	Recall@20	nDCG@20	Recall@20	nDCG@20	Recall@20	nDCG@20	Recall@20	nDCG@20
	MIMIC-III, masking 25%		MIMIC-III, masking 75%		MIMIC-IV, masking 25%		MIMIC-IV, masking 75%	
CGL	0.3448 (−6.28%)	0.3792 (−5.73%)	0.2649 (−28.01%)	0.2882 (−28.33%)	0.3774 (−10.09%)	0.3644 (−12.83%)	0.2168 (−48.35%)	0.2003 (−52.08%)
T-LSTM	0.3503 (−6.20%)	0.3828 (−5.61%)	0.2715 (−27.31%)	0.2934 (−27.66%)	0.3919 (−8.18%)	0.3943 (−6.82%)	0.3165 (−25.85%)	0.2342 (−44.67%)
AdaCare	0.3443 (−6.18%)	0.3740 (−4.95%)	0.2890 (−21.25%)	0.3173 (−19.37%)	0.3816 (−10.39%)	0.3659 (−13.31%)	0.1749 (−58.94%)	0.1489 (−64.71%)
SHy	**0.3604*** (−5.92%)	**0.3941*** (−3.58%)	**0.3020*** (−21.17%)	**0.3298*** (−19.32%)	**0.3945*** (−7.61%)	**0.4150*** (−2.10%)	**0.3213*** (−24.75%)	**0.3293*** (−22.31%)

a. The percentages reflect the relative performance drop from the setting with no past diagnoses masked.

Table 3: Evaluation of Explanation Quality

Model	Faithfulness	Complexity	Distinctness
		MIMIC-III	
RETAIN	0.1867	17.6564	–
Dipole	0.3028	–	–
AdaCare	0.2084	17.6564	–
Timeline	0.2499	17.6564	–
SHy w/o. $\mathcal{L}_{fidelity}$	0.5206	11.1233	0.7470
SHy w/o. $\mathcal{L}_{distinct}$	0.1139	97.9960	0.2398
SHy w/o. $\mathcal{L}_{alpha}$	**0.7381**	<u>5.0899</u>	**1.0000**
SHy w/o. $\mathcal{L}_{fidelity,distinct}$	0.0789	98.1879	0.2420
SHy w/o. $\mathcal{L}_{fidelity,alpha}$	<u>0.7269</u>	**5.0155**	**1.0000**
SHy w/o. $\mathcal{L}_{distinct,alpha}$	0.0834	103.2148	0.2295
SHy (only $\mathcal{L}_{pred}$)	0.0801	103.6230	0.2293
SHy	0.5915	10.1839	<u>0.7749</u>

a. "–" means unavailable.

was not confirmed by the ground-truth label. However, considering the patient's recent diagnosis of cirrhosis of liver—a condition affecting insulin sensitivity—and the chronic nature of diabetes mellitus, this prediction still holds merit.

SHy offered compelling rationales for its predictions. **Each extracted phenotype reflected a unique facet of the patient's health status.** Phenotype 1 indicated that the patient might have experienced endocrine system abnormalities, possibly even severe hypoglycemia, due to diabetes; Phenotype 2, which consistently highlighted pneumonia at all visits, symbolized worsening respiratory health and a compromised immune system; Phenotype 3 underlined pneumonia complications and congestive heart failure, suggesting a high probability of hypoxia; Phenotype 4 indicated significant liver impairment and resulting coronary artery lesions; Pheno-
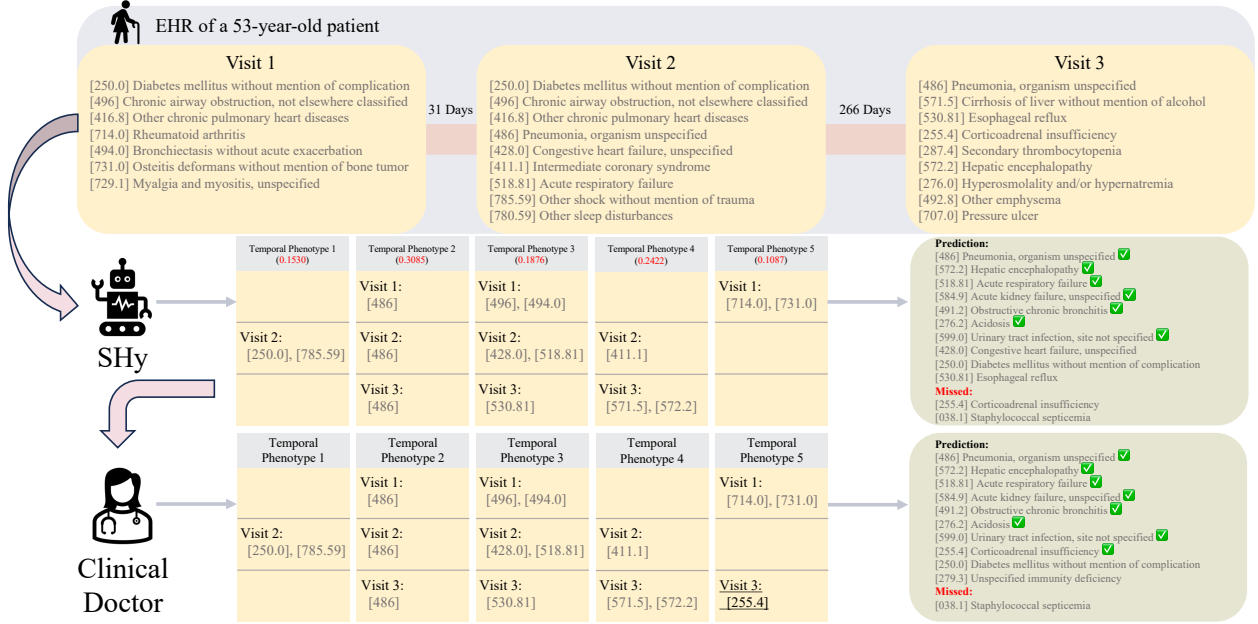
Figure 3: An illustration of how SHy extracts five temporal phenotypes from the EHR of a 53-year-old female patient and how a clinician can refine the prediction by directly adjusting the generated phenotypes. Underlined text highlights human modifications, while check marks indicate correct predictions. The red numbers indicate the importance weights.

type 5 implied that the abnormal immune response was detrimentally affecting the bones and joints. **Every phenotype contributed to the prediction result.** Pneumonia, chronic obstructive bronchitis and acute respiratory failure could be inferred from Phenotype 2; urinary tract infection, closely associated with diabetes and acute kidney failure, with hypoxia as the main driving force, could be anticipated based on Phenotype 1 and 3; hepatic encephalopathy could be predicted through Phenotype 4; Phenotype 5, reflecting the patient's waning immune system, played a suggestive role in predicting most lesions. **The weights assigned to phenotypes were justifiable.** Phenotype 2, which received the highest weight, was the most critical because the patient had diagnoses related to the respiratory system across all past and subsequent visits, and a consistent diagnosis of pneumonia suggested a weakening immune system and a critical health condition; Phenotype 5, assigned the lowest weight, was less relevant compared to the other phenotypes, as the patient did not receive diagnoses similar to those within this phenotype during

the last two visits. More analysis on the case study can be found in Appendix E.

5. Conclusions

In this study, we introduce SHy, a self-explaining model for diagnosis prediction. It represents each patient's longitudinal EHR as a hypergraph and employs message passing to derive personalized disease embeddings. With these embeddings, SHy offers tailored explanations by extracting temporal phenotypes upon which predictions are based. With a novel combination of objectives, we ensure that these temporal phenotypes are concise, faithful, distinct, and easy to be edited by domain experts. Quantitative evaluations highlight SHy's competitive predictive performance and superior explanatory capabilities, advocating its potential as an AI assistant. Future endeavors include employing timestamps in the prediction process and devising more robust metrics for evaluating the explanation quality of diagnosis prediction models.

References

David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018.

Devanshu Arya, Deepak K Gupta, Stevan Rudinac, and Marcel Worring. Hypersage: Generalizing inductive representation learning on hypergraphs, 2020.

Song Bai, Feihu Zhang, and Philip HS Torr. Hypergraph convolution and hypergraph attention. *Pattern Recognition*, 110:107637, 2021.

Tian Bai, Shanshan Zhang, Brian L Egleston, and Slobodan Vucetic. Interpretable representation learning for healthcare via capturing disease progression through time. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 43–51, 2018.

Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017.

Derun Cai, Moxian Song, Chenxi Sun, Baofeng Zhang, Shenda Hong, and Hongyan Li. Hypergraph structure learning for hypergraph neural networks. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1923–1929, 2022a.

Derun Cai, Chenxi Sun, Moxian Song, Baofeng Zhang, Shenda Hong, and Hongyan Li. Hypergraph contrastive learning for electronic health records. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 127–135. SIAM, 2022b.

Qingpeng Cai, Kaiping Zheng, Beng Chin Ooi, Wei Wang, and Chang Yao. Elda: Learning explicit dual-interactions for healthcare analytics. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 393–406. IEEE, 2022c.

Zhengping Che, David Kale, Wenzhe Li, Mohammad Taha Bahadori, and Yan Liu. Deep computational phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 507–516, 2015.

Eli Chien, Chao Pan, Jianhao Peng, and Olgica Milenkovic. You are allset: A multiset function framework for hypergraph neural networks. In *International Conference on Learning Representations*, 2022.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR, 2016a.

Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. Multi-layer representation learning for medical concepts. In *proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1495–1504, 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. *Advances in neural information processing systems*, 29, 2016c.

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

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

Suhan Cui, Junyu Luo, Muchao Ye, Jiaqi Wang, Ting Wang, and Fenglong Ma. Medskim: Denoised health risk prediction via skimming medical claims data. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 81–90. IEEE, 2022.

Yihe Dong, Will Sawin, and Yoshua Bengio. Hnhn: Hypergraph networks with hyperedge neurons, 2020.

Mengnan Du, Ninghao Liu, and Xia Hu. Techniques for interpretable machine learning. *Communications of the ACM*, 63(1):68–77, 2019.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3558–3565, 2019.

Junyi Gao, Cao Xiao, Yasha Wang, Wen Tang, Lucas M Glass, and Jimeng Sun. Stagenet: Stage-aware neural networks for health risk prediction. In *Proceedings of The Web Conference 2020*, pages 530–540, 2020.

Yue Gao, Yifan Feng, Shuyi Ji, and Rongrong Ji. Hgnn+: General hypergraph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3181–3199, 2022.

Srishti Gautam, Ahcene Boubekki, Stine Hansen, Suaiba Salahuddin, Robert Jenssen, Marina Höhne, and Michael Kampffmeyer. Protovae: A trustworthy self-explainable prototypical variational model. *Advances in Neural Information Processing Systems*, 35:17940–17952, 2022.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, pages e215–e220, 2000.

Jing Huang and Jie Yang. Unignn: a unified framework for graph and hypergraph neural networks. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-21*, 2021.

Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of NAACL-HLT*, 2019.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017.

Jaehyeong Jo, Jinheon Baek, Seul Lee, Dongki Kim, Minki Kang, and Sung Ju Hwang. Edge representation learning with hypergraphs. *Advances in Neural Information Processing Systems*, 34:7534–7546, 2021.

Alistair E. W. Johnson, Tom J. Pollard, and Roger G. Mark. MIMIC-III clinical database (version 1.4), 2016a.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(160035), 2016b. doi: https://doi.org/10.1038/sdata.2016.35.

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

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

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 Journal of Biomedical and Health Informatics*, November 2023. URL https://doi.org/10.1109/JBHI.2022.3224727.

Chuanren Liu, Fei Wang, Jianying Hu, and Hui Xiong. Temporal phenotyping from longitudinal electronic health records: A graph based framework. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 705–714, 2015.

Chang Lu, Chandan K Reddy, Prithwish Chakraborty, Samantha Kleinberg, and Yue Ning. Collaborative graph learning with auxiliary text for temporal event prediction in healthcare. In *Proceedings of the Thirty-First International Joint*

Conference on Artificial Intelligence, IJCAI-21, 2021a.

Chang Lu, Chandan K Reddy, and Yue Ning. Self-supervised graph learning with hyperbolic embedding for temporal health event prediction. *IEEE Transactions on Cybernetics*, 2021b.

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

Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 647–656, 2020.

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 *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1903–1911, 2017.

Fenglong Ma, Jing Gao, Qiuling Suo, Quanzeng You, Jing Zhou, and Aidong Zhang. Risk prediction on electronic health records with prior medical knowledge. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018a.

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

Liantao Ma, Junyi Gao, Yasha Wang, Chaohe Zhang, Jiangtao Wang, Wenjie Ruan, Wen Tang, Xin Gao, and Xinyu Ma. Adacare: Explainable clinical health status representation learning via scale-adaptive feature extraction and recalibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 825–832, 2020a.

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

Gao, and Junyi Gao. Concare: Personalized clinical feature embedding via capturing the healthcare context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020b.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.

Chuizheng Meng, Loc Trinh, Nan Xu, James Enouen, and Yan Liu. Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. *Scientific Reports*, 2022.

Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan, Mitesh M Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. Towards transparent and explainable attention models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4206–4216, 2020.

Seyedeh Neelufar Payrovnaziri, Zhaoyi Chen, Pablo Rengifo-Moreno, Tim Miller, Jiang Bian, Jonathan H Chen, Xiuwen Liu, and Zhe He. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *Journal of the American Medical Informatics Association*, 27(7):1173–1185, 2020.

Xueping Peng, Guodong Long, Tao Shen, Sen Wang, and Jing Jiang. Sequential diagnosis prediction with transformer and ontological representation. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 489–498. IEEE, 2021a.

Xueping Peng, Guodong Long, Sen Wang, Jing Jiang, Allison Clarke, Clement Schlegel, and Chengqi Zhang. Mipo: Mutual integration of patient journey and medical ontology for healthcare representation learning, 2021b.

Zhi Qiao, Shiwan Zhao, Cao Xiao, Xiang Li, Yong Qin, and Fei Wang. Pairwise-ranking based collaborative recurrent neural networks for clinical event prediction. In *Proceedings of the twenty-seventh international joint conference on artificial intelligence*, 2018.

Zhi Qiao, Zhen Zhang, Xian Wu, Shen Ge, and Wei Fan. Mhm: Multi-modal clinical data based hierarchical multi-label diagnosis prediction. In *Proceed-*

ings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1841–1844, 2020.

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.

Houxing Ren, Jingyuan Wang, Wayne Xin Zhao, and Ning Wu. Rapt: Pre-training of time-aware transformer for learning robust healthcare representation. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2021.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Sofia Serrano and Noah A Smith. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, 2019.

Yuqi Si, Jingcheng Du, Zhao Li, Xiaoqian Jiang, Timothy Miller, Fei Wang, W Jim Zheng, and Kirk Roberts. Deep representation learning of patient data from electronic health records (ehr): A systematic review. *Journal of biomedical informatics*, 115:103671, 2021.

Huan Song, Deepta Rajan, Jayaraman Thiagarajan, and Andreas Spanias. Attend and diagnose: Clinical time series analysis using attention models. In *Proceedings of the AAAI conference on artificial intelligence*, 2018.

Lihong Song, Chin Wang Cheong, Kejing Yin, William K Cheung, Benjamin CM Fung, and Jonathan Poon. Medical concept embedding with multiple ontological representations. In *IJCAI*, volume 19, pages 4613–4619, 2019.

Yanchao Tan, Chengjun Kong, Leisheng Yu, Pan Li, Chaochao Chen, Xiaolin Zheng, Vicki S Hertzberg, and Carl Yang. 4sdrug: Symptom-based set-to-set small and safe drug recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.

Yanchao Tan, Zihao Zhou, Leisheng Yu, Weiming Liu, Chaochao Chen, Guofang Ma, Xiao Hu, Vicki S Hertzberg, and Carl Yang. Enhancing personalized healthcare via capturing disease severity, interaction, and progression. In *2023 IEEE International Conference on Data Mining (ICDM)*, 2023.

Yanchao Tan, Hengyu Zhang, Zihao Zhou, Guofang Ma, Fan Wang, Weiming Liu, Xinting Liao, Vicki S Hertzberg, and Carl Yang. Enhancing progressive diagnosis prediction in healthcare with continuous normalizing flows. In *Companion Proceedings of the ACM on Web Conference 2024*, 2024.

Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*. PMLR, 2019.

Maarten Van Smeden, Johannes B Reitsma, Richard D Riley, Gary S Collins, and Karel GM Moons. Clinical prediction models: diagnosis versus prognosis. *Journal of clinical epidemiology*, 132:142–145, 2021.

Yipei Wang and Xiaoqian Wang. Self-interpretable model with transformation equivariant interpretation. *Advances in Neural Information Processing Systems*, 34:2359–2372, 2021.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

Hanrui Wu and Michael K Ng. Hypergraph convolution on nodes-hyperedges network for semi-supervised node classification. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(4):1–19, 2022.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2018.

Ran Xu, Yue Yu, Chao Zhang, Mohammed K Ali, Joyce C Ho, and Carl Yang. Counterfactual and factual reasoning over hypergraphs for interpretable clinical predictions on ehr. In *Machine Learning for Health*, pages 259–278. PMLR, 2022.

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

Naganand Yadati, Madhav Nimishakavi, Prateek Yadav, Vikram Nitin, Anand Louis, and Partha Talukdar. Hypergcn: A new method for training graph convolutional networks on hypergraphs. *Advances in neural information processing systems*, 32, 2019.

Borui Zhang, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Bort: Towards explainable neural networks with bounded orthogonal constraint. In *The Eleventh International Conference on Learning Representations*, 2023.

Chaohe Zhang, Xin Gao, Liantao Ma, Yasha Wang, Jiangtao Wang, and Wen Tang. Grasp: generic framework for health status representation learning based on incorporating knowledge from similar patients. In *Proceedings of the AAAI conference on artificial intelligence*, 2021.

Jinghe Zhang, Kamran Kowsari, James H Harrison, Jennifer M Lobo, and Laura E Barnes. Patient2vec: A personalized interpretable deep representation of the longitudinal electronic health record. *IEEE Access*, 6:65333–65346, 2018.

Tianran Zhang, Muhao Chen, and Alex AT Bui. Diagnostic prediction with sequence-of-sets representation learning for clinical events. In *Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings 18*, pages 348–358. Springer, 2020a.

Xi Sheryl Zhang, Fengyi Tang, Hiroko H Dodge, Jiayu Zhou, and Fei Wang. Metapred: Meta-learning for clinical risk prediction with limited patient electronic health records. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019.

Xianli Zhang, Buyue Qian, Shilei Cao, Yang Li, Hang Chen, Yefeng Zheng, and Ian Davidson. Inprem: An interpretable and trustworthy predictive model for healthcare. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 450–460, 2020b.

Yue Zhao, Zhi Qiao, Cao Xiao, Lucas Glass, and Jimeng Sun. Pyhealth: A python library for health predictive models, 2021.

Appendix A. Deep Learning on Hypergraphs

The success of graph neural networks (Kipf and Welling, 2017) has spurred hypergraph-based deep learning methods. Models updating node embeddings of hypergraphs through clique-expansion or its variants were proposed (Feng et al., 2019; Gao et al., 2022; Bai et al., 2021; Yadati et al., 2019). HyperSAGE (Arya et al., 2020) was developed for inductive hypergraph learning, capturing relations within and across hyperedges. HNHN (Dong et al., 2020) employed a normalization strategy that could be adjusted according to datasets. Jo et al. (2021) innovated edge representation learning on graphs by introducing dual hypergraphs. HCNH (Wu and Ng, 2022) utilized the hypergraph reconstruction loss for semi-supervised node classification. Chien et al. (2022) introduced AllSet, a generalized framework encapsulating most existing propagation rules on hypergraphs. For EHR-based tasks, HCL, ProCare, and CACHE emerged as three hypergraph neural network models (Xu et al., 2022; Tan et al., 2023; Cai et al., 2022b), with the former two lacking interpretability and the last one prone to false negative diagnoses.

Appendix B. Comparing Hypergraph Neural Networks

To optimize the performance of SHy in diagnosis prediction, we experimented with 8 state-of-the-art hypergraph neural network architectures. As SHy represents each patient through a distinct hypergraph and all patient hypergraphs utilize a shared message-passing mechanism, only inductive hypergraph neural network models can be considered. The results in Table 4 indicate that UniGIN (Huang and Yang, 2021), AllSetTransformer (Chien et al., 2022), and UniGAT outperform others on the MIMIC-III dataset, whereas UniGIN, UniGCN, and HyperGCN (Yadati et al., 2019) are the top performers on MIMIC-IV. Given UniGIN's consistently high performance across both datasets and its fewer parameters relative to

UniGAT and AllSetTransformer, we adopted UniGIN as the message-passing mechanism for SHy.

Appendix C. The GRU Decoder

SHy utilizes a GRU decoder to reconstruct patient hypergraphs from the embeddings of the temporal phenotypes, denoted as $[\mathbf{U}_1 \,\|\, \mathbf{U}_2 \,\|\, ... \,\|\, \mathbf{U}_K] \in \mathbb{R}^{K d_{hid}}$. Specifically, the reconstructed patient hypergraph, $\hat{\mathbf{P}}$, is derived as follows:

$$\mathbf{H}_1, ..., \mathbf{H}_T = \mathrm{GRU}([\mathbf{U}_1 \,\|\, ... \,\|\, \mathbf{U}_K], ..., [\mathbf{U}_1 \,\|\, ... \,\|\, \mathbf{U}_K])$$
$$\hat{\mathbf{P}} = \sigma(\mathbf{H}\mathbf{W}_{\mathrm{recon}} + \boldsymbol{b}_{\mathrm{recon}})^\top,$$

where σ denotes the sigmoid function, $\mathbf{H}_t$ is the hidden state corresponding to the t-th visit, and $\mathbf{W}_{\mathrm{recon}}$ and $\boldsymbol{b}_{\mathrm{recon}}$ are learnable parameters. In essence, SHy duplicates the concatenated phenotype embedding T times, providing the same input to the decoder GRU at each time step and reconstructing the columns of $\hat{\mathbf{P}}$ using the derived hidden states.

Appendix D. More on Experimental Setup

D.1. Dataset Statistics

Details on the employed datasets are in Table 5.

D.2. Baseline Descriptions

We compare SHy with 13 representative state-of-the-art models on diagnosis prediction:

- Doctor AI (Choi et al., 2016a) utilizes a GRU for learning patient representation.
- RETAIN (Choi et al., 2016c) employs RNNs and an attention mechanism for interpretable predictions.
- Dipole (Ma et al., 2017) leverages bidirectional RNNs with an attention mechanism for interpretability.
- T-LSTM (Baytas et al., 2017) uses a time-aware long short-term memory to learn patient representation.
- GRAM (Choi et al., 2017) infuses knowledge from medical ontologies into disease embeddings via attention.
- CGL (Lu et al., 2021a) achieves accurate diagnosis predictions via collaborative graph learning.

- Timeline (Bai et al., 2018) ensures interpretability and prediction accuracy through learning time decay factors.
- AdaCare (Ma et al., 2020a) models correlations in diagnoses to retain interpretability and prediction accuracy.
- StageNet (Gao et al., 2020) extracts disease stage information for accurate diagnosis prediction.
- ConCare (Ma et al., 2020b) models feature interactions with attention for personalized diagnosis prediction.
- Chet (Lu et al., 2022) exploits transition functions on dynamic graphs to model disease progressions.
- SETOR (Peng et al., 2021a) employs neural ordinary differential equations for learning patient representation.
- MIPO (Peng et al., 2021b) leverages Transformer and medical ontologies for accurate diagnosis prediction.

For a fair comparison, while CGL utilizes unstructured texts in the EHR dataset to enhance prediction, we have opted to use it without its textual feature modeling module. Baseline models are implemented using the original authors' code or through PyHealth (Zhao et al., 2021).

D.3. Hyperparameters

The datasets are split into training, validation, and testing sets using an 0.8:0.1:0.1 ratio. We employ the Adam optimizer and standardize the batch size to 128 across all models. We carefully tune the learning rate and specific hyperparameters of the baseline models to optimize their performance. Through grid search, hyperparameters for SHy are finalized: $K = 5$ and $Z = 2$ for both MIMIC-III and MIMIC-IV. Details for other hyperparameters are available in our repository. The models are trained on a server with NVIDIA A40 GPUs. For all experiments, we present the average results from 5 runs with random model initializations.

Appendix E. More on Case Study

SHy enhanced its explanatory capability by including false negatives. For Phenotype 1, a diagnosis of pneumonia was retrospectively added to the initial visit, reflecting the patient's respiratory

Table 4: The Comparison of Different Hypergraph Neural Network Architectures

Model	Recall@10	Recall@20	nDCG@10	nDCG@20	# Params	Recall@10	Recall@20	nDCG@10	nDCG@20	# Params
	MIMIC-III					MIMIC-IV				
SHय (UniGIN)	0.2775	0.3831	0.4135	**0.4088**	2.69M	**0.3344**	**0.4270**	**0.4236**	**0.4239**	3.59M
SHय (UniSAGE)	0.2744	0.3799	0.4112	0.4050	2.69M	0.3311	0.4231	0.4205	0.4199	3.59M
SHय (UniGAT)	**0.2780**	0.3828	**0.4141**	0.4080	2.70M	0.3321	0.4240	0.4212	0.4230	3.59M
SHय (UniGCN)	0.2758	0.3735	0.4132	0.4055	2.69M	0.3338	0.4268	0.4208	0.4232	3.59M
SHय (UniGCNII)	0.2659	0.3722	0.4098	0.4009	2.63M	0.3328	0.4251	0.4206	0.4228	3.53M
SHय (HyperGCN)	0.2648	0.3671	0.4024	0.3965	2.50M	0.3318	0.4234	0.4216	0.4229	3.49M
SHय (AllDeepSets)	0.2693	0.3812	0.4051	0.4036	3.08M	0.3305	0.4216	0.4200	0.4215	3.73M
SHय (AllSetTransformer)	0.2762	**0.3835**	0.4123	0.4080	3.06M	0.3303	0.4224	0.4205	0.4226	3.73M

Table 5: Dataset Statistics

Dataset	MIMIC-III	MIMIC-IV
# of patients	7493	45250
# of visits	19894	162372
Avg. # of visits per patient	2.66	3.59
Max # of visits per patient	42	95
Avg. # of codes per visit	13.06	10.27
Max # of codes per visit	39	39
# of unique ICD-9 codes	4880	8402

problems, such as bronchiectasis, in the same visit, and subsequent pneumonia diagnoses; also, this adjustment further emphasized the patient's enduring respiratory and immune system problems.

A clinician observed that SHय initially overlooked corticoadrenal insufficiency, and none of the phenotypes included this diagnosis in the third visit. Since rheumatoid arthritis is an autoimmune disease, which could be a major cause of corticoadrenal insufficiency, this clinician decided to intervene by incorporating corticoadrenal insufficiency in the third visit into Phenotype 5 and proceed with the predictive process using the adjusted phenotypes. Interestingly, this intervention allowed SHय to accurately predict corticoadrenal insufficiency in the next visit, demonstrating its superior flexibility by allowing explicit interventions from domain experts in the prediction process.