PATNet: Propensity-Adjusted Temporal Network for Joint Imputation and Prediction using Binary EHRs with Observation Bias

Kejing Yin, Dong Qian, William K. Cheung *Member, IEEE*

Abstract—Predictive analysis of electronic health records (EHR) is a fundamental task that could provide actionable insights to help clinicians improve the efficiency and quality of care. EHR are commonly recorded in binary format and contain inevitable missing data. The nature of missingness may vary by patients, clinical features, and time, which incurs observation bias. It is essential to account for the binary missingness and observation bias or the predictive performance could be substantially compromised. In this paper, we develop a propensity-adjusted temporal network (PATNet) to conduct data imputation and predictive analysis simultaneously. PATNet contains three subnetworks: 1) an imputation subnetwork that generates the initial imputation based on historical observations, 2) a propensity subnetwork that infers the patient-, feature-, and time-dependent propensity scores, and 3) a prediction subnetwork that produces the missing-informative prediction using the propensity-adjusted imputations and the missing probabilities. To allow the propensity scores to be inferred from data, we use the expectation-maximization (EM) algorithm to learn the imputation and propensity subnetworks and incorporate a low-rank constraint via PARAFAC2 approximation. Extensive evaluation using the MIMIC-III and eICU datasets demonstrates that PATNet outperforms the state-of-the-art methods in terms of binary data imputation, disease progression modeling, and mortality prediction tasks.

Index Terms—Electronic health records, clinical risk prediction, disease progression modeling, missing data, binary data imputation, propensity score, missing at random

1 Introduction

Electronic Health Records (EHR) have been increasingly available during the past decade. Consisting of longitudinal health data about patient, including demographics, diagnoses, laboratory tests and medication prescriptions, EHR data has triggered countless data-driven researches on secondary use for predictive analysis [1], [2]. Disease progression modeling tries to capture the disease developing process from historical records to predict future risk [3], [4], [5]. Disease onsite risk prediction [6], [7] and in-hospital mortality prediction [8], [9], [10] provide insights for patient risk stratification and priority setting in the allocation of limited medical resources.

In such predictive analysis tasks, it is common that features like presence of diagnosis codes or prescriptions are represented in *binary format* [7], [8]. However, the observed records inevitably contain significant missingness due to reasons like difficulties in performing diagnosis and operational causes. What makes it even worse is that the missingness is typically not completely at random in real practice, which introduces ***observation bias***. In other words, the probability of observing a particular feature (ground truth) varies among different features, patients, and time. For example, patients with a fever generally have higher chances to have the laboratory test of blood count performed and recorded, but those without a fever are more likely to have this laboratory test missing in the records [11]. Besides, diagnosing some diseases may require laboratory tests which are expensive or time-consuming. As a result, the diagnosis records of those diseases may have higher chances to be missing at the early stages.

Existing methods on predictive analysis using EHR typically ignore the missingness and observation bias by directly conditioning their model predictions on the partial observations [7], [8], [12]. In this paper, we conjecture that *ignoring the observation bias would result in substantially degraded performance*, and we propose to estimate the complete data and simultaneously train the predictive model using the estimated complete data. Our conjecture is in line with recent studies, where missing data could decrease the accuracy and the reliability of clinical predictive analysis, and data imputation is a promising way to alleviate this issue [11], [13], [14], [15], [16]. However, despite the great effort and encouraging results reported in the literature, it remains an open challenge to impute binary EHR with observation bias for predictive analysis due to several issues.

First, the majority of the existing data imputation methods are developed for continuous-valued data rather than binary data [17], [18], [19], [20], [21]. They require a *masking matrix* to indicate which features are missing for each subject, and mask out the missing values during learning. Unfortunately, this is not possible for partially observed binary data due to the lack of explicit negative records [22]. In other words, such binary data are mixtures of positive and unobserved features, and it is unknown what unobserved features in the data are the missing ones. For example, in EHR, a diagnosis record with a value of one recorded indicates that the diagnosis is confirmed by the clinician.

• *Kejing Yin, Dong Qian and William K. Cheung are with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong.*
E-mail: {cskjyin, dongqian, william}@comp.hkbu.edu.hk

But a value of zero only indicates that the diagnosis is not recorded, and the disease diagnosis is in fact unknown as it can be missing. So, the goal of binary data imputation is to distinguish the missing positive features from the true negative features.

Second, although a few binary data imputation techniques have recently been proposed, they cannot well handle the observation bias incurred by the not-completely-at-random missingness. To alleviate such bias, one approach is to estimate the probabilities of observing positive features in the data, which are defined as the *propensity scores*. Existing methods either assume that the missing is completely at random with a constant missing probability [22], [23], or rely on simple heuristics to determine the propensity scores based on some basic statistical quantities of the data [24]. In practice, however, the ground-truth propensity scores could exhibit much more variations and such naive methods could lead to suboptimal performance.

To summarize, the observation bias arising from varying probabilities in observing different clinical features across time and subjects should be considered in both data imputation and predictive analysis when applied to binary EHR, which has not been fully addressed by existing research.

To tackle this issue, we develop the Propensity-Adjusted Temporal Network (PATNet), which is composed of three subnetworks. Specifically, we use a Transformer-based *imputation subnetwork* to generate an initial imputation based on historically observed features. To alleviate the observation bias, we utilize an LSTM-based *propensity subnetwork* coupled with a low-rank approximation method called PARAFAC2 to infer the propensity scores from historical observations. Then, we compute the final imputation by deriving the posterior distribution of the complete data given the partial observations and the two subnetworks. To conduct missing-informative prediction, we use the third *prediction subnetwork* which takes the derived posterior estimation as well as the propensity scores as input and generates the final prediction. Due to the lack of ground truth, we cannot directly learn the propensity scores in a supervised manner. We make use of the expectation-maximization (EM) framework to learn the imputation and propensity subnetworks. We evaluate the proposed model using two publicly available EHR datasets. Extensive experiments show that PATNet yields significantly more accurate imputations for binary missing data for all levels of missing ratios and observation bias. PATNet is also more robust against higher missing ratios due to explicitly estimating the propensity scores simultaneously. As a result of more accurate imputation, PATNet obtains up to 23% relative improvement for the disease progression modeling task and 12% for the mortality prediction task.

2 Background and Related Work

2.1 Missing Data in Electronic Health Records

It has been widely recognized that EHR data contain substantial missing information, which could potentially undermine the validity of conclusions drawn from EHR if such missingness is not carefully addressed [25]. Depending on their generation mechanisms, missing data can be divided into three categories: missing-completely-at-random (MCAR), missing-at-random (MAR) and missing-not-at-random (MNAR) [26]. MCAR indicates that the probability of a variable being missing is independent of itself and all other covariates, *i.e.*, being a constant across subjects, times and features. MAR means that the probability of having a missing variable is independent of the value of itself, but could depend on other covariates. MNAR means that the probability of missing data depends only on the values of the variables subject to missing.

It is generally believed that missing data could have negative impact on predictive analysis using EHR data and data imputation is a feasible solution to tackle this problem. For example, it is shown in a simulation study that the point estimates of odds ratio (OR) for predictors in a logistic regression are significantly over- or under-estimated in the case of MAR. Even in MCAR, their confidence interval becomes noticeably wider [25]. In several clinical predictive studies including vesicoureteral reflux and recurrent urinary tract infection prediction [14], septic shock prediction [13], and heart failure prediction [15], it is observed that missing data imputation in conjunction with predictive modeling leads to improved prediction accuracy.

Irregular time series modeling has been gaining increasing popularity and shown promising in tackling missing data. Some representative models include T-LSTM [27], HiTANet [12], mTAN [28], RAINDROP [29], Neural ODEs [30], [31], and CATNet [32]. Such methods treat the observed EHR data as data points irregularly sampled from an underlying continuous time series. By directly capturing the dynamics of the underlying time series over irregular time intervals, the issue of missing values can be partially alleviated.

2.2 Binary Data Imputation

A large portion of the EHR data are represented in binary form, *e.g.*, presence or absence of diagnosis codes, abnormal laboratory tests and medication prescriptions. During the process of collecting and recording the binary clinical data, it is quite common that when a patient does not have certain features (*e.g.*, symptom/comorbidity) or the patient is not asked about the features, the data fields are left blank instead of labeled as negative [25]. Despite its ubiquitousness, the issue of missingness in such binary data is relatively under-researched, compared with the continuous missing data imputation problem. Recently, several low-rank models were developed to tackle the binary data imputation issue. [23] applies the positive-unlabeled (PU) classification method to complete binary data organized in matrices based on low-rank constraints and [22] extends it to handle binary temporal data with missing values. Both assume that the missing is completely at random, and could fail when applied to real-world data where observation bias often exists (*i.e.*, missing-not-completely-at-random). Also, [22] is limited in modeling the underlying nonlinear and complex temporal dependency of the temporal EHR data, which could be crucial for accurate clinical data imputation. [24] aims to alleviate the bias arising from missing-not-completely-at-random in the application of recommender systems. It utilizes the propensity scores to represent the probability of observing present features and adjusts for

them during training. However, it assumes that 1) the propensity is proportional to frequency of features, and 2) the propensity of a feature is the same for all subjects. Both assumptions may not hold for applications other than recommender systems. Besides, it only handles static matrix data but no temporal information can be utilized.

To summarize, developing data imputation methods that accounts for the observation bias and underlying temporal dependency remains an important and open challenge.

3 Data Observation Model

Our binary EHR data contains medical records of N individual patients. The partially observed records of the n-th patient is represented as a sequence of multi-hot vectors $\{\mathbf{x}_1^{(n)}, \mathbf{x}_2^{(n)} \ldots, \mathbf{x}_{T_n}^{(n)}\}$, where $\mathbf{x}_t^{(n)} \in \{0,1\}^J$ denotes the observed binary clinical features of this patient at the t-th time step (*e.g.*, a clinical visit or one day in ICU), T_n is the number of time steps of the n-th patient, and J is the number of clinical features. At the t-th time step, we denote the historical observation of the n-th patient by $\mathcal{H}_t^{(n)} = \left\{\mathbf{x}_1^{(n)}, \ldots, \mathbf{x}_{t-1}^{(n)}\right\}$, a sequence consisting of the past observed clinical features. We denote the underlying complete data by $\mathbf{z}_t^{(n)} \in \{0,1\}^J$ for the n-th patient at the t-th time step. To characterize the observation bias, we borrow the concept of *propensity score* from causal inference literature [33], which is formally defined as follows.

Definition 1 (Propensity Score)**.** For the n-th patient at the t-th time step, the propensity score $\rho_{tj}^{(n)} \in (0,1)$ is defined as the probability of observing the j-th feature if its ground truth is positive, *i.e.*,

$$\rho_{tj}^{(n)} = p(x_{tj}^{(n)} = 1 | z_{tj}^{(n)} = 1).$$

Depending on the prediction task, a target label $y^{(n)}$ may be associated with the n-th patient. For example, in mortality prediction, a binary in-hospital decease outcome will be recorded as the target label; in disease progression modeling, it is common to train the predictive model in an autoregressive way, *i.e.*, using the clinical features of the next time step as labels for the current time step. Fig. 1 shows a graphical comparison between the observation model of the existing methods and that of our proposed PATNet. Existing predictive models (*left*) for EHR ignore the observation bias by treating the observed data as complete and assume that the label is generated from the observations; therefore, they directly condition their model prediction on the observations. Our model (*right)*, in contrast, assumes that the observations are only a part of the latent complete data ($\mathbf{z}$) that are recorded according to their corresponding propensity scores ($\boldsymbol{\rho}$). The labels are generated from the complete data instead of the partial observations. This observation model motivates us to learn the latent complete data and the propensity scores to make more informed predictions that are conditioned on the complete data.

To ease the notations, we omit the patient index "(n)" in the superscripts whenever it does not cause confusion.

4 Propensity-Adjusted Temporal Network

Given the partial observation of the binary EHR data $\mathbf{X}$, our goal is to (1) estimate the latent complete data $\mathbf{Z}$, and (2) simultaneously predict their corresponding target labels y using the estimated $\mathbf{Z}$. In this section, we describe the architecture of PATNet that is designed to achieve this goal. Fig. 2 shows an overview of our proposed PATNet method. It comprises three subnetworks. The **imputation subnetwork** f_{imp} takes the historical observation as input and generates an initial imputation. The **propensity subnetwork** f_{prop} estimates the propensity scores subject to a low-rank approximation based on the historical observations. Then, the posterior probability of the ground truth is computed by adjusting the initial imputation by the estimated propensity scores. The **prediction subnetwork** then takes the propensity-adjusted imputation and the propensity scores as input to perform missing-informative predictions. We introduce each component in detail below.

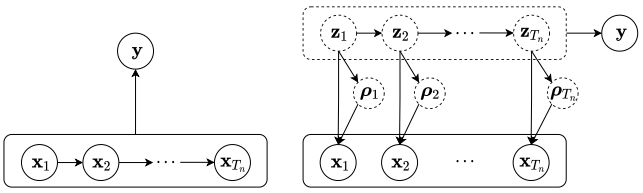


Fig. 1. A graphical comparison between the observation model of the existing methods (*left*) and that of ours (*right*) for EHR predictive analysis. Existing EHR predictive models do not consider the observation bias and assume that the temporal features are generated in an autoregressive way and labels are generated from the observed features. On the contrary, our model considers the observation bias by assuming that the records only reflect a part of the complete data according to their corresponding propensity scores, and the labels are generated from the latent ground-truth features, instead of the partial observations.

4.1 Imputation Subnetwork

The goal of the imputation subnetwork is to produce an initial imputation based on the historical observations, *i.e.*,

$$\widetilde{\mathbf{x}}_t := p(\mathbf{z}_t = 1 | \mathcal{H}_t) = f_{\text{imp}}\left(\mathcal{H}_t; \Theta_{\text{imp}}\right) \quad \forall t, \tag{1}$$

where Θ_{imp} denotes its parameters to be learned. In principle, any sequential modeling method can be used to parameterize f_{imp}. In PATNet, we use a masked Transformer [34] for the imputation subnetwork due to its advantages in modeling long-term dependencies, which is particularly important in modeling clinical data.

The binary observations before the last time step are first encoded into an m-dimensional continuous space by applying a linear transformation and adding the positional encoding as follows:

$$\mathbf{e}_t = \frac{1}{\|\mathbf{x}_t\|_1} \mathbf{W}_{\text{emb}} \mathbf{x}_t + PE_t \quad t = 1, \ldots, T-1, \tag{2}$$

where $\mathbf{W}_{\text{emb}} \in \mathbb{R}^{m \times J}$ is the embedding matrix and m is the dimension of the embeddings. The positional encoding $PE_t \in \mathbb{R}^m$ is given by:

$$\begin{aligned} PE_{(t,2d)} &= \sin(t/10000^{2t/m}), \\ PE_{(t,2d+1)} &= \cos(t/10000^{2t/m}), \end{aligned} \tag{3}$$

where d is the index of the embedding dimensions. The embedded inputs are concatenated as the input to the first Transformer layer, where a zero vector is inserted before

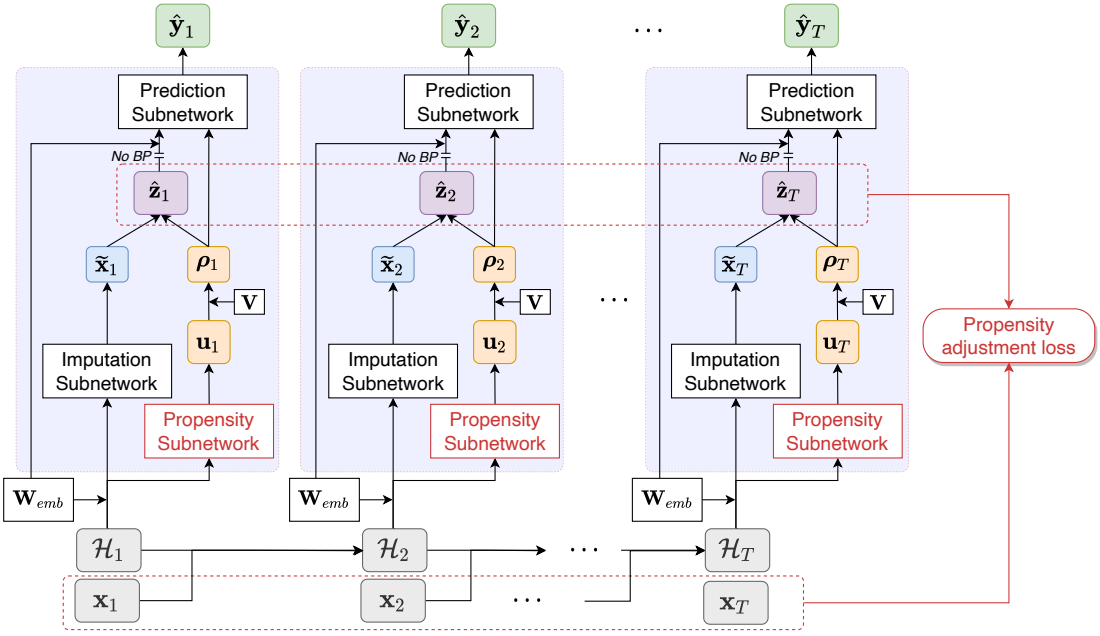


Fig. 2. The framework overview of PATNet. It consists of three subnetworks: the imputation subnetwork, the propensity subnetwork, and the prediction subnetwork. At each time step, the imputation subnetwork and the propensity subnetwork take the historical data as input and generate as output the initial imputation and the low-rank propensity factors, respectively. A global factor matrix is multiplied to the propensity factors to recover the propensity score estimations. The posteriors then are computed as the final estimation of the complete features at the corresponding time step. The embedded posteriors and the propensity scores are concatenated as the input to carry out the missingness-informed prediction using the prediction subnetwork. The propensity adjustment loss is minimized together with the prediction loss to drive the learning of propensity scores.

the first time step to indicate that there is no historical observation available for the first time step:

$$\mathbf{H}_0 = [\mathbf{0}, \mathbf{e}_1, \ldots, \mathbf{e}_{T-1}] \in \mathbb{R}^{T\times m}, \tag{4}$$

where $[\cdot]$ denotes concatenation. The imputation subnetwork contains L identical layers. Each layer first transforms the output from the previous layer ($\mathbf{H}_0$ for the first layer) into query, key, and value matrices by linear transformations as follows.

$$\begin{aligned}
\mathbf{Q}_{(l,i)} &= \mathbf{H}_{l-1}\mathbf{W}_i^Q \in \mathbb{R}^{T\times s} \quad l = 1, \ldots, L,\\
\mathbf{K}_{(l,i)} &= \mathbf{H}_{l-1}\mathbf{W}_i^K \in \mathbb{R}^{T\times s} \quad l = 1, \ldots, L,\\
\mathbf{V}_{(l,i)} &= \mathbf{H}_{l-1}\mathbf{W}_i^V \in \mathbb{R}^{T\times s} \quad l = 1, \ldots, L,
\end{aligned} \tag{5}$$

where $\mathbf{W}_i^Q \in \mathbb{R}^{m\times s}$, $\mathbf{W}_i^K \in \mathbb{R}^{m\times s}$, and $\mathbf{W}_i^V \in \mathbb{R}^{m\times s}$ are the learnable projection matrices, i is the index of the attention heads, l is the index of the layers, and s is the dimension of each head. The embedding dimensions are evenly split into I attention heads, *i.e.*, $s \times I = m$. Each attention head computes its own output $\widetilde{\mathbf{H}}_{(l,i)} \in \mathbb{R}^{T\times s}$ by:

$$\widetilde{\mathbf{H}}_{(l,i)} = \text{softmax}\left(\frac{\mathbf{Q}_{(l,i)}\mathbf{K}_{(l,i)}^\top}{\sqrt{m}} + \mathbf{M}\right)\mathbf{V}_{(l,i)} \quad i = 1, \ldots, I, \tag{6}$$

where I is the number of attention heads and $\mathbf{M} \in \mathbb{R}^{T\times T}$ is the patient-specific masking matrix added to ensure that no future information is used when making imputations at each time step, with its entries given by:

$$M_{ij} = \begin{cases} 0 & \text{if } i \geq j \\ -\infty & \text{otherwise} \end{cases}. \tag{7}$$

The final output of the l-th layer is then computed by concatenating the outputs of all attention heads and applying two fully connected layers with ReLU activation function. Residual connections and layer norms are used in between.

$$\begin{aligned}
\widetilde{\mathbf{H}}_l &= \text{LN}\left(\mathbf{H}_{l-1} + \left[\widetilde{\mathbf{H}}_{(l,1)}, \ldots, \widetilde{\mathbf{H}}_{(l,I)}\right]\right),\\
\mathbf{H}_l &= \text{LN}\left(\widetilde{\mathbf{H}}_l + \text{ReLU}\left(\widetilde{\mathbf{H}}_l\mathbf{W}_1 + \mathbf{1}_T\mathbf{b}_1^\top\right)\mathbf{W}_2 + \mathbf{1}_T\mathbf{b}_2^\top\right),
\end{aligned} \tag{8}$$

where LN denotes the operation of layer norm and $\mathbf{1}_T$ denotes a T-dimensional vector of all ones.

Finally, we use the output of the last Transformer layer to compute the initial imputation using a multilayer perceptron (MLP) followed by a sigmoid transformation:

$$[\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_T] = \sigma\big(\text{MLP}(\mathbf{H}_l)\big). \tag{9}$$

4.2 Propensity Subnetwork

Due to the presence of observation bias as introduced in Section 3 in real-world data, we need to also estimate the propensity scores based on historical observations:

$$\widehat{\rho}_{tj} := p\big(\mathbf{x}_{tj} = 1 | z_{tj} = 1, \mathcal{H}_t\big) = f_{\text{prop}}\left(\mathcal{H}_t; \Theta_{\text{prop}}\right) \quad \forall t. \tag{10}$$

However, this is not possible without further assumption: if the propensity score can take any arbitrary value, then we cannot distinguish whether a feature with zero value in the observation ($x_{tj} = 0$) is because of low probability of being observed ($\rho_{tj} \to 0$) or it is because of negative ground truth ($z_{tj} = 0$). Similar issue exists in the application of positive-unlabeled classification where the positive samples are selected according to some propensity scores. [35] proposes to learn the propensity scores using only a part of the features. However, EHR data are usually of high dimensions, and it is equally difficult to discover the feature set that determines the propensity scores. As such, PATNet will make use of

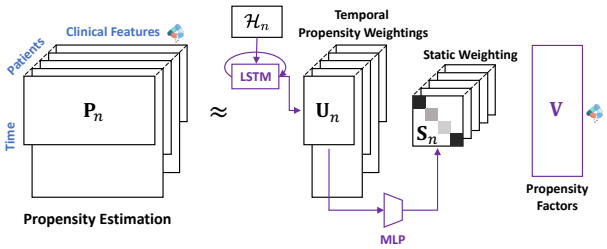


Fig. 3. The illustration of PARAFAC2 approximation of the propensity scores. The sequence of the propensity scores of all patients in the training set are naturally represented by an irregular tensor $\mathcal{P}$ where its n-th slice represents the propensity scores for the n-th patient. PARAFAC2 is a low-rank approximation of the irregular tensor, where $\mathrm{P}^{(n)}$ is approximated by the patient-dependent temporal propensity weighting $\mathbf{U}^{(n)}$, patient-dependent static weighting $\mathbf{S}^{(n)}$, and the patient-independent propensity factors $\mathbf{V}$.

all features available to infer the propensity scores, but we impose a low-rank constraint to the estimation of propensity scores. It has been shown that any sufficiently large and even full-rank matrix generated from a bounded latent variable model can be approximated by a low rank matrix up to a small element-wise error [36]. Moreover, matrix and tensor approximations have been shown particularly promising in modeling EHR data in practice [10], [22], [37]. In PATNet, we estimate low-rank propensity scores using a PARAFAC2 approximation combined with an LSTM model, which are detailed below in turn.

4.2.1 PARAFAC2 Approximation of Propensity Scores

While low-rank constraints can be easily imposed by minimizing regularization terms like nuclear norm [38], [39], they cannot be applied in our scenario due to the temporal irregularity of clinical data: patients have different numbers of visits or lengths of stays, making it impossible to construct a regular tensor and regularize its nuclear norm. Therefore, we use an irregular tensor to represent the propensity scores and the PARAFAC2 as its low-rank approximation. We use $\mathcal{P} = \left\{\mathrm{P}^{(n)} \in \mathbb{R}^{T_n \times J}\right\}_{n=1}^{N}$ to denote the irregular tensor for the propensity scores, which comprises a set of N matrices. $\mathrm{P}^{(n)}$ is the n-th slice of it, representing the temporal propensity scores of the n-th patient, and $\rho_{tj}^{(n)}$ is its entry indexed by t and j.

PARAFAC2 is a variant of tensor CP factorization [40] that can be applied to irregular tensors by allowing temporal weightings to vary between different patients while the low-rank components of the input irregular tensor are shared for all patients. Fig. 3 illustrates the logistic PARAFAC2 factorization [22] as an approximation of the propensity scores that take values between zero and one. Formally, the propensity scores of the n-th patient is approximated by:

$$\mathrm{P}^{(n)} \approx \widehat{\mathrm{P}}^{(n)} = \sigma\left(\mathbf{U}^{(n)}\mathbf{S}^{(n)}\mathbf{V}^{\top}\right) \text{ s.t. } {\mathbf{U}^{(n)}}^{\top}\mathbf{U}^{(n)} = \mathbf{\Phi} \ \forall n, \quad (11)$$

where $\widehat{\mathrm{P}}^{(n)}$ is the low-rank approximation of the propensity scores, $\sigma(\cdot)$ denotes the sigmoid function, $\mathbf{U}^{(n)} \in \mathbb{R}^{T_n \times R}$ is the temporal propensity weightings for the n-th patient and R is the target approximation rank. $\mathbf{S}_n = \mathrm{diag}\left(\mathbf{s}_n\right) \in \mathbb{R}^{R \times R}$ is the patient-specific and time-independent diagonal static weighting matrix. $\mathbf{V} \in \mathbb{R}^{J \times R}$ is the propensity factors that are shared for all patients. The invariance constraint ${\mathbf{U}^{(n)}}^{\top}\mathbf{U}^{(n)} = \mathbf{\Phi}$ is imposed to enforce uniqueness of the approximations [37], [41], where $\mathbf{\Phi} \in \mathbb{R}^{R \times R}$ is invariant for all patients. To enable the model to make predictions for patients out of the training set, we learn a subnetwork to produce the patient-dependent components $\mathbf{U}^{(n)}$ and $\mathbf{s}^{(n)}$ based on observations. The factor matrix $\mathbf{V}$ is directly learned as a part of model parameters and kept frozen during test time.

4.2.2 Learning Patient-Dependent Propensity Weightings

Although PARAFAC2 has shown promising in multiple domains [42], including healthcare [22], [37], [43], existing PARAFAC2 methods are limited in modeling the underlying temporal dependency by independently learning the temporal weightings at each time step without considering historical information. This could be suboptimal when applied to infer propensity scores in our setting as the propensity scores could strongly depend on past observations. Take the diagnosis of an existing disease as an example, the propensity score, *i.e.*, the probability of recording the disease in data, could be low at the beginning due to insufficient information available for the doctors to diagnose; over time, the propensity score may become higher, as other features like laboratory test results are gradually observed. On the other hand, if the diagnosis appeared once in the record, the probability of misdiagnosing could be significantly reduced. In other words, the propensity scores may be much higher.

To model such complex temporal dependency, we propose to use the long short-term memory (LSTM) model [44] to parameterize the temporal propensity weightings. It summarizes the historical information of each patient using a hidden state $\mathbf{h}_t$ at the t-th time step and produces the temporal propensity weightings using an MLP, given by:

$$\mathbf{u}_t = \mathrm{MLP}(\mathbf{h}_t) \quad t = 1, \ldots, T, \quad (12)$$

$$[\mathbf{h}_t, \mathbf{c}_t] = \begin{cases} [\mathbf{0}, \mathbf{0}] & \text{if } t = 1, \\ \mathrm{LSTM}(\mathbf{x}_t, [\mathbf{h}_{t-1}, \mathbf{c}_{t-1}]) & \text{otherwise}, \end{cases} \quad (13)$$

where $\mathbf{h}_t$ and $\mathbf{c}_t$ are the hidden states and the cell states of the LSTM for the t-th time step. By stacking the temporal weightings at all time steps, we obtain the temporal weighting matrix by:

$$\mathbf{U}^{(n)} = \left[\mathbf{u}_1^{(n)}; \ldots; \mathbf{u}_{T_n}^{(n)}\right]. \quad (14)$$

Then, we compute the time-independent propensity weightings $\mathbf{s}^{(n)}$ using an MLP with the average temporal weightings over time as its input, by:

$$\mathbf{s}^{(n)} = \mathrm{MLP}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{u}_t\right). \quad (15)$$

In this paper, we consider the regular setting where the intervals between time steps are fixed. Having said so, PATNet can be generalized to irregular settings by replacing the LSTM in Eq. (12) by a time-aware model, *e.g.*, T-LSTM [27]. Due to the difficulties in optimization with the hard invariance constraint on the temporal weightings, we follow the recent work [22] to minimize the following soft uniqueness regularization term:

$$\mathcal{R} = \sum_{n=1}^{N}\frac{\mu}{2}\left\|{\mathbf{U}^{(n)}}^{\top}\mathbf{U}^{(n)} - \Phi\right\|_F^2, \quad (16)$$

where we treat Φ as parameter to be learned during optimization, and μ is a hyperparameter controlling the strength of the regularization.

4.3 Propensity-Adjusted Imputations

Based on our data observation model defined in Section 3 and the Bayes' rule, we can derive the posterior probability of a feature in the latent complete data being positive, *i.e.*, $z_{tj} = 1$, given the observation and historical input by:

$$\begin{aligned}\widehat{z}_{tj} :=& q\left(z_{tj}=1|x_{tj},\mathcal{H}_t\right)\\ =& \mathbb{1}\left[x_{tj}=1\right]p\left(z_{tj}=1|x_{tj}=1,\mathcal{H}_t\right)\\ &+\mathbb{1}\left[x_{tj}=0\right]p\left(z_{tj}=1|x_{tj}=0,\mathcal{H}_t\right)\\ \overset{(a)}{=}& x_{tj}+(1-x_{tj})p\big(z_{tj}=1|x_{tj}=0,\mathcal{H}_t\big),\end{aligned} \tag{17}$$

where $\mathbb{1}[\cdot]$ is an indicator function, and (a) follows from $p\left(z_{tj}=1|x_{tj}=1\right) = 1$. In Eq. (17), when a feature is observed, its ground truth must be one. On the other hand, when the feature is not observed, we compute its estimation $p\big(z_{tj}=1|x_{tj}=0,\mathcal{H}_t\big)$ by:

$$\begin{aligned}p\big(z_{tj}=1|x_{tj}=0,\mathcal{H}_t\big) =& \frac{p\big(z_{tj}=1|\mathcal{H}_t\big)p\big(x_{tj}=0|z_{tj}=1,\mathcal{H}_t\big)}{1-p\big(x_{tj}=1|\mathcal{H}_t\big)}\\ =& \frac{\widetilde{x}_{tj}\left(1-\widehat{\rho}_{tj}\right)}{1-\widetilde{x}_{tj}\widehat{\rho}_{tj}}.\end{aligned} \tag{18}$$

Eq. (18) shows that the propensity estimations play a role of "downscaling adjustment" in computing Eq. (18), and the effect of downscaling becomes heavier with increasing $\widehat{\rho}_{tj}$. This aligns with our intuition that propensity scores for the missing features should be low in order to explain the missingness, or otherwise they would have been observed in the data. Thus, we name imputation obtained by Eq. (17) the propensity-adjusted imputations. We visualize its contour plot in Appendix A, available in the supplemental materials.

Note that directly learning the network that produces $\mathbf{z}_t$ given the current and historical observations ($\mathbf{x}_t$ and $\mathcal{H}_t$) is not feasible due to the lack of supervised information, *i.e.*, ground-truth $\mathbf{z}_t$.

4.4 Missing-Informative Prediction

As shown in Fig. 1, we assume that the label associated with each patient is generated from the complete data $\mathbf{z}$ rather than the partial observations $\mathbf{x}$. Besides, the missing data in clinical setting is often informative as it could indirectly reflect clinicians' judgement of the health conditions of patients at the time of data recorded. As an example [45], missing cholesterol measurements during early visits to a general practitioner (GP) may be because the GP considers the cardiovascular risk too low to request for cholesterol measurement. Therefore, PATNet produces the final prediction based on the propensity-adjusted imputation in Eq. (17) and the propensity scores estimated using Eq. (11-15).

To achieve this, we use another Transformer model which has the same architecture as the imputation subnetwork's. We first obtain a dense representation of the propensity-adjusted imputations by reusing the embedding layer in Eq. (2). Then we concatenate it with the estimated propensity scores as the input to the Transformer model. Formally, we have:

$$\mathbf{e}'_t = \left[\frac{1}{\|\widehat{\mathbf{z}}_t\|_1}\mathbf{W}_{\text{emb}}\widehat{\mathbf{z}}_t + PE_t,\ \widehat{\boldsymbol{\rho}}_t\right] \quad \forall t, \tag{19}$$

$$\widehat{\mathbf{y}}_t = \text{Transformer}\left(\left[\mathbf{e}'_1,\ldots,\mathbf{e}'_{t-1}\right];\Theta_{\text{pred}}\right) \quad \forall t, \tag{20}$$

where Transformer encapsulates Eq. (4-9). The embedding matrix $\mathbf{W}_{\text{emb}}$ and the positional encoding PE are shared with the imputation subnetwork, and Θ_{pred} denotes the set of parameters of the prediction subnetwork. Our preliminary study shows that allowing the predictive task to influence the imputation subnetwork is not beneficial, but instead it decreases the imputation performance, and hence makes the prediction even worse. Thus, we block the backpropagation to $\mathbf{z}$ as indicated in Fig. 2.

The goal of disease progression modeling (DPM) task is to make predictions of the complete data of future events, *i.e.*, $\mathbf{z}$, so we use the propensity-adjusted imputations at the next time step as labels for training the prediction subnetwork by minimizing the cross-entropy loss as follows.

$$\mathcal{L}_{\text{DPM}} = \sum_{n=1}^{N}\sum_{t=1}^{T_n-1}\widehat{\mathbf{z}}_{t+1}\log\widehat{\mathbf{y}}_t + (1-\widehat{\mathbf{z}}_{t+1})\log\left(1-\widehat{\mathbf{y}}_t\right). \tag{21}$$

For the mortality prediction, patients are associated with the binary labels at the end of their observation windows. So we take the prediction at the last time step as the final prediction and minimize the following loss function:

$$\mathcal{L}_{\text{MR}} = \sum_{n=1}^{N} y_n\log\widehat{\mathbf{y}}_{T_n} + (1-y_n)\log\left(1-\widehat{\mathbf{y}}_{T_n}\right). \tag{22}$$

4.5 Learning Algorithms

Due to the lack of ground-truth complete data, we cannot directly learn the imputation subnetwork and the propensity subnetwork. Thus, we adopt an iterative strategy to learn the model. First, we learn the parameters of the imputation and propensity subnetworks via the expectation-maximization (EM) algorithm [35], [46]. It iterates between the E-step and the M-step where the former estimates the posterior distribution of the hidden complete data, and the latter maximizes the expected log likelihood of the models given the estimation found in the E-step. We have presented the estimation of the posterior distribution of $\mathbf{z}_t$ in Eq. (17). In the M-step, the expected log likelihood is maximized as follows:

$$\underset{\Theta_{\text{imp}},\Theta_{\text{prop}}}{\arg\max}\ \frac{1}{N}\sum_{n=1}^{N}\frac{1}{T_n}\sum_{t=1}^{T_n}\sum_{j=1}^{J}\mathbb{E}_{z_{tj}^{(n)}\sim q\left(z_{tj}^{(n)}=1|x_{tj}^{(n)},\mathcal{H}_t^{(n)}\right)}\mathcal{L}_{tj}^{(n)} + \mathcal{R}, \tag{23}$$

with

$$\begin{aligned}\mathcal{L}_{tj}^{(n)} =& \log p\left(z_{tj}^{(n)}, x_{tj}^{(n)}, \mathcal{H}_t^{(n)}, |\Theta_{\text{imp}},\Theta_{\text{prop}}\right)\\ \propto& \widehat{z}_{tj}^{(n)}\log p\left(z_{tj}^{(n)}=1|\mathcal{H}_t^{(n)},\Theta_{\text{imp}}\right)\\ &+\left(1-\widehat{z}_{tj}^{(n)}\right)\log p\left(z_{tj}^{(n)}=0|\mathcal{H}_t^{(n)},\Theta_{\text{imp}}\right)\\ &+\widehat{z}_{tj}^{(n)}\log p\left(x_{tj}^{(n)}|z_{tj}^{(n)}=1,\mathcal{H}_t^{(n)},\Theta_{\text{prop}}\right),\end{aligned} \tag{24}$$

where $\widehat{z}_{tj}^{(n)}$ is the estimation of the posterior distribution given in Eq. (17), and $\mathcal{R}$ is defined in Eq. (16).

After the M-step, we freeze Θ_{imp} and Θ_{prop}, and update the prediction subnetwork by minimizing $\mathcal{L}_{\text{DPM}}$ or $\mathcal{L}_{\text{MR}}$ depending on the prediction task. Note that $\Theta_{\text{imp}}, \Theta_{\text{prop}}$ and Θ_{pred} all contain the embedding matrix $\mathbf{W}_{\text{emb}}$. So it is updated when maximizing the expected log likelihood and minimizing the prediction loss.

We iterate the E-step, M-step, and the predictive subnetwork update step to train the model until it converges. Algorithm 1 summarizes the overall learning procedures. The most computationally demanding component of PATNet is the prediction subnetwork, where the Transformer architecture is used, which is similar to that in HiTANet [12]. Therefore, the overall time complexity of PATNet remains the same as that of HiTANet.

Algorithm 1: Learning Procedure of PATNet

Input: Binary EHR data $\{\mathbf{x}_1^{(n)}, \ldots, \mathbf{x}_{T_n}^{(n)}\}_{n=1}^N$;
Randomly initialize parameters $\Theta_{\text{prop}}, \Theta_{\text{imp}}$, and Θ_{pred};
repeat
  for *each mini-batch* **do**
    /* E-step */
    Estimate $p\big(z_{tj} = 1 | x_{tj} = 0, \mathcal{H}_t\big)$ using Eq. (18);
    Compute the posterior using Eq. (17):

$$\widehat{z}_{tj} \leftarrow x_{tj} + (1 - x_{tj}) p\big(z_{tj} = 1 | x_{tj} = 0, \mathcal{H}_t\big);$$

    /* M-step */
    Compute the patient-dependent propensity weightings $\mathbf{U}^{(n)}$ and $\mathbf{s}^{(n)}$ using Eq. (12–15);
    Approximate the propensity scores by Eq. (11);
    Update Θ_{prop} and Θ_{imp} by maximizing the log-likelihood in Eq. (23) using gradient descent;
    /* Update predictive subnetwork */
    Freeze Θ_{prop} and Θ_{imp}, update Θ_{pred} by minimizing the predictive loss Eq. (21–22);
  end
until *converge or reach maximum number of iterations*;

5 Experiments and Results

5.1 Datasets and Experimental Settings

We evaluate PATNet using two real-world, large-scale, and ICU-related datasets, **MIMIC-III** [47] and **eICU** [48]. MIMIC-III contains data related to over 40,000 patients who stayed in the ICU units at Beth Israel Deaconess Medical Center from 2001 to 2012. eICU contains data from a number of critical care units throughout the continental U.S. in 2014 and 2015. We use the abnormal laboratory tests and the medications collected during the ICU stays of the patients as clinical features. For MIMIC-III, we rely on the abnormality flag in the database for abnormal lab test results extraction. Since neither the reference values nor the abnormality flags are available for laboratory tests in eICU, we treat a laboratory test result as "abnormal" if the result value is smaller than the 20-percentile or greater than the 80-percentile of all records for the corresponding laboratory test item. For both datasets, we extract adult patients having 20 to 100 time steps, and accumulate features recorded during every consecutive eight hours as one time step.

Finally, we extract data of $13,112$ patients and 411 clinical features from MIMIC-III and $10,162$ patients and 365 clinical features from eICU. The average number of time steps of patients in MIMIC-III and eICU datasets are 20 and 16.7, respectively. For the disease progression modeling task, we use the data during the last 40 hours before hospital discharge as the prediction window (held out for evaluation), and use data between hospital admission and the start of the prediction window to train the model. For the mortality prediction task, we use the in-hospital death as the labels. Patients are divided into training, validation and test sets with ratios of 7:1:2 for both tasks. The details of hyperparameter settings are summarized in Appendix C, available in the supplemental materials.

5.2 Missing Data Sampling

To evaluate the performance of PATNet with observation bias, we treat the extracted data as the complete data $\mathbf{Z}$, and manually hold out some positive features as the missing data. We follow a similar procedure used by [49] to generate the element-wise missing weights:

$$\widetilde{\rho}_{tj}^{(n)} = (\mathbf{a}_n^\top \mathbf{w}_1 + \mathbf{b}_t^\top \mathbf{w}_2 + \mathbf{c}_j^\top \mathbf{w}_3)/3,$$
$$\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3 \sim \mathcal{U}[0,1]^M, \quad \mathbf{b}_t, \mathbf{c}_j \sim \mathcal{N}(0,1)^M,$$
$$\mathbf{a}_n \sim \mathcal{N}(\mu_{d_n}, 1)^M \quad (\mu_1 = 1, \ \ \mu_2 = 0, \ \ \mu_3 = -1),$$

where M is the dimension of the random variables $\mathbf{a}_n, \mathbf{b}_t$ and $\mathbf{c}_j$, generated for the n^{th} patient, the t^{th} time step, and the j^{th} clinical feature, respectively. We set $M = 10$ in all experiments. Patients are grouped into three clusters by KMeans using the aggregated features over all time steps, *i.e.*, $\sum_{n=1}^{T_n} \mathbf{z}_t^{(n)}$, and d_n represents the cluster index of the n-th patient. Then, we transform their mean and standard deviation to desired values to control the levels of missing and observation bias, and finally sample the observations by:

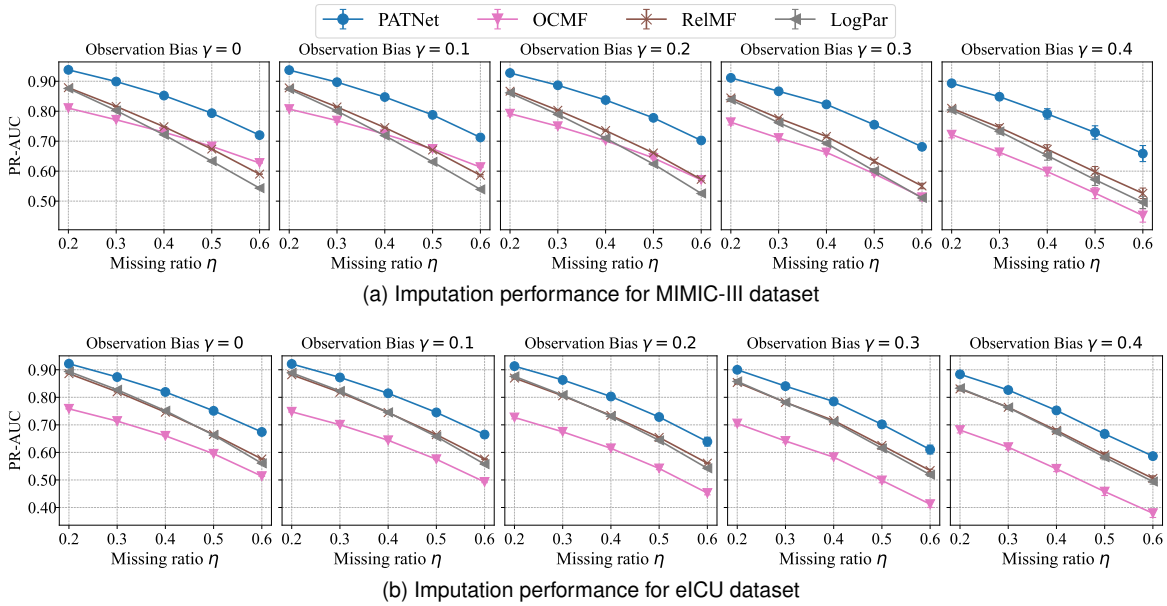
$$x_{tj}^{(n)} \sim \text{Bernoulli}\left(z_{tj}^{(n)} \rho_{tj}^{(n)}\right),$$
$$\rho_{tj}^{(n)} = 1 - \sigma\left(\frac{\widetilde{\rho}_{tj}^{(n)} - \text{mean}(\widetilde{\boldsymbol{\rho}})}{\text{std}(\widetilde{\boldsymbol{\rho}})} * \alpha + \beta\right).$$

We tweak α and β to control $\eta = 1 - \text{mean}(\boldsymbol{\rho})$ and $\gamma = \text{std}(\boldsymbol{\rho})$, where the former is the mean value of the final propensity scores, representing the expected missing ratio. A larger γ indicates that the propensity scores of different features at different time steps for different patients spread out over a wider range, thus indicating a heavier observation bias.

We generate datasets with five different levels of missing ($\eta = 0.2, 0.3, 0.4, 0.5, 0.6$) and five levels of observation bias ($\gamma = 0, 0.1, 0.2, 0.3, 0.4$). When $\gamma = 0$, the propensity scores take constant value of $1 - \eta$, which represents the scenario of missing completely-at-random. The upper bound of γ is close to 0.5 given its range. By inspecting the generated propensity scores, most of the values of $\boldsymbol{\rho}$ are quite close to either zero or one when $\gamma = 0.4$; therefore, we consider $\gamma = 0.4$ the extreme case of heavy observation bias.

5.3 Evaluation of Missing Data Imputation

We first evaluate the imputation performance of PATNet. We train PATNet using training sets with different levels of missing and observation bias. Then we freeze the trained model and generate the propensity-adjusted imputations $\widehat{\mathbf{z}}$ for the test set given their partial observations $\mathbf{x}$. Similar to the existing work [22], we use the Precision-Recall Area Under Curve (PR-AUC) as the evaluation metric due to the binary and imbalanced nature of the data. We measure the PR-AUC score between the ground-truth $\mathbf{z}$ and the imputations $\widehat{\mathbf{z}}$ for each clinical feature and take the average over all clinical features.



(a) Imputation performance for MIMIC-III dataset

(b) Imputation performance for eICU dataset

Fig. 4. Imputation performance measured by PR-AUC score with increasing missing ratio under different levels of observation bias. For both MIMIC-III and eICU datasets, PATNet consistently outperforms baselines and is the most robust model against increasing missing ratio.

5.3.1 Baselines

We compare the data imputation performance of PATNet against the existing binary data imputation methods, including:

- **One-class MF (OCMF)** [50] is a binary matrix factorization/completion model based on sampling zero entries as negative features.
- **LogPar** [22] is a binary data completion method based on PARAFAC2 factorization. It assumes MCAR and enforces local smoothness by a temporal variance regularization.
- **Rel-MF** [24] is a matrix-factorization model to learn recommender systems from binary data with MNAR. It assume that the propensity scores are subject-independent and proportional to the feature frequency.

Note that OCMF and Rel-MF do not handle multiple subjects with temporal input, so we construct a single matrix as their input by concatenating the observed temporal features of all patients.

5.3.2 Results and Discussions

We visualize the PR-AUC scores of missing data imputation for MIMIC-III and eICU datasets under different levels of observation bias in Fig. 4a and Fig. 4b, respectively. PATNet consistently outperforms all baselines in terms of imputation for both datasets in all levels of missingness and observation bias. Particularly, when the levels of missingness and observation bias are both highest, PATNet still obtains PR-AUC of 0.53 for MIMIC-III and 0.51 for eICU, which shows 25.2% and 16% relative improvement against the best performing baselines for MIMIC-III and eICU, respectively. Besides, PATNet is the most robust model against the missing ratio. For example, in MIMIC-III data with the highest level of observation bias ($\gamma = 0.4$), the imputation performance of PATNet only drops 25.8% (from 0.89 to 0.66) as η increases from 0.2 to 0.6, which is the smallest among all models. On the contrary, the second most robust model is Rel-MF, which obtains a relative performance drop of 35.8% (from 0.81 to 0.52). This demonstrates that by explicitly considering the propensity scores and inferring it as a part of the model, PATNet achieves its goal of better handling the observation bias. Rel-MF also considers the propensity score and performs imputation based on the propensity scores, making it the second-best performance when the missing ratio or the observation bias is high for MIMIC-III data. However, it does not infer the propensity score from data during training, but rather use simple frequency statistics to determine the propensity score prior the model learning. The large gap between the performance of Rel-MF and that of PATNet clearly demonstrates that such heuristics is insufficient for complex data like EHR and leads to suboptimal solutions. On the other hand, LogPar and OCMF both assume that missing are completely at random, which leads to the worst imputation performance for large missing ratios and observation bias.

TABLE 1
The propensity recovery error measured by mean absolute error (MAE) when $\eta = 0.4$. Smaller values indicate more accurate recovery of the latent propensity scores.

	$\gamma = 0.1$	$\gamma = 0.2$	$\gamma = 0.3$	$\gamma = 0.4$
Constant 1	0.40±.003	0.40±.004	0.40±.006	0.42±.015
Constant $\rho^\star$	0.11±.001	0.16±.046	0.21±.078	0.23±.114
Rel-MF	0.23±.002	0.27±.005	0.33±.006	0.41±.006
PATNet	**0.09**±.005	**0.13**±.008	**0.17**±.009	**0.17**±.010

We further investigate how accurately we can estimate the propensity scores. As pointed out by the literature in the field of positive-unlabeled (PU) learning with selection bias [35], the accuracy of estimating the propensity scores plays an important role in reducing the bias of the estimator. To empirically evaluate the capability of PATNet to recover the propensity scores, we use the MIMIC-III dataset, set the missing ratio to be $\eta = 0.4$, collect the propensity scores estimated or constructed by different models, and compare them with the ground-truth propensity scores using the mean absolute error (MAE). Table 1 summarizes the results obtained. Higher values indicate larger error in estimating the propensity scores. When a model fails to adjust for

the observation bias as reflected by a large propensity recovery error, it could lead to worse imputation and degraded downstream prediction performance. The first row of Table 1 is obtained by setting the propensity score to be the constant one, *i.e.*, $\rho = 1$. This setting is equivalent to an assumption that the data is fully observed, namely no missing data at all. The second row sets $\rho = \rho^\star$, where $\rho^\star$ is the mean value of the ground-truth propensity scores. This corresponds to the assumption of missing-completely-at-random (adopted by LogPar and OCMF) and the value of $\rho^\star$ represents propensity prior that a model assuming missing-at-random would attain the best propensity recovery error. The third and fourth rows correspond to the propensity scores constructed by Rel-MF baseline and learned by our proposed PATNet, respectively. The results show that PATNet can recover the propensity scores much more accurately than baselines and the existing heuristics. Although Rel-MF considers the entry-wise propensity score, it pre-constructs the propensity scores deviates too much from the ground truth. This also explains the observation in Fig. 4a that LogPar outperforms Rel-MF even though the former assumes missing completely at random.

5.4 Disease Progression Prediction

Predicting disease progression is a fundamental task that could enable timely interventions to lower the risk of patients and improve the quality of care. In this section, we evaluate the performance of PATNet for the disease progression modeling task. Given the historical data within the observation window, the goal is to predict the clinical events that will happen in the subsequent prediction window.

5.4.1 Baselines and Evaluation Metric

We compare the data imputation performance of PATNet against the existing binary data imputation methods, including:

- **Long Short-Term Memory (LSTM)** [44] is a sequential prediction model which has been widely adopted in clinical prediction tasks.
- **RETAIN** [6] is a predictive model tailored for EHR data. It utilizes a recurrent neural network (RNN) model and a two-level reverse time attention model for prediction.
- **Dipole** [7] is a diagnosis prediction model based on attentional bidirectional RNN.
- **Transformer** [34] is a widely adopted sequential modeling method purely based on attention mechanism.
- **HiTANet** [12] is a hierarchical attention networks for predictive analysis of EHR data, which achieves the state-of-the-art performance for EHR prediction.
- **CATNet** [32] is an attention-based medical event prediction model tailored for irregular binary electronic health records data.
- **ODE-LSTM** [31] is a variant of LSTM with its hidden state parameterized by the ordinary differential equation (ODE) which handles the irregularity in time series data.

We measure the performance using the mean average precision (mAP) which is defined as

$$\mathrm{mAP} = \frac{1}{J}\sum_{j=1}^{N} \mathrm{AP}_j,$$

where AP_j is the average precision for the j-th feature.

5.4.2 Results and Discussions

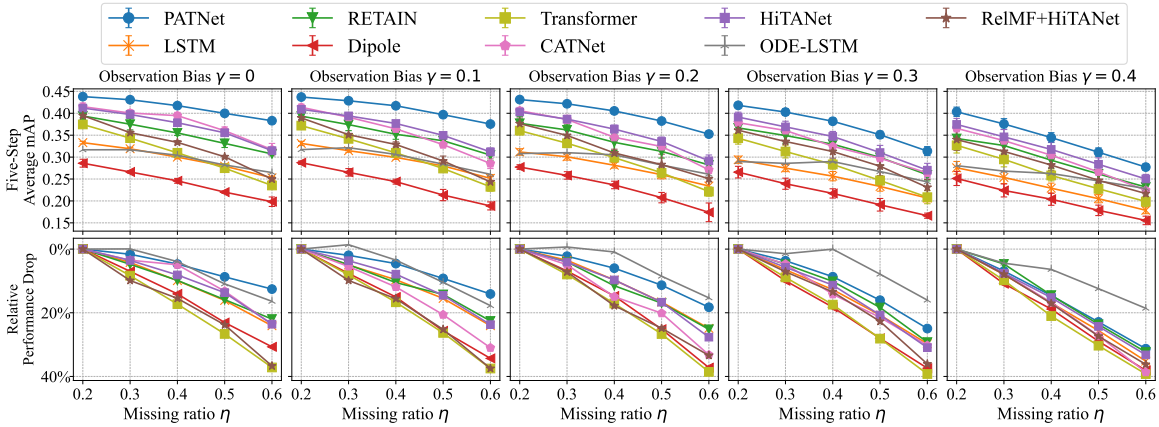
Fig. 5 visualizes the performance of disease progression modeling under different levels of missing ratio and observation bias for MIMIC-III and eICU, respectively. We show the mAP score of the first prediction step and the average mAP score over the five prediction steps, along with the relative performance drop when the missing ratios increase for different levels of observation bias. The results show that PATNet consistently outperforms the baseline models by a large margin for both MIMIC-III and eICU datasets across different levels of missingness and observation bias. In the presence of mild and moderate observation bias, PATNet obtains significant improvement over the best baselines. For example, when $\gamma = 0.1$, PATNet improves the five-step average mAP over the best baseline model by up to 20.3% and 22.7%, for MIMIC-III and eICU, respectively. When $\gamma = 0.2$, the relative improvements are even slightly enlarged to 21% and 23%. When the observation bias further increase, it becomes exceptionally challenging to accurately estimate the propensity scores due to the extreme variance in the missing data sampling process. Having said that, PATNet still achieves up to 11% and 17% relative improvement of the five-step average mAP score over the best baseline for MIMIC-III and eICU, respectively. Another interesting observation is that when $\gamma = 0$, meaning that no extra bias is introduced in the missing data sampling process, PATNet also outperforms all baseline models by a large margin. This indicates that the effectiveness of PATNet in handling the observation bias also applies to the bias that exists in the raw data.

The bottom row of each sub-figure in Fig. 5 shows the relative performance drop for each level of missing ratio compared to the smallest one ($\eta = 0.2$). They clearly demonstrate that PATNet is the most robust one, among all models compared, against increasing missing ratio, when $\gamma < 0.4$. In the extreme case where $\gamma = 0.4$, despite the narrowing gap of the relative performance drop between PATNet and baseline models, PATNet still obtains the smallest drop when the missing ratio exceeds 0.4.

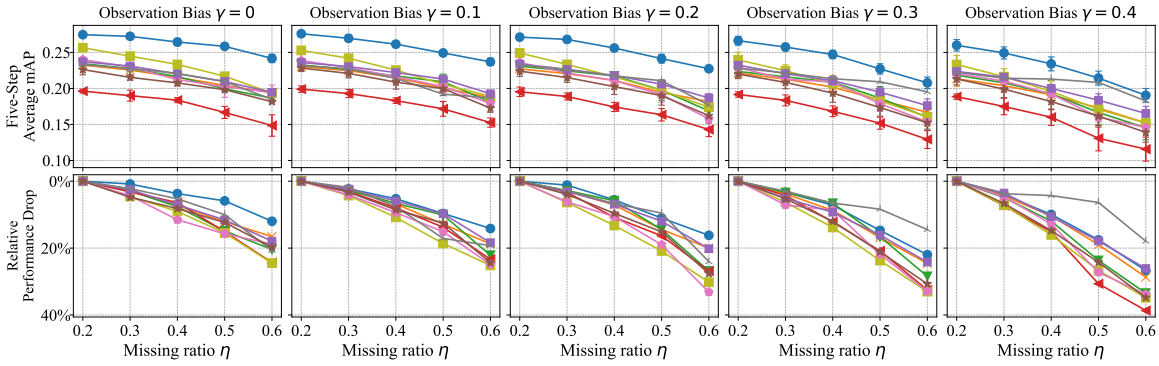
To gain more insights of the significant performance boost, we divide the features into five groups evenly according to their feature-wise observation bias and visualize the distribution of the absolute performance gaps of each group using box plots in Fig. 6. A positive performance gap indicates that PATNet outperforms HiTANet. We also plot a dashed red line representing a linear regression of the median value of each group. The Pearson correlation of the linear regression is 0.95 (with p-value of 0.012), showing a strong positive correlation between the median values of the absolute performance gap and feature-wise observation bias. This shows that the performance boost obtained by PATNet mainly attributes to improving the prediction performance for features with heavy observation bias, implying that PATNet is effective in adjusting for the observation bias by estimating and modeling the propensity scores. We visualize the performance gap for each individual feature in Appendix B, available in the supplemental materials.

5.5 Mortality Prediction

We further conduct the mortality prediction to evaluate the predictive power of the proposed model. Patients in MIMIC-



(a) Disease progression modeling performance for MIMIC-III dataset



(b) Disease progression modeling performance for eICU dataset

Fig. 5. The prediction performance of the disease progression modeling task obtained for MIMIC-III and eICU datasets. The upper rows of two subfigures denote the mAP scores averaged over five prediction steps. The bottom rows of the subfigures show the relative performance drop when the missing ratio increases compared with the lowest missing ratio, which reflects the robustness against increasing missing ratio. PATNet consistently outperforms all baseline models. With mild to moderate observation bias ($\gamma < 0.4$), PATNet achieves the best robustness against missing ratios. With extreme observation bias ($\gamma = 0.4$), the robustness of PATNet is not worse than baselines while achieving much higher absolute performance scores.

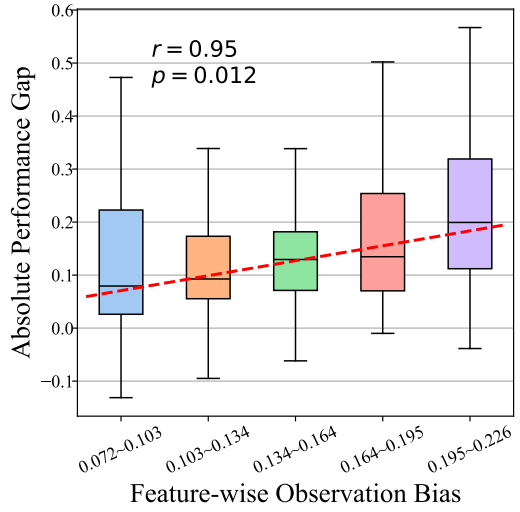
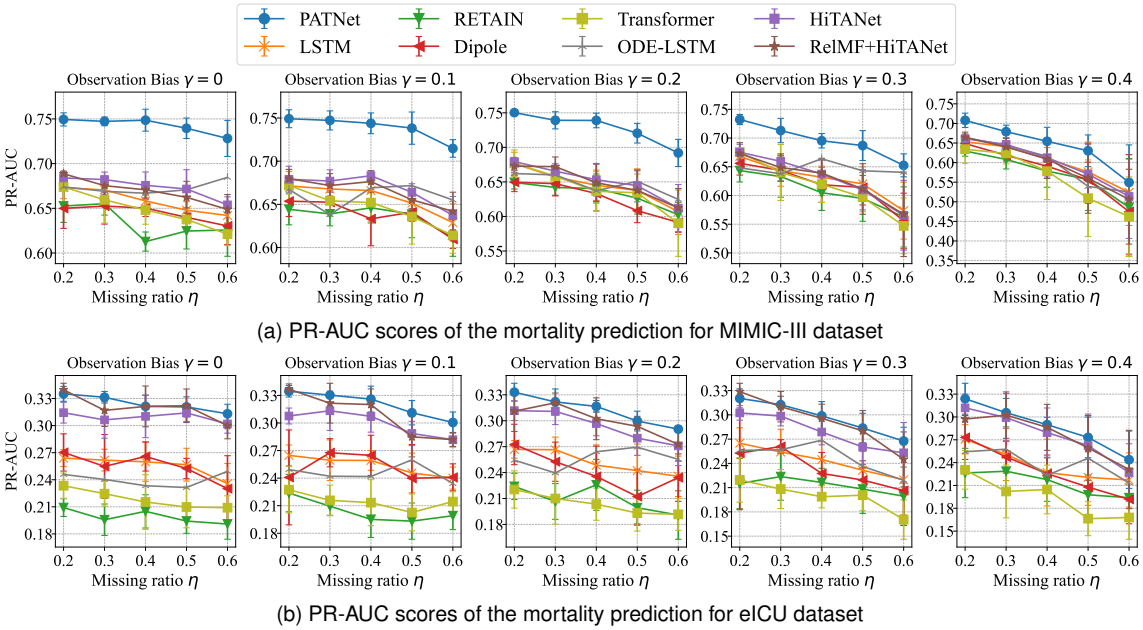


Fig. 6. The box plot of average performance gaps of feature groups with different feature-wise observation bias between PATNet and HiTANet for MIMIC-III dataset when $\eta = 0.4$ and $\gamma = 0.2$. Each box represents a feature group with its range of feature-wise observation bias annotated in below the X axis. The red dashed line shows a linear regression of the median value of the performance gaps. $r = 0.95$ is the Pearson correlation coefficient of the linear regression and the p-value of $p = 0.012$ indicates the significance of the correlation.

III and eICU datasets are associated with labels indicating their hospital discharge status (*i.e.*, alive or expired). We use the records within the observation window to train the models and predict the in-hospital mortality. We use PR-AUC as the evaluation metric.

Fig. 7 visualizes the mortality prediction results obtained for MIMIC-III and eICU datasets. Due to the joint imputation of the missing data, PATNet consistently outperforms all baseline models under all settings of observation bias and missing ratios. In particular, PATNet outperforms the best baseline by a large margin for MIMIC-III dataset with mild to moderate observation bias. For example, when $\gamma = 0.2$, PATNet achieves relative improvement up to 13% over HiTANet, the state-of-the-art and best performing baseline model. However, with extreme observation bias ($\gamma = 0.4$), the performance gaps significantly narrows. We conjecture that this is due to increasing difficulties in accurately estimating the propensity scores and hence the missing data imputation. Nevertheless, PATNet still demonstrates its superiority by achieving the best prediction PR-AUC scores and the smallest variance. On the other hand, mortality prediction for eICU dataset is much more challenging due to more sparse observations: the prediction PR-AUC scores obtained by all models are significantly lower than that obtained for MIMIC-III dataset. Having said that, the results show that PATNet also achieves the best prediction perfor-



(a) PR-AUC scores of the mortality prediction for MIMIC-III dataset

(b) PR-AUC scores of the mortality prediction for eICU dataset

Fig. 7. The performance of mortality prediction task obtained for MIMIC-III and eICU datasets. PATNet outperforms all baselines for both datasets under all levels of missing ratio and observation bias.

TABLE 2
Ablation study results ($\eta = 0.4, \gamma = 0.2$).

		Disease Progression		Mortality Prediction	
		MIMIC-III	eICU	MIMIC-III	eICU
#1	HiTANet	0.36±.009	0.22±.003	0.66±.023	0.30±.014
#2	PATNet-**1**	0.37±.006	0.22±.003	0.68±.009	0.29±.013
#3	PATNet-$\boldsymbol{\rho}^\star$	0.37±.007	0.23±.005	0.70±.007	0.30±.016
#4	PATNet-noLR	0.37±.003	0.23±.002	0.71±.010	0.29±.018
#5	PATNet-noMI	0.40±.003	**0.26**±.006	0.72±.003	0.31±.005
#6	PATNet	**0.41**±.004	**0.26**±.004	**0.74**±.011	**0.32**±.007
#7	HiTANet-C	0.42±.002	0.25±.001	0.74±.001	0.31±.001
#8	PATNet-C	0.44±.001	0.28±.002	0.75±.002	0.33±.002

mance for eICU dataset, implying that joint imputation of missing data is beneficial for prediction tasks.

5.6 Ablation Study

5.6.1 *Effectiveness of Each Component*

To further understand the contributions of different components to the overall performance gain, we conduct a set of ablation studies. We set the missing ratio η to 0.4 and the observation bias γ to 0.2, run PATNet with different modifications, and measure the performance of the disease progression prediction task and the mortality prediction by the five-step average mAP score and the PR-AUC score, respectively. Table 2 summarizes the results. For easy reference, we include the performance obtained by the best performing baseline model, HiTANet, in Row #1. We also run HiTANet and PATNet with complete data (no missingness) and report their performance in Rows #7 (HiTANet-C) and #8 (PATNet-C), respectively. They could serve as upper bounds of the performance in the presence of missing data during the ablation studies for easier comparison.

We first examine how much the missing data affects the predictive analysis by setting the propensity score to constant one in PATNet (PATNet-**1**). In this case, the propensity-adjusted imputation $\hat{z}_{ij}$ shown in Eq. (17) reduces to the input x_{ij}, which is equivalent to the assumption of no missing data. Comparing Rows #1 and #2 of Table 2, we can see that the performance of PATNet-**1** becomes close to HiTANet, which does not account for missing data. Further comparing Rows #2 and #6, it is clear that by joint data imputation, HiTANet significantly improves the performance for both disease progression modeling and mortality prediction. The marginal differences between #7 and #8 also suggest that PATNet's performance is quite close to HiTANet's when the data is complete, also confirming that the major source of gain lies in the accurate imputation of the missing data.

PATNet performs missing data imputation by inferring the low-rank propensity scores. So we further examine the effect of inferring the propensity score and the role of the low rank approximation of the propensity scores. We replace the propensity subnetwork by a constant of the average propensity score $\rho^\star = 0.6$ and train the rest of the model (PATNet-$\boldsymbol{\rho}^\star$). This corresponds to the assumption that all features are missing with the same probability which is time- and feature-independent. The comparison between Rows #2 and #3 of Table 2 reveals that the missing-completely-at-random assumption helps little for disease progression modeling task. This shows that inferring the propensity scores is the key for the performance improvement in PATNet. Mortality prediction for MIMIC-III, on the other hand, is less sensitive to the observation bias; therefore, up to 6% relative improvement can be observed. Yet, it can be further boosted by another 5.7% via inferring the propensity scores as shown in Row #6. We also remove the low rank assumption by replacing the PARAFAC2 approximation with an LSTM (PATNet-noLR). Row #4 of Table 2 clearly shows that the performance is almost identical to PATNet-**1** and PATNet-$\boldsymbol{\rho}^\star$, indicating that the propensity scores cannot be effectively inferred without the low-rank assumption. Row #5 shows the results obtained by another variant of PATNet where the missing-informative prediction is removed (PATNet-noMI), *i.e.*, only the propensity-adjusted imputations are used to make predictions. It shows

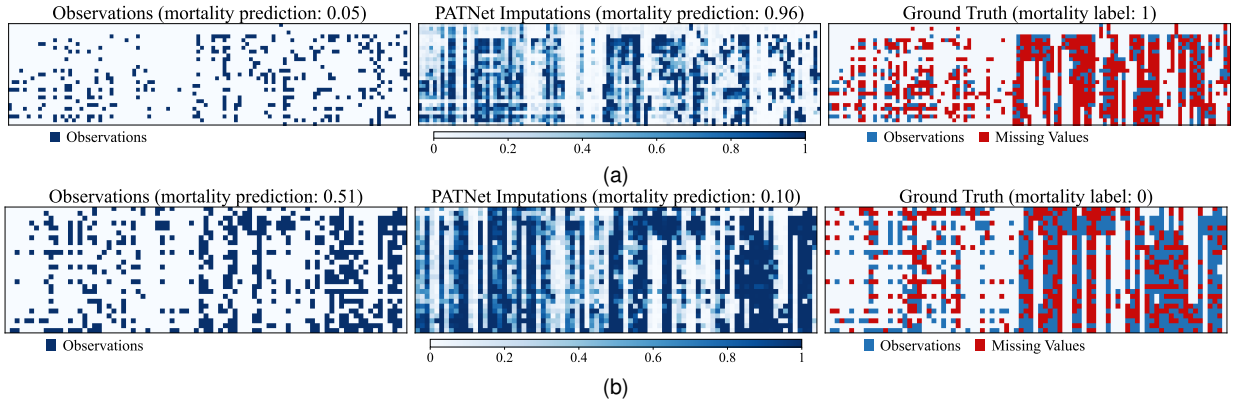


Fig. 8. Examples of the partial observations and the imputations obtained by PATNet corresponding to two patients in MIMIC-III dataset. The rows and columns are time points and clinical features, respectively. The examples show that PATNet can generate accurate imputations and improve prediction performance.

TABLE 3
Comparison between PATNet and RelMF+HiTANet ($\eta = 0.4$).

	$\gamma = 0$	$\gamma = 0.1$	$\gamma = 0.2$	$\gamma = 0.3$	$\gamma = 0.4$
	Mortality Prediction for MIMIC-III				
HiTANet	0.68±.014	0.68±.005	0.66±.023	0.64±.024	0.61±.049
RelMF+HiTANet	0.68±.009	0.68±.012	0.66±.027	0.65±.025	0.62±.043
PATNet	0.75±.012	0.74±.012	0.74±.011	0.70±.012	0.65±.035
	Mortality Prediction for eICU				
HiTANet	0.31±.023	0.31±.017	0.30±.014	0.28±.015	0.28±.032
RelMF+HiTANet	0.32±.022	0.32±.017	0.30±.024	0.30±.017	0.29±.030
PATNet	0.32±.012	0.33±.014	0.32±.007	0.30±.017	0.29±.009
	Disease Progression for MIMIC-III				
HiTANet	0.38±.003	0.38±.002	0.36±.009	0.35±.011	0.32±.017
RelMF+HiTANet	0.36±.003	0.34±.005	0.31±.008	0.31±.015	0.28±.013
PATNet	0.42±.002	0.42±.001	0.41±.004	0.38±.007	0.34±.011
	Disease Progression for eICU				
HiTANet	0.22±.003	0.22±.002	0.22±.003	0.21±.006	0.20±.005
RelMF+HiTANet	0.21±.004	0.21±.009	0.20±.010	0.19±.012	0.18±.011
PATNet	0.26±.005	0.26±.003	0.26±.004	0.25±.005	0.23±.009

that utilizing the informative missing is beneficial in predictive analysis, especially for the mortality prediction task.

5.6.2 *Benefits of Joint Imputation and Prediction*

An alternative and commonly adopted approach to handle the missing data is to impute the missing values first and make predictions based on the frozen imputations. To gain deeper insights into the benefits of the proposed joint imputation and prediction method, we set the missing ratio η to 0.4, impute the missing observations using RelMF which is the baseline model with the best performance in binary data imputation, and then separately training the HiTANet model using the imputations. The results are presented in Table 3. The mortality prediction is relatively less affected by observation bias and RelMF+HiTANet slightly outperforms HiTANet. However, observation bias has a much stronger negative impact on the disease progression modeling task, and RelMF+HiTANet yields worse results than HiTANet for both datasets. This suggests that a two-phase imputation could harm the performance in the presence of severe observation bias. Notably, PATNet outperforms the two-phase approach by a large margin. This can be attributed to its joint imputation and prediction methodology that enables gradual enhancement of the accuracy of propensity score estimation and the imputation during training, which is partially driven by the predictive loss and in turn leads to improved predictive performance.

5.7 Case Studies

We conduct case studies to gain more insights into the performance improvement. Fig. 8 illustrates two examples from the MIMIC-III dataset with the partial observations (we set η=0.4 and γ=0.2 for the case study) and the imputations obtained by PATNet. The three plots in each subfigure are the partial observations, the imputations, and the ground truth of the held-out missing data, respectively. The numbers above them show the predicted probability of in-hospital mortality using the partial observations and PATNet imputations, and the ground-truth mortality labels. The figures show that PATNet generates accurate imputations and greatly improves the mortality prediction performance. Take Fig. 8a as an example. The most heavily missing features include abnormal laboratory test results of *blood pH* (100% missing), *blood glucose* (86% missing), *blood RDW* (79% missing), *blood hemoglobin* (69% missing), *blood pO2* (74% missing) and *blood pCO2* (88% missing). All of them have been shown closely related to mortality in ICU [51], [52], [53]. With these important predictors heavily missing, the patient is predicted to have a mortality probability of 0.05. Among these missing values, PATNet successfully imputes 93% of the *pH*, 86% of the *glucose*, 100% of the *RDW*, 100% of the *hemoglobin*, 90% of the *pO2*, and 75% of the *pCO2*. Thus, the patient with PATNet imputation was predicted a mortality probability of 0.96, which is in line with the ground-truth label. The example in Fig. 8b, on the other hand, has several medication prescriptions heavily missing (e.g., *calcium gluconate*, *potassium chloride* and *insulin*). Without knowing that these medications are used, the model outputs a mortality probability of 0.51 for this patient. Again, PATNet successfully imputes at least 90% of these missing medication prescriptions. With this information available to the predictive model, the output of mortality probability prediction is lowered to 0.1.

6 Conclusion

In this paper, we introduce PATNet, a propensity-adjusted temporal network for missing data imputation and predictive analysis of partially observed binary electronic health

records with observation bias. Unlike existing predictive models that assume complete observation or missing data that are completely at random, we explicitly take into account the observation bias arising from the not-completely-random missingness by modeling and estimating the propensity scores. We propose to learn a transformer-based imputation subnetwork to generate an initial imputation given historical observations and learn a propensity subnetwork coupled with a low-rank PARAFAC2 approximation to adjust for the observation bias incurred by varying missing probabilities. Finally, we train the prediction subnetwork to produce the missing-informative predictions, conditioning on the propensity-adjusted imputations and the inferred feature-, time- and patient-dependent missing probabilities. We conduct extensive experiments using two publicly available datasets, and the results show that PATNet achieves better performance in data imputation than existing binary data imputation methods. PATNet also consistently outperforms the state-of-the-art predictive analysis models in terms of disease progression modeling and mortality prediction tasks. The ablation study confirms that the performance improvement mainly attributes to the consideration and estimation of the propensity scores. One limitation of PATNet is that it assumes regular input and does not tackle the issue of irregularly sampled observations with varying time intervals. We will address this in future work. Besides, we will also focus on further modeling the relationship between the patients' health states and the not-at-random missing mechanism to enhance the imputation and prediction performance in the presence of extreme observation bias.

ACKNOWLEDGMENTS

This research is partially supported by General Research Fund RGC/HKBU12201219 and RGC/HKBU12202117 from the Research Grants Council of Hong Kong.

REFERENCES

[1] X. Zhang, B. Qian, Y. Li, S. Cao, and I. Davidson, "Context-aware and time-aware attention-based model for disease risk prediction with interpretability," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[2] Y. An, L. Zhang, H. Yang, L. Sun, B. Jin, C. Liu, R. Yu, and X. Wei, "Prediction of treatment medicines with dual adaptive sequential networks," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[3] Y. Zhang, X. Yang, J. Ivy, and M. Chi, "Attain: attention-based time-aware LSTM networks for disease progression modeling," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 4369–4375.

[4] A. M. Alaa and M. van der Schaar, "Attentive state-space modeling of disease progression," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 11 338–11 348.

[5] K. Yin, W. K. Cheung, B. C. M. Fung, and J. Poon, "TedPar: Temporally dependent PARAFAC2 factorization for phenotype-based disease progression modeling," in *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*. SIAM, 2021.

[6] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun, "RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016, pp. 3512–3520.

[7] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. 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*, 2017, pp. 1903–1911.

[8] T. Fu, T. N. Hoang, C. Xiao, and J. Sun, "DDL: Deep dictionary learning for predictive phenotyping," in *28th International Joint Conference on Artificial Intelligence, IJCAI 2019*. International Joint Conferences on Artificial Intelligence, 2019, pp. 5857–5863.

[9] K. Yin, D. Qian, W. K. Cheung, B. C. M. Fung, and J. Poon, "Learning phenotypes and dynamic patient representations via RNN regularized collective non-negative tensor factorization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 1246–1253.

[10] K. Yin, W. K. Cheung, B. C. Fung, and J. Poon, "Learning inter-modal correspondence and phenotypes from multi-modal electronic health records," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 09, pp. 4328–4341, 2022.

[11] M. H. Gorelick, "Bias arising from missing data in predictive models," *Journal of Clinical Epidemiology*, vol. 59, no. 10, pp. 1115–1123, 2006.

[12] J. Luo, M. Ye, C. Xiao, and F. 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*, 2020.

[13] J. C. Ho, C. H. Lee, and J. Ghosh, "Septic shock prediction for patients with missing data," *ACM Transactions on Management Information Systems (TMIS)*, vol. 5, no. 1, pp. 1–15, 2014.

[14] T. Köse, S. Özgür, E. Coşgun, A. Keskinoğlu, and P. Keskinoğlu, "Effect of missing data imputation on deep learning prediction performance for vesicoureteral reflux and recurrent urinary tract infection clinical study," *BioMed Research International*, 2020.

[15] Z. Hu and D. Du, "A new analytical framework for missing data imputation and classification with uncertainty: Missing data imputation and heart failure readmission prediction," *PloS One*, vol. 15, no. 9, p. e0237724, 2020.

[16] S. W. J. Nijman, T. K. J. Groenhof, J. Hoogland, M. L. Bots, M. Brandjes, J. J. Jacobs, F. W. Asselbergs, K. G. Moons, and T. P. Debray, "Real-time imputation of missing predictor values improved the application of prediction models in daily practice," *Journal of Clinical Epidemiology*, vol. 134, pp. 22–34, 2021.

[17] W. Cao, D. Wang, J. Li, H. Zhou, Y. Li, and L. Li, "BRITS: bidirectional recurrent imputation for time series," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 6776–6786.

[18] K. Yin and W. K. Cheung, "Context-aware imputation for clinical time series," in *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 2019, pp. 1–3.

[19] E. Jun, A. W. Mulyadi, J. Choi, and H.-I. Suk, "Uncertainty-gated stochastic sequential model for EHR mortality prediction," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 9, pp. 4052–4062, 2020.

[20] Y. Ren, J. Lou, L. Xiong, and J. C. Ho, "Robust irregular tensor factorization and completion for temporal health data analysis," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 1295–1304.

[21] M. Gupta, T.-L. T. Phan, H. T. Bunnell, and R. Beheshti, "Concurrent imputation and prediction on ehr data using bi-directional GANs: Bi-GANs for EHR imputation and prediction," in *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2021, pp. 1–9.

[22] K. Yin, A. Afshar, J. C. Ho, W. K. Cheung, C. Zhang, and J. Sun, "LogPar: Logistic PARAFAC2 factorization for temporal binary data with missing values," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2020, pp. 1625–1635.

[23] C.-J. Hsieh, N. Natarajan, and I. Dhillon, "PU learning for matrix completion," in *International Conference on Machine Learning*. PMLR, 2015, pp. 2445–2453.

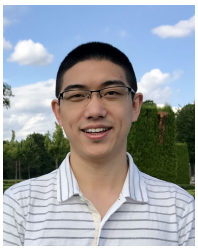
[24] Y. Saito, S. Yaginuma, Y. Nishino, H. Sakata, and K. Nakata, "Unbiased recommender learning from missing-not-at-random implicit feedback," in *Proceedings of the 13th International Conference on Web Search and Data Mining*, 2020, pp. 501–509.

[25] B. J. Wells, K. M. Chagin, A. S. Nowacki, and M. W. Kattan, "Strategies for handling missing data in electronic health record derived data," *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, vol. 1, no. 3, 2013.

[26] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2019, vol. 793.

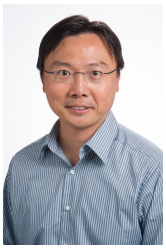
[27] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou, "Patient subtyping via time-aware LSTM networks," in *Proceedings*

of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 65–74.
[28] S. N. Shukla and B. Marlin, "Multi-time attention networks for irregularly sampled time series," in *Proceedings of the International Conference on Learning Representations*, 2021.
[29] X. Zhang, M. Zeman, T. Tsiligkaridis, and M. Zitnik, "Graph-guided network for irregularly sampled multivariate time series," in *International Conference on Learning Representations*, 2022.
[30] M. Habiba and B. A. Pearlmutter, "Neural odes for informative missingess in multivariate time series," in *2020 31st Irish Signals and Systems Conference (ISSC)*. IEEE, 2020, pp. 1–6.
[31] M. Lechner and R. Hasani, "Learning long-term dependencies in irregularly-sampled time series," *arXiv preprint arXiv:2006.04418*, 2020.
[32] S. Liu, X. Wang, Y. Xiang, H. Xu, H. Wang, and B. Tang, "CAT-Net: Cross-event attention-based time-aware network for medical event prediction," *Artificial Intelligence in Medicine*, vol. 134, p. 102440, 2022.
[33] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.
[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
[35] J. Bekker, P. Robberechts, and J. Davis, "Beyond the selected completely at random assumption for learning from positive and unlabeled data," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2019, pp. 71–85.
[36] M. Udell and A. Townsend, "Why are big data matrices approximately low rank?" *SIAM Journal on Mathematics of Data Science*, vol. 1, no. 1, pp. 144–160, 2019.
[37] A. Afshar, I. Perros, E. E. Papalexakis, E. Searles, J. Ho, and J. Sun, "COPA: Constrained PARAFAC2 for sparse & large datasets," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 2018, pp. 793–802.
[38] M. Yuan and C.-H. Zhang, "On tensor completion via nuclear norm minimization," *Foundations of Computational Mathematics*, vol. 16, no. 4, pp. 1031–1068, 2016.
[39] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2862–2869.
[40] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.
[41] R. Bro, C. A. Andersson, and H. A. Kiers, "PARAFAC2–Part II. Modeling chromatographic data with retention time shifts," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 13, no. 3-4, pp. 295–309, 1999.
[42] B. M. Wise, N. B. Gallagher, and E. B. Martin, "Application of PARAFAC2 to fault detection and diagnosis in semiconductor etch," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 15, no. 4, pp. 285–298, 2001.
[43] I. Perros, E. E. Papalexakis, R. Vuduc, E. Searles, and J. Sun, "Temporal phenotyping of medically complex children via PARAFAC2 tensor factorization," *Journal of Biomedical Informatics*, vol. 93, p. 103125, 2019.
[44] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[45] R. H. Groenwold, "Informative missingness in electronic health record systems: the curse of knowing," *Diagnostic and Prognostic Research*, vol. 4, no. 1, pp. 1–6, 2020.
[46] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.
[47] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160035, 2016.
[48] T. J. Pollard, A. E. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, "The eICU collaborative research database, a freely available multi-center database for critical care research," *Scientific Data*, vol. 5, 2018.
[49] W. Ma and G. H. Chen, "Missing not at random in matrix completion: The effectiveness of estimating missingness probabilities under a low nuclear norm assumption," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
[50] H.-F. Yu, M. Bilenko, and C.-J. Lin, "Selection of negative samples for one-class matrix factorization," in *Proceedings of the 2017 SIAM International Conference on Data Mining*. SIAM, 2017, pp. 363–371.
[51] S. Samanta, R. K. Singh, A. K. Baronia, P. Mishra, B. Poddar, A. Azim, and M. Gurjar, "Early ph change predicts intensive care unit mortality," *Indian Journal of Critical Care Medicine*, vol. 22, no. 10, p. 697, 2018.
[52] S. J. Finney, C. Zekveld, A. Elia, and T. W. Evans, "Glucose control and mortality in critically ill patients," *Journal of the American Medical Association*, vol. 290, no. 15, pp. 2041–2047, 2003.
[53] Z. Zhang, X. Xu, H. Ni, and H. Deng, "Red cell distribution width is associated with hospital mortality in unselected critically ill patients," *Journal of Thoracic Disease*, vol. 5, no. 6, p. 730, 2013.



Kejing Yin received the Ph.D. degree in Computer Science from Hong Kong Baptist University in Hong Kong in 2021. He is currently a Research Assistant Professor of the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His current research interests include machine learning and data mining, and their applications to healthcare. He has published papers in refereed journals and conference proceedings, including IEEE TKDE, SIGKDD, SDM, AAAI, IJCAI, etc.

Dong Qian is a Ph.D. student in the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His research interests include machine learning, deep learning and their applications to sequence generation.



William K. Cheung received the Ph.D. degree in computer science from the Hong Kong University of Science and Technology in Hong Kong in 1999. He is currently a Professor of the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His current research interests include artificial intelligence, data mining, collaborative information filtering, social network analysis, and healthcare informatics. He has served as the Co-Chairs and Program Committee Members for a number of international conferences and workshops, as well as Guest Editors of journals on areas including artificial intelligence, Web intelligence, data mining, Web services, e-commerce technologies, and health informatics. From 2002-2018, he was on the Editorial Board of the IEEE Intelligent Informatics Bulletin. He is currently a Track Editor of Web Intelligence Journal and an Associate Editor of Journal of Health Information Research, and Network Modeling and Analysis for Health Informatics and Bioinformatics.