Resolving the Bias in Electronic Medical Records

Kaiping Zheng
National University of Singapore
kaiping@comp.nus.edu.sg

Jinyang Gao
National University of Singapore
jinyang.gao@comp.nus.edu.sg

Kee Yuan Ngiam
National University Health System
kee_yuan_ngiam@nuhs.edu.sg

Beng Chin Ooi
National University of Singapore
ooibc@comp.nus.edu.sg

Wei Luen James Yip
National University Health System
james_yip@nuhs.edu.sg

ABSTRACT

Electronic Medical Records (EMR) are the most fundamental resources used in healthcare data analytics. Since people visit hospital more frequently when they feel sick and doctors prescribe lab examinations when they feel necessary, we argue that there could be a strong bias in EMR observations compared with the hidden conditions of patients. Directly using such EMR for analytical tasks without considering the bias may lead to misinterpretation.

To this end, we propose a general method to resolve the bias by transforming EMR to regular patient hidden condition series using a Hidden Markov Model (HMM) variant. Compared with the biased EMR series with irregular time stamps, the unbiased regular time series is much easier to be processed by most analytical models and yields better results. Extensive experimental results demonstrate that our bias resolving method imputes missing data more accurately than baselines and improves the performance of the state-of-the-art methods on typical medical data analytics.

CCS CONCEPTS

• **Mathematics of computing** → *Time series analysis*; • **Computing methodologies** → *Learning in probabilistic graphical models*; • **Applied computing** → *Health informatics*;

KEYWORDS

Healthcare, Time series, Data analytics

ACM Reference format:
Kaiping Zheng, Jinyang Gao, Kee Yuan Ngiam, Beng Chin Ooi, and Wei Luen James Yip. 2017. Resolving the Bias in Electronic Medical Records. In *Proceedings of KDD '17, Halifax, NS, Canada, August 13-17, 2017,* 10 pages.
https://doi.org/10.1145/3097983.3098149

1 INTRODUCTION

A large amount of heterogeneous medical data has become available in various healthcare organizations. Electronic Medical Records (EMR) are the fundamental resource to help derive healthcare insights and provide more effective healthcare. In the past, medical professionals performed most clinical tasks based on their rich experiences, and medical researchers and clinicians conducted clinical researches via painstaking designed and costly experiments. In recent years, the rapidly increasing availability of EMR is becoming the driving force for adopting data-driven approaches, which provides greater opportunities to automate medical practices. Expected benefits may include more accurate diagnosis as well as prognosis, clinical research breakthrough and improved patient management.

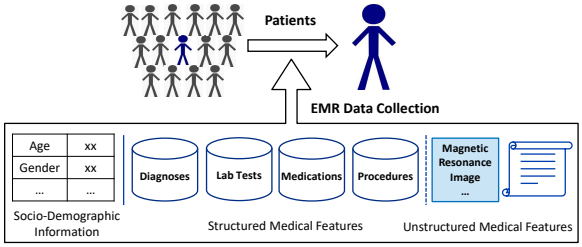

Figure 1: EMR data including socio-demographic information, structured and unstructured medical features.

EMR data[1] is time series data that records patients' visits to hospitals. As illustrated in Figure 1, EMR data typically includes socio-demographic information, heterogeneous medical features such as diagnoses, lab tests, medications, procedures, unstructured data like image data (e.g., magnetic resonance imaging (MRI) data) and text data (e.g., doctors' notes), etc. EMR data is commonly abstracted as a multivariate time series where each time point is a patient's visit, and the variable dimension consists of all the medical features. As a consequence, voluminous researches [3, 22] have been devoted to using advanced time series models to analyze EMR data for various tasks such as ICU patient mortality prediction, ICU patient diagnosis, and disease progression modelling.

However, compared with the traditional time series data, EMR data has its own peculiar characteristics. For most time series data, the time points can be viewed as regularly or randomly sampled from the timeline. In contrast, for EMR data, patients tend to visit hospital more often when they feel sick, and doctors tend to prescribe the lab examinations that show abnormality. This can be considered as a kind of **"bias"**. Hence, the sampling process for each medical feature in the timeline is not only irregular, but also biased. Without understanding such phenomenon and performing a remedy, the analytical models may result in serious misinterpretation about the input EMR data. As an illustration, a young patient who has two visits about respiratory infections with an interval of

[1]For ease of reference, we shall refer to EMR as EMR data.

KDD '17, August 13-17, 2017, Halifax, NS, Canada

ACM ISBN 978-1-4503-4887-4/17/08.
https://doi.org/10.1145/3097983.3098149

six months should not be considered having the respiratory infection status over six months, as "no visits during the six months" is actually a strong indicator that the disease may have been cured. Meanwhile, for some other features such as chronic kidney disease (CKD), the conclusion of cure cannot be made even there are no visits for years since it is generally believed that the disease cannot be cured unless given renal transplantation. In essence, the bias and irregularity for different medical features may have different types of characteristics, depending on the severity or pain extent of diseases, the sensitivity and specificity of lab tests, etc.

In this paper, we propose a general method to transform the biased irregularly sampled EMR data into unbiased regular time series. We create multivariate time series with a regular time interval, and the hidden condition for each medical feature at each time point is learned by an inference model. Since there exists natural uncertainty for medical features at those time points when there are no explicit observations, the output of each medical feature is represented as a time series of distributions over its possible values. By doing such a transformation, we can make the best from all state-of-the-art time series analytical models such as Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU) and their variants, which take regular unbiased time series as input. Therefore, the transformation can serve as a basis for various analytical tasks that take EMR time series as input.

Our proposed inference model mainly considers two characteristics about medical features, namely **Condition Change Rate** and **Observation Rate**. The condition change rate refers to how a medical feature is likely to change from its condition in previous observations. As an illustration, the high glucose level is a medical feature that may change even during a short period, as food can easily affect glucose level. In contrast, diabetes, being a chronic disease, is a medical feature that is hard to change. The observation rate refers to the probability of one medical feature being observed at a time point based on its actual condition (e.g. negative/normal, positive/abnormal, etc.). For instance, the observation rate for acute kidney injury (AKI) will be much higher in the active period, as the patient may have many significant symptoms such as fatigue, loss of appetite, headache, nausea and vomiting, and may visit the hospital in such a situation. However, this phenomenon may be much less significant for the first mildest stage of CKD (CKD Stage I), as CKD is initially without specific symptoms. Our inference model is an HMM variant that estimates both the condition change rate and the observation rate for each medical feature, and infers the hidden condition variable for each time point. Roughly speaking, to predict the hidden condition of one medical feature at a certain time point, the model is likely to infer more from the previous and subsequent observations when the condition change rate is low, and tends to assign the default normal value when the observation rate for the abnormal condition is much higher than the normal condition.

To summarize, our paper has made the following contributions.

- We identify and formalize the bias and irregularity characteristics in EMR data, which is a major challenge on improving the healthcare analytical performance based on traditional time series analytical models.
- We propose a general method to transform the biased and irregularly sampled EMR data into unbiased regular time series. The inference model takes evidence from two parts of observations. One is condition change rate reflecting how the actual condition of one medical feature is likely to change from past observations. The other one is observation rate measuring how one medical feature is likely to be observed based on its actual condition.
- We evaluate the effectiveness of our bias resolving technique in two designed experiments. The first evaluates that our imputation method achieves the highest accuracy. The second demonstrates our method's benefits for analytical tasks including ICU patient mortality prediction, ICU patient diagnosis by category and CKD patient disease progression modelling. Compared with baselines, with the same analytical model, the performance can be improved by using unbiased, regular time series generated by our bias resolving method.

The rest of this paper is organized as follows. Section 2 reviews related works. In Section 3, we define the problem. Section 4 describes our model for resolving bias in EMR data, which considers the condition change rate and the observation rate. In Section 5, we summarize the experimental evaluation of our proposed method against existing proposals. Finally, we conclude in Section 6.

2 RELATED WORK

2.1 EMR Data Analytics

EMR data captures patients' visits to hospitals. The fact that patients tend to visit hospitals when they feel unwell causes EMR data to be highly irregular. For example, the time interval between consecutive visits of a patient varies greatly, resulting in an inherently diverse frequency of visits. Existing works alleviate such a problem using the following three broad categories of methods.

The first category utilizes patients' baseline features (i.e., medical features recorded during patients' first visit to the hospital) for analytical tasks. Some of them are based on regression models [8, 20, 21]. Duchesne et al. [8] employ a robust linear regression model to use baseline MRI features for predicting patients' one-year changes of MMSE scores (a lab test related to patients' mild cognitive impairment). Stonnington et al. [21] utilize MRI scans to predict patients' clinical scores via a relevance vector regression model. Similarly, Schulze et al. [20] collect patients' baseline features and then employ a Cox regression model to predict their development of Type 2 Diabetes. A different line of research [16, 23, 24] is based on multi-task learning [2]. The key idea of multi-task learning in clinical analytics is to capture the intrinsic relationship between tasks, i.e., patients' severity at consecutive future time points. Zhou et al. [24] target at predicting patients' severity and selecting a common set of features significant to all tasks. Zhou et al. [23] add a functionality of selecting task-specific features and Nie et al. [16] take into account the consistency in prediction results among multiple modalities. The performance of this method category may be affected by under-utilization of time-related features. Since patients' medical conditions tend to change over time, such methods tend to benefit more from utilizing more time-related features.

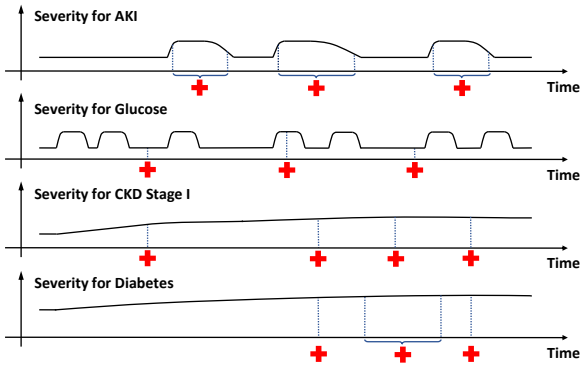

Figure 2: Examples of four representative medical features.

The second category of proposals aims at incorporating as many time-related medical features as possible for data analytics. For instance, Pham et al. [17] modify the standard LSTM model which is specified for regular sequential data to add an irregular time interval term in the form of a monotonic decay or a full time-parameterization in the model computation. Similarly, Che et al. [4] propose to add a time interval term based on GRU which has been designed for regular time series so that the model could handle irregular EMR time series data. However, these works process the "time" information in a heuristic manner, by either using a monotonically decreasing function or a full matrix for learning time's weights. Hence, such methods may cause over-parameterization or under-parameterization in modelling time intervals. Furthermore, these works are based on an end-to-end implementation, which may be specific to modifying a certain kind of models. Therefore, the generality of these works may be affected.

The third category of related works, to which our work belongs[2], focuses on transforming the irregular EMR time series data into a regular one through resampling the data in the time dimension into disjoint windows [5, 14, 15, 22] or overlapping windows [3]. Due to the peculiar characteristics of EMR data, this transformation process will introduce missing data problem. While some proposals employ relatively simple imputation methods, including simple combination [22], majority value imputation for binary variables and mean imputation for others [5], forward and backward imputation [3, 14], some others employ more advanced methods such as adding a missing data indicator to forward imputation of features to capture the missingness in EMR data [15]. However, these methods fail to consider the bias in EMR data and fill the missing data without understanding the reason for missingness, and hence, may cause misinterpretation of EMR data. As a consequence, the overall analytical performance might be degraded.

Our proposed method is different from existing works in that we take the bias in EMR data into consideration, capture the two representative behaviours, condition change rate and observation rate of medical features, and impute the missing data in a more accurate manner.

[2]We propose our method in this category based on two considerations: (i) we hope to avoid the possible under-utilization of time-related features involved in the first category; (ii) instead of proposing an end-to-end implementation, we aim to propose a general method which does not depend on specific models.

2.2 Time Series Analytics via Deep Learning

Deep learning [9, 13] has attracted a huge amount of interests from both industry and academia in recent years due to its excellent accuracy in analyzing and recognizing images, audios, videos, speeches, etc. One category of deep learning models that may be of use to EMR data is Recurrent Neural Networks (RNN), as it has been designed specifically for modelling sequential data and capturing dynamic behaviour in data. Among the RNN models, LSTM [10] and GRU [6] are widely applied and are proven effective for time series. To be specific, in the area of EMR data analytics, researchers have also shown interests in employing deep learning models in various applications, ranging from ICU patient in-hospital mortality prediction, classification and diagnosis [3–5, 14, 15] to analytical tasks for general patients such as unplanned readmission prediction [17] etc.

2.3 Bias in EMR Data

Bias in EMR data is caused by the fact that the data is not captured with fixed frequency due to the natural occurrence of medical events. For example, patients only visit the hospitals when they are sick, and clinicians tend to measure sick patients on more related features.

In [18], the biases in laboratory test results are identified via checking the relationship between lab test value and the time to the next same lab test, and then mitigated through separating different lab test patterns. However, this can only resolve the coarse-grained bias and leave the intra-pattern bias unresolved. In [11], three proposed time parameterization methods are proposed and compared, but the proposed method is heuristic in nature.

To the best of our knowledge, there are no existing works that not only identify the bias challenge of EMR data, but also solve it. In this paper, we propose a general method to transform the biased, irregular EMR time series into an unbiased, regular one, as a means to reduce the effect of misinterpretation of EMR data and further improve the overall analytical performance.

3 PROBLEM FORMULATION

3.1 Examples

We observe that different medical features have different intrinsic characteristics, for instance, (i) whether the feature tends to change frequently and sharply over time; (ii) if the value of feature indicates the abnormal condition, and whether this will cause the patient to visit the hospital.

We shall elaborate our observations through four representative medical features as shown in Figure 2. For each medical feature, we show the changing trend of the severity of its value over time. We draw a red cross in the timeline to denote that the patient visits the hospital at that time point and if a patient stays in the hospital for a period (in-patient case), we draw a red cross beneath the period.

For both AKI and glucose, the severity has a higher probability of changing from the previous condition. One difference between these two lies in that once AKI is severe, the patient tends to visit the hospital for this reason. However, for glucose, this likelihood is relatively lower and sometimes, the patient's visit to the hospital is not related to glucose. For both CKD Stage I and diabetes, the

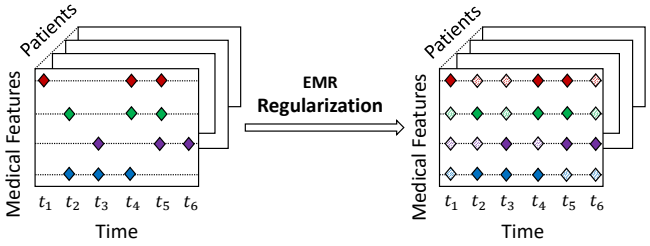

Figure 3: EMR regularization.

severity will not change frequently or sharply, and when the medical feature indicates abnormal, patients may not visit hospital due to this reason.

3.2 Problem Definition

To resolve the bias in EMR data, the ideal case is to use the exact patients' hidden conditions as the input of analytical models.

Definition 1. *(Patients' Hidden Conditions Φ) Given a set of patients Ω_P with cardinality P, a set of regular time points Ω_T with cardinality T and a set of distinct medical features Ω_D with cardinality D, patients' hidden states for each time point and for each medical feature is defined as $\Phi = \{< p, t, d, v > | \forall p \in \Omega_P, \forall t \in \Omega_T, \forall d \in \Omega_D,$ v is the value of feature d for patient p at time point $t\}$.*

However, EMR data only contains a subset of Φ, which contains records when patients visit the hospital with lab tests that are taken. We define the observed part of Φ from EMR as EMR Series Ψ.

Definition 2. *(EMR Series Ψ) Observed EMR data can be viewed as a multivariate time series for P^{EMR} patients $\Omega_P^{EMR} \subseteq \Omega_P$, where T^{EMR} regular time points $\Omega_T^{EMR} \subseteq \Omega_T$ correspond to patients' visits and variable dimension contains D^{EMR} distinct medical features $\Omega_D^{EMR} \subseteq \Omega_D$. This multivariate time series is a subset of Φ and defined as "EMR series" Ψ.*

Our work is inspired by the fact that Ψ is not a randomly sampled subset of Φ. Instead, the probability that one tuple $< p, t, d, v >$ is observed may depend on the medical feature d and its value v. Therefore, the target of our work is to estimate the unobserved hidden conditions $\Phi - \Psi$ using EMR series Ψ (illustrated in Figure 3), while considering that observations in Ψ may not be randomly sampled. This process is formally defined as EMR regularization.

Definition 3. *(EMR Regularization) EMR regularization is the process of predicting the values of the unobserved hidden conditions $\Phi - \Psi$. Given an EMR series Ψ as the input, suppose that for each tuple $< p, t, d, v >$ in Φ, its probability to be observed in Ψ is a function depending on the medical feature d and its value v. Since there is high uncertainty in the prediction, for each patient p, time point t, and medical feature d, we learn the distribution over all possible values v instead of generating one simple prediction.*

The EMR regularization is a general method designed to transform biased, irregular multivariate time series data into unbiased, regular ones, hence, can generate time series data of high quality and avoid misinterpretation. We will illustrate the generality of our method with relevant applications in Section 5.3.

The notations used in the remaining sections of this paper are summarized in Table 1.

Table 1: Notations

Notation	Description
P, Ω_P, p	Number of patients, patient set, each patient in Ω_P
T, Ω_T, t	Number of regular time points, regular time point set, each time point in Ω_T
D, Ω_D, d	Number of distinct medical features, medical feature set, each medical feature in Ω_D
Φ	Patients' hidden conditions
Ψ	EMR series
$y_t^{d,s}$	Observed value for feature d at time point t in an observation sequence s
$m_t^{d,s}$	Mask indicator for feature d at time point t in an observation sequence s
$Y^{d,s}$	An observation sequence (corresponding to one patient) for feature d composed of $y_t^{d,s}$
Ω_S^d	A set of observation sequence for feature d, $\Omega_S^d = \{Y^{d,s}\}$
$q_t^{d,s}$	Hidden state value for feature d at time point t in a hidden state sequence s
$Q^{d,s}$	A hidden state sequence for feature d composed of $q_t^{d,s}$
θ^d, θ	Condition change rate of feature d, prior of θ^d
ϕ^d, ϕ	Observation rate of feature d, prior of ϕ^d
$Beta(a, b)$	Beta function
N	Number of states in HMM
M	Number of observation symbols in HMM
Z	Hidden state set in HMM
V	Vocabulary set in HMM
Π^d	Initial hidden state distribution for feature d
A^d	Transition probability matrix for feature d
B^d	Emission probability matrix for feature d

4 EMR REGULARIZATION MODEL

In this section, we propose our EMR regularization model through the modelling and inference steps in detail.

4.1 Model Description

The graphical representation of our EMR regularization model is illustrated in Figure 4, which is based on the dynamic Bayesian networks. In the model structure, we show the interaction between all variables (either observed or hidden) in "Time Slide 1" in the left part and the transitioning relationship between consecutive time points (in this case, between "Time Slice t" and "Time Slice $t+1$") in the right part.

In the model structure, Ω_D denotes all distinct medical features and d represents each medical feature. $Y^d = y_1^d \cdots y_t^d \cdots$ in filled circles denote observed EMR series for feature d, where $Q^d = q_1^d \cdots q_t^d \cdots$ in open circles represents the hidden states, which are the distribution of feature d's possible values. As discussed in Section 3.2, the objective of EMR regularization is to infer Q^d based on Y^d for each feature d.

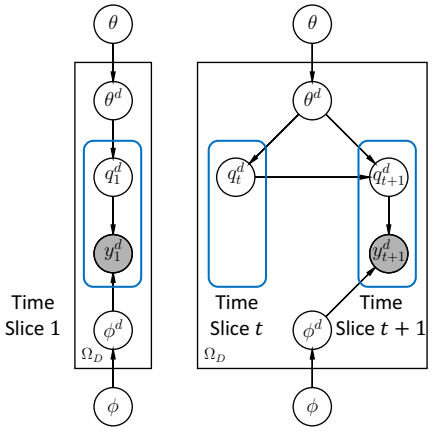

Figure 4: Graphical representation of proposed model.

We model two patterns in our EMR regularization to describe the characteristics of medical features.

- **Condition Change Rate (CCR)**

  CCR measures how a medical feature is likely to change from its condition in the previous observation. θ^d in Figure 4 represents CCR and θ on top is the prior of feature d's CCR. Different medical features have different CCRs. As in our representative examples in Section 3.1, AKI and glucose tend to have a higher CCR, whereas CKD Stage I and diabetes may have a lower CCR.

- **Observation Rate (OR)**

  OR measures the probability that a medical feature is exposed at a time point based on its actual condition at that time point. OR is denoted as ϕ^d in our graphical representation and ϕ on the bottom is the prior of feature d's OR. OR of AKI can be quite high during the active period as it will cause much pain to patients. However, OR of glucose, CKD Stage I and diabetes may be much less significant.

Assigning prior to CCR and OR for medical features, which can guide the model's learning, is of vital importance in the EMR regularization. We tend to assign CCR's prior such that there is a higher probability of being in the same state as in the previous one. Furthermore, we tend to assign OR's prior such that patients are more likely to visit the hospital when their condition indicates abnormal.

Specifically, for ease of computation, we assume that both CCR and OR follow the Beta distribution[3]

Assumption 1. *Given a medical feature d, its CCR's probability density function is:*

$$f(\theta^d = x) = x^{a_{ccr}-1}(1-x)^{b_{ccr}-1}/Beta(a_{ccr}, b_{ccr})$$

Similarly, feature d's OR's probability density function is:

$$f(\phi^d = x) = x^{a_{or}-1}(1-x)^{b_{or}-1}/Beta(a_{or}, b_{or})$$

[3]As will be discussed in Section 4.2 later, CCR and OR follow the Binomial distribution. Therefore, we choose the corresponding conjugate prior, the Beta distribution, for ease of computation.

In the rest of this section, we will discuss how to learn the parameters involved and infer the true hidden state sequences based on our proposed EMR regularization model.

4.2 Learning and Inference

We employ the Baum-Welch algorithm [19], which is an instantiation of the Expectation-Maximization (EM) algorithm [7] for HMM, to find the maximum a posterior estimated values of model parameters. The overall procedure is summarized in Algorithm 1.

Before we dive into details of Algorithm 1, we introduce some key concepts used in our HMM variant. HMM is proposed and used to model sequential observations based on the assumption that such observations are generated under a hidden stochastic process. In our case, the HMM is characterized by the following elements.[4]

- N denotes the number of states. In our case, $N = 2$ and the state set is represented as $Z = \{z_i\}$ which is $\{-1, +1\}$ where "-1" denotes abnormal state and "$+1$" denotes normal state.
- M denotes the number of distinct observation symbols. In our HMM model, $M = 3$ and the vocabulary set is $V = \{v_k\}$ which is $\{-1, 0, +1\}$ denoting abnormal, missing and normal observation respectively.
- $\Pi = \{\Pi_i\}$ is the initial state distribution and $\Pi_i = P(q_1 = z_i)$ corresponds to the probability of being in state z_i initially in the hidden state sequence.
- $A = \{A_{i,j}\}$ is the state transition probability distribution and $A_{i,j} = P(q_{t+1} = z_j | q_t = z_i)$ for $\forall 1 \le i, j \le N$. This A corresponds to the modelling of our CCR and represents how much the previous hidden state influences the current state.

  Based on $N = 2$, A (i.e., CCR) follows the Binomial distribution and A's prior follows the Beta distribution. [5]
- $B = \{B_{j,v_k}\}$ is the emission probability distribution and $B_{j,v_k} = P(y_t = v_k | q_t = z_j)$ for $\forall 1 \le j \le N$ and $\forall 1 \le k \le M$. B is the modelling of our OR, which denotes patient's probability of visiting the hospital based on the actual hidden state.

  The difference between our model and HMM is our "No Misdiagnosis" assumption. That is, given that feature d's actual state q_t is normal ($q_t = +1$), it can only be emitted to either normal or missing observation, i.e., $y_t = +1$ or $y_t = 0$. Analogously, given $q_t = -1$, it can only be emitted to $y_t = -1$ or $y_t = 0$.

  Based on the assumption, we model B to follow the Binomial distribution and B's prior to follow the Beta distribution. [6]

Next, we describe the procedure for learning and inference shown in Algorithm 1 in detail. As shown, the training input is composed of all distinct medical features Ω_D with all corresponding observation sequences (each sequence is $Y^{d,s}$), and the prior

[4]For simplicity, we omit the superscript of medical feature d and sequence s for q and y, and omit the superscript of d for Π, A and B in the notations here.
[5]For multi-state situation, we can replace the Binomial distribution and the Beta distribution to the more general Multinomial distribution and the Dirichlet distribution.
[6]Without the "No Misdiagnosis" assumption or with more complicated mapping between hidden states and observations, we can replace the Binomial distribution and the Beta distribution to the more general Multinomial distribution and the Dirichlet distribution accordingly.

Algorithm 1: EMR regularization with smoothing

Input: medical features Ω_D, observation sequences $\Omega_S^d = \{Y^{d,s} | Y^{d,s} = y_1^{d,s}, \cdots, y_T^{d,s}\}$ for each feature d and for each sequence s. A's prior for feature d is $Beta(a_A^d, b_A^d)$, B's prior for feature d is $Beta(a_B^d, b_B^d)$.

Output: parameters $\lambda^d = (\Pi^d, A^d, B^d)$ for each $d \in \Omega_D$, hidden state probability sequence $P(q_t^{d,s} = z_i | Y^{d,s}, \lambda^d)$.

1: For each medical feature $d \in \Omega_D$
2: Initialize $\lambda^d = (\Pi^d, A^d, B^d)$
3: Iterate EM process until convergence
4: **E-Step**:
5: For each observation sequence $s \in \Omega_S^d$
6: Compute $\xi_t(q_t^{d,s} = z_i, q_{t+1}^{d,s} = z_j)$ (Equation 3)
7: Compute $\gamma_t(q_t^{d,s} = z_j)$ (Equation 4)
8: **M-Step**:
9: Update $\hat{\Pi}_i^d$ (Equation 5)
10: Update transition matrix $\hat{A}_{i,j}^d$ (Equation 6)
11: Update emission matrix $\hat{B}_{j,v_k}^d$ (Equation 7)
12: Compute $P(q_t^{d,s} = z_i | Y^{d,s}, \lambda^d)$ (Equation 8)
13: **return** $\lambda^d = (\Pi^d, A^d, B^d)$, $P(q_t^{d,s} = z_i | Y^{d,s}, \lambda^d)$

parameters for A and B for each feature d. Then the model's output consists of learned parameters λ^d and hidden state probability sequence $P(q_t^{d,s} = z_i | Y^{d,s}, \lambda^d)$.

Then in Line 1-11, we iteratively compute the expectation of latent variables needed in the E-step (expectation step) and then update the parameter values in the M-step (maximization step) to maximize the joint log-likelihood of all observation sequences.

E-Step (Line 4-7 in Algorithm 1). For each feature d's each observation sequence $Y^{d,s}$ (corresponding to one patient), we need to compute the forward probability $\alpha_t(q_t^{d,s} = z_i)$ and the backward probability $\beta_t(q_t^{d,s} = z_j)$ as follows.

$$\alpha_t(q_t^{d,s} = z_i) = P(y_1^{d,s}, y_2^{d,s}, \cdots, y_t^{d,s}, q_t^{d,s} = z_i | \lambda^d) \tag{1}$$

denotes the probability of seeing partial observation sequence until t and staying in state z_i at time point t given λ^d.

$$\beta_t(q_t^{d,s} = z_j) = P(y_{t+1}^{d,s}, y_{t+2}^{d,s}, \cdots, y_T^{d,s} | q_t^{d,s} = z_j, \lambda^d) \tag{2}$$

represents the probability of seeing the partial observation after t given λ^d and being in state z_j at time point t.

Next, we compute the following two probability distributions described below:

$$\begin{aligned} \xi_t(q_t^{d,s} = z_i, q_{t+1}^{d,s} = z_j) &= P(q_t^{d,s} = z_i, q_{t+1}^{d,s} = z_j | Y^{d,s}, \lambda^d) \\ &= \frac{\alpha_t(q_t^{d,s} = z_i) A_{i,j}^d B_{j,y_{t+1}^{d,s}}^d \beta_{t+1}(q_{t+1}^{d,s} = z_j)}{\sum_{i=1}^N \sum_{j=1}^N \alpha_t(q_t^{d,s} = z_i) A_{i,j}^d B_{j,y_{t+1}^{d,s}}^d \beta_{t+1}(q_{t+1}^{d,s} = z_j)} \end{aligned} \tag{3}$$

for $\forall t, z_i$ and z_j, which is the probability of being in z_i at time point t and in z_j at time point $t+1$ given the observation and the model.

Then we can compute the following:

$$\begin{aligned} \gamma_t(q_t^{d,s} = z_j) &= P(q_t^{d,s} = z_j | Y^{d,s}, \lambda^d) \\ &= \frac{\alpha_t(q_t^{d,s} = z_j) \beta_t(q_t^{d,s} = z_j)}{\sum_{j=1}^N \alpha_t(q_t^{d,s} = z_j) \beta_t(q_t^{d,s} = z_j)} \end{aligned} \tag{4}$$

for $\forall t$ and z_j, which is the probability of being in z_j at time point t given the observation and the model.

M-Step (Line 8-11 in Algorithm 1). We need to update the parameters summarizing computed expectation values from all observation sequences in Ω_S^d, including the initial state distribution (Equation 5), the transition probability matrix (Equation 6) and the emission probability matrix (Equation 7). With A^d's prior $Beta(a_A^d, b_A^d)$, we assign a_A^d, b_A^d to corresponding entries in A^d and obtain $u_{i,j}^A$ for $\forall i, j \in Z$. Similarly, for B^d, using a_B^d, b_B^d, we get u_{j,v_k}^B for $\forall j \in Z$ and $\forall v_k \in V$ [1].

$$\hat{\Pi}_i^d = \frac{1}{|\Omega_S^d|} \sum_{s \in \Omega_S^d} \gamma_1(q_1^{d,s} = z_i) \tag{5}$$

$$\hat{A}_{i,j}^d = \frac{\sum_{s \in \Omega_S^d} \sum_{t=1}^{T-1} \xi_t(q_t^{d,s} = z_i, q_{t+1}^{d,s} = z_j) + u_{i,j}^A - 1}{\sum_{s \in \Omega_S^d} \sum_{t=1}^{T-1} \sum_{j=1}^N \xi_t(q_t^{d,s} = z_i, q_{t+1}^{d,s} = z_j) + \sum_j (u_{i,j}^A - 1)} \tag{6}$$

$$\hat{B}_{j,v_k}^d = \frac{\sum_{s \in \Omega_S^d} \sum_{t=1 s.t. y_t^{d,s} = v_k}^{T} \gamma_t(q_t^{d,s} = z_j) + u_{j,v_k}^B - 1}{\sum_{s \in \Omega_S^d} \sum_{t=1}^{T} \gamma_t(q_t^{d,s} = z_j) + \sum_{v_k} (u_{j,v_k}^B - 1)} \tag{7}$$

Inference. After the EM algorithm has converged, we can infer the hidden state distribution. We first compute the forward probability and the backward probability using learned parameters after convergence, then based on the Bayesian rule, the hidden state distribution is as follows:

$$P(q_t^{d,s} = z_i | Y^{d,s}, \lambda^d) = \frac{\alpha_t(q_t^{d,s} = z_i) \beta_t(q_t^{d,s} = z_i)}{P(Y^{d,s} | \lambda^d)} \tag{8}$$

The expectation of this $P(q_t^{d,s} = z_i | Y^{d,s}, \lambda^d)$ is the desired output of our EMR regularization model.

5 EXPERIMENTS

We design two experiments to evaluate the effectiveness of our proposed EMR regularization model: (i) imputation accuracy evaluation, which is to randomly hide positions in the time series, and then to compare the imputation results with the ground truth values; (ii) benefits for analytical tasks, which is to apply data imputed by different methods into further analytical tasks and to compare the corresponding analytical results. We report the results of these two experiments in Section 5.2 and Section 5.3 respectively.

5.1 Experimental Set-Up

5.1.1 Datasets, Applications and Evaluation Metrics. We conduct experiments using two real-world EMR datasets to evaluate the performance of our proposed method. We perform different analytical tasks in the datasets described as follows, with corresponding evaluation metrics.

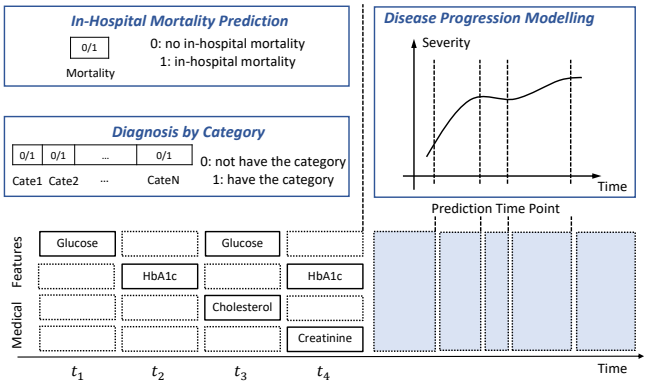

Figure 5: Analytical applications including in-hospital mortality prediction, diagnosis by category, and disease progression modelling. Medical features for illustration include glucose, HbA1c (Hemoglobin A1c), cholesterol and creatinine.

Table 2: Dataset Statistics

Dataset	# of Features	# of Samples	Missing Rate
MIMIC-III	53	46776	92.87%
NUH-CKD	332	18344	99.03%

MIMIC-III Dataset is a public EMR database [12] consisting of records of over 40,000 ICU patients who are admitted to the hospital between 2001 and 2012. This dataset includes patients' lab tests, medications, procedures, demographics, and vital sign measurements, etc.

In this dataset, we use each admission as a sample, which refers to one of a patient's visits to the hospital. For each admission, there is an in-hospital mortality label denoting whether the patient dies in the hospital. Furthermore, there are several ICD-9 (Ninth Revision of the International Classification of Diseases) diagnosis codes assigned by doctors denoting which diseases the patient has. We extract the admissions with a time span longer than 48 hours as samples and perform two prediction tasks: (i) **In-hospital mortality prediction** is to predict whether the patient will die in the hospital in this admission and this prediction task is modelled as a binary classification problem; (ii) **Diagnosis by category** is to predict the disease categories the patient has. We categorize available ICD-9 diagnosis codes into 20 categories by separating E codes and V codes[7]. After removing one category which does not appear in our data, we formalize this task as a 19-label classification problem.

For evaluation, we report the AUROC value (area under the ROC curve) for classification. We report AUROC for in-hospital mortality prediction and the average AUROC value of the 19 labels for diagnosis by category.

NUH Dataset is a real-world longitudinal EMR dataset from National University Hospital in Singapore. We conduct experiments in a sample dataset of the NUH dataset with over 100,000 patients admitted in NUH from 2011 to 2012. Patients' medical features such as diagnoses, lab tests, medications and procedures are collected in this dataset.

In this dataset, we perform **disease progression modelling (DPM)** task, which is to predict the future severity of patients suffering from a target disease. We choose CKD patients who are in Stage 3 or latter stages[8] as our cohort and denote this dataset as the **NUH-CKD** dataset. We predict the patients' future GFR values and formalize the DPM task as a regression problem. GFR[9] is a lab test short for Glomerular Filtration Rate, which measures the severity of CKD patients. The lower the GFR value is, the more severe the patient is. After consulting medical experts, according to *The Renal Association*[10], to examine the loss of GFR value over one year is of vital importance. To be specific, if a patient suffers from a loss of $5ml/min/1.73m^2$ in GFR value in less than one year, this patient is having a deteriorating state and needs assessment from medical specialists. As a consequence, our DPM task is medically meaningful in the NUH-CKD dataset.

We use MSE (mean squared error) and R value (Pearson product-moment correlation coefficient) as our evaluation metrics. R value with range $[-1, 1]$ measures the linear relationship between the predicted value and the true value, where $R > 0$ denotes a positive relationship and $R < 0$ represents a negative relationship. A good regression model will have a small MSE value and a large R value.

We illustrate the three analytical tasks described above in Figure 5. In the bottom, the left part denotes the longitudinal EMR data we can make use of, and the right part denotes data unavailable in the future. As shown in Figure 5, the in-hospital mortality prediction task and the diagnosis by category task in the MIMIC-III dataset target at classifying admission sequences and the DPM task in the NUH-CKD dataset aims at predicting the future severity trajectory of CKD patients.

We show some statistics information including number of features, number of samples and missing rate[11] about the MIMIC-III dataset and the NUH-CKD dataset in Table 2. Compared with the MIMIC-III dataset, the NUH-CKD dataset has more features but fewer samples. Moreover, NUH-CKD dataset has a larger missing rate.

5.1.2 Baseline Methods and Implementation Details. We compare our proposed method with several basic methods and advanced methods. The corresponding details are as follows.

We employ the following methods to impute the missing data in the time series sequence:

- **Forw:** use the nearest previous value for the missing time point, called forward imputation
- **Mean:** use the mean value for the missing time point, called mean imputation
- **Zero:** use zero for the missing time point, called zero imputation

Though basic, these three methods for handling missing data are widely applied in the area of EMR data analytics[3, 5, 14].

[7] https://en.wikipedia.org/wiki/List_of_ICD-9_codes

[8] http://www.renal.org/information-resources/the-uk-eckd-guide/ckd-stages

[9] https://labtestsonline.org/understanding/analytes/gfr/tab/test

[10] http://www.renal.org/information-resources/the-uk-eckd-guide/deteriorating-function

[11] In this table, "missing rate" is computed for lab tests and is defined as the ratio of non-zero entries over all entries, where the total number of entries is (# of features) × (# of time points) × (# of samples).

Recently, an advanced model [15] proposes to employ a mask indicator to denote whether a position has value or not, to improve the performance of EMR data analytics. The mask indicator is $m_t^{d,s}$, where $m_t^{d,s} = 1$ indicates the value for feature d in sequence s at time point t is missing, and $m_t^{d,s} = 0$ indicates not missing. Hence, this method can be considered as adding mask indicators to original input features based on the Forw method. Through adding such mask indicators, the information loss of the original data caused by imputation can be reduced, and Lipton et al. show improved predictive performance in their results [15].

Hence, we add the following methods as our baselines to compare with [15] as well.

- **Forw-Mask:** further add mask indicators based on **Forw**
- **Mean-Mask:** further add mask indicators based on **Mean**
- **Zero-Mask:** further add mask indicators based on **Zero**

For the MIMIC-III dataset, we make use of the LABEVENTS modality, which records patients' lab test results. We extract the first 48 hours of each admission sequence as input and then divide the sequence into two-hour time windows for utilization. For both applications, in the deep learning model structure, we input the time series features through an input layer, then connect it to an RNN (GRU, LSTM or Vanilla RNN) layer, and next to a dense layer followed by the activation layer serving as the output layer. For in-hospital mortality prediction, we use one *sigmoid* unit as the output layer in the model. For diagnosis by category, we use one *sigmoid* unit per diagnosis category, ending up with a 19-dimension output layer. We aim to minimize the cross-entropy loss during the training of the constructed deep learning model.

For the NUH-CKD dataset, we extract each patient's lab tests in the first 200 days and aggregate them by week for utilization. We feed the extracted time series data into an RNN layer and then a dense layer, followed by a ReLU (rectified linear units) activation layer. Furthermore, we select the GFR lab tests after each patient's 200-day time point as our prediction target, incorporate the time span between the prediction and the utilized time series data as an input feature into our deep learning model and use a dense layer with one unit to predict the target GFR values as a regression problem.

For all application experiments, we randomly partition the whole dataset into 80%, 10%, 10% representing training data, validation data, and testing data respectively. We train the deep learning model and keep the hyper-parameters which achieve the best performance in the validation data, and then apply the trained model in the testing data for reporting the experimental results.

5.2 Imputation Accuracy Evaluation

We conduct experiments in both the MIMIC-III dataset and the NUH-CKD dataset to show that our proposed method is more accurate in filling in missing data than baseline methods including Forw, Mean and Zero[12].

In this experiment, we cannot evaluate the performance of different methods directly as we do not have access to the ground truth, i.e., EMR data at certain time points when the data is missing, meaning when the patients do not visit the hospital. As a consequence, we perform **random hiding** for evaluating the imputation accuracy, which is to randomly hide one position's value in a time sequence and then apply different methods to fill in this position. Then through comparing the difference between the imputed value and the original value which is hidden, we can evaluate different methods' imputation accuracy.

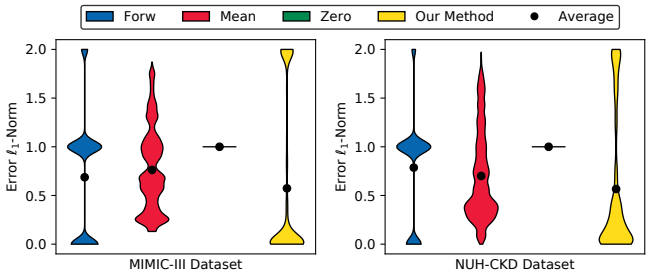

Figure 6: Imputation accuracy evaluation results.

After hiding a position with value in a sequence randomly, we employ our method and baseline methods on randomly selected 10% sequences and then compare the imputed results. We evaluate different methods' performance by calculating the ℓ_1-norm of the difference between the prediction result and the ground truth value. The ℓ_1-norm of the difference ($\in [0, 2]$) provides a direct comparison for the imputation accuracy of different methods. The smaller this calculated ℓ_1-norm is, the more accurate the method is.

We illustrate the imputation results of different methods in both the MIMIC-III dataset and the NUH-CKD dataset respectively in Figure 6. This figure is a violin plot[13] with the ability to show data's probability density in different values. For each dataset in Figure 6, we show all testing sequences' Error ℓ_1-norm, which is computed as the difference (in terms of ℓ_1-norm) between the filled value and the corresponding ground truth value. The results of four methods are shown from left to right, namely Forw, Mean, Zero and our method. The filled dot represents the average of all testing sequences' Error ℓ_1-norm for each method (i.e., average Error ℓ_1-norm for short). In the MIMIC-III dataset, the average Error ℓ_1-norm for four methods is 0.6872, 0.7625, 1.0 and 0.5732. In the NUH-CKD dataset, the average value is 0.7862, 0.7009, 1.0 and 0.5664 respectively.

As illustrated in Figure 6, the Error ℓ_1-norm of the Forw method is 1.0 in many sequences. This might be because there exists few entries with values in each sequence. Hence, there is a high probability that there are no previous values before the hidden position for the Forw method and in this case, the Error ℓ_1-norm for this hidden position is 1.0. The Error ℓ_1-norm of the Mean method is computed using the difference between the mean value and the filled value in this hidden position. The shape of the Mean method results shows the deviation of the original values in the hidden positions from the mean value. For the Zero method, the Error ℓ_1-norm is always 1.0 for all sequences. As shown, our proposed EMR regularization method fills most hidden positions with a smaller Error ℓ_1-norm and therefore, imputes the hidden positions more accurately than other methods.

[12] We do not compare with **Forw-Mask**, **Mean-Mask** and **Zero-Mask** in this experiment because these methods based on mask indicators do not impute missing data directly. We compare with these methods for specific analytical tasks in Section 5.3.

[13] https://en.wikipedia.org/wiki/Violin_plot

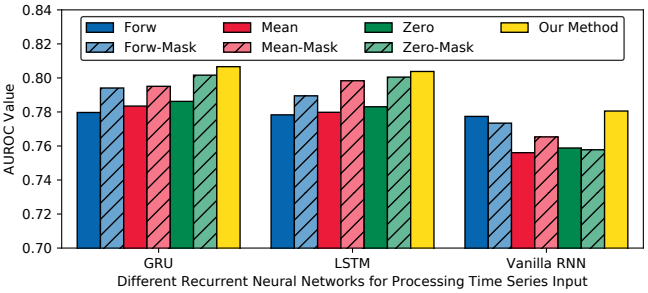

Figure 7: MIMIC-III in-hospital mortality prediction results.

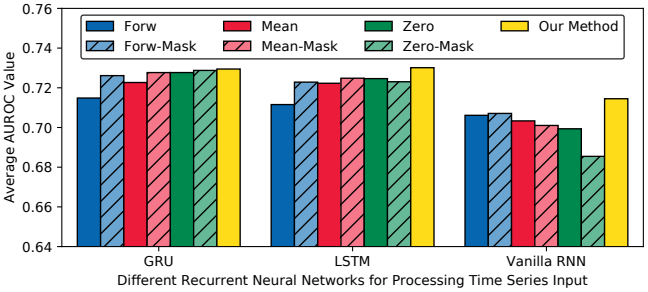

Figure 8: MIMIC-III diagnosis by category results.

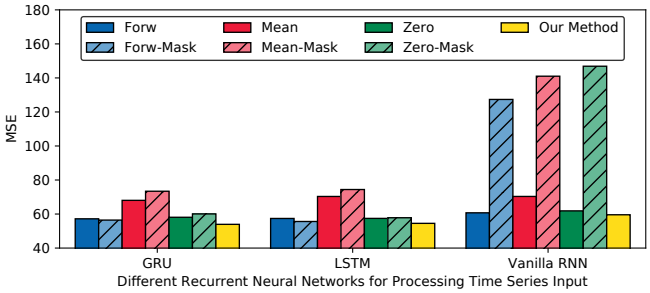

Figure 9: MSE for NUH-CKD disease progression modelling.

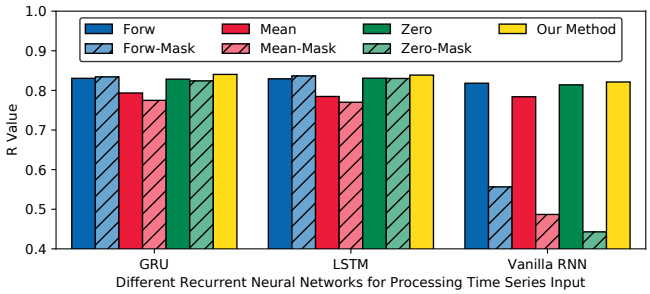

Figure 10: R value for NUH-CKD disease progression modelling.

5.3 Benefits for Analytical Tasks

5.3.1 MIMIC-III In-Hospital Mortality Prediction. We compare our proposed method with all six baseline methods in predicting in-hospital mortality in the MIMIC-III dataset illustrated in Figure 7.

In this experiment, we change three different commonly employed RNN models (GRU, LSTM and Vanilla RNN) for processing input time series data, and we compare our proposed method with three basic baseline methods as well as three advanced methods incorporating mask indicators.

For each RNN model, methods incorporating mask indicators (mask-based methods) outperform corresponding basic baselines in most cases (for instance, Forw-Mask is more accurate than Forw). This observation indicates that analytical tasks can benefit from incorporating mask indicators which reduce the information loss caused by specific imputation methods. GRU and LSTM can achieve competitive performance in the AUROC value of this task and both models outperform Vanilla RNN. For all three RNN models, our proposed EMR regularization method achieves the highest AUROC value in testing data among all methods, indicating that our consideration of CCR and OR helps resolve the bias in EMR data and contributes to more accurate in-hospital mortality prediction.

5.3.2 MIMIC-III Diagnosis by Category. In this experiment, we employ the three RNN models as in Section 5.3.1, but target at the diagnosis by category task in the MIMIC-III dataset. We report the average AUROC value for 19 diagnosis categories in Figure 8.

The insight we can obtain from this experiment are similar to that from the in-hospital mortality prediction task. First, mask-based methods perform better in the average AUROC value than corresponding basic methods in most cases. This means mask indicators provide more information beneficial to the task. Second, the average AUROC values achieved by GRU and LSTM are similar, but are higher than that achieved by Vanilla RNN. Third, within all three RNN models, our proposed EMR regularization method outperforms other methods in terms of the average AUROC value. This indicates our method manages to help resolve the bias and hence is beneficial to the diagnosis by category task.

From these two analytical tasks in the MIMIC-III dataset, one interesting observation is that the most competitive baseline method of our proposed method is Zero-Mask when employing GRU and LSTM. This is a bit surprising, as intuitively we believe that Zero imputation (or Zero-Mask) may not convey as much information as Forw (or Forw-Mask) and Mean (or Mean-Mask). However, the results may indicate that using EMR data recorded explicitly is better than imputing missing data without considering the bias, even if mask indicators are added as auxiliary information. The advantage of Zero-Mask method shown in our experiments agrees with [15].

5.3.3 NUH-CKD Disease Progression Modelling. As described in Section 5.1, DPM in the NUH-CKD dataset is modelled as a regression problem and MSE, R value are used as evaluation metrics. A better regression model gives a smaller MSE value and a larger R value. We illustrate the comparison results between our proposed method and all baseline methods in Figure 9 and Figure 10 for MSE and R value respectively.

For advanced RNN models (GRU and LSTM), we observe that mask-based methods achieve similar performance (in terms of both MSE and R value) to basic methods without much superiority. This phenomenon is different from the results in the MIMIC-III dataset which show the advantages of mask-based methods over basic ones. The reason may be that the MIMIC-III dataset is an ICU EMR dataset which collects patients' more dense information. However, the NUH-CKD dataset is a general EMR dataset which has a more severe issue of missing data (as shown in Table 2). As a consequence,

the superiority of mask-based methods in providing useful information of missingness may be weakened in the NUH-CKD dataset. For Vanilla RNN model which might suffer from gradient exploding or vanishing when processing long-term dependencies [10], the performance tends to degrade. Moreover, compared with the MIMIC-III dataset, the NUH-CKD dataset has a larger number of features but a smaller number of samples. Vanilla RNN is likely to overfit without enough training samples and when adding mask indicators, the feature dimension further increases, mask-based methods are therefore badly influenced by the overfitting problem in Vanilla RNN.

Our proposed EMR regularization method achieves the smallest MSE and the largest R value among all methods when employing three RNN models. This further demonstrates that our bias resolving method provides benefits for analytical tasks.

6 CONCLUSION AND FUTURE WORK

In this paper, we identify and formalize the bias in EMR data, which may degrade the analytical performance if not carefully handled. Then we propose a general method to resolve the bias through transforming biased, irregularly sampled EMR time series into an unbiased, regular one and define this process as EMR regularization. Specifically, we employ an HMM variant as the inference model to conduct the transformation by considering the two characteristics of medical features, i.e., condition change rate and observation rate.

We conduct two experiments to evaluate the proposed EMR regularization method. In the imputation accuracy evaluation, we use random hiding to select the positions for imputation and the experimental results show our proposed method outperforms baselines. This demonstrates that our bias resolving approach imputes missing values more accurately. When checking our method's benefits for analytical tasks, we use our method as a pre-processing step for transforming EMR data and then feed the data into further analytical tasks. We compare with three basic methods and their corresponding advanced methods incorporating mask indicators. The experimental results demonstrate that our method achieves the highest accuracy and hence, our method can improve the analytical performance by resolving the bias.

This work can be further extended towards multiple directions. For instance, instead of modelling different diseases independently, how to take into account modelling them jointly in the probabilistic graphical model for capturing the relationship between diseases is interesting and medically meaningful. Furthermore, different patients might behave differently in terms of condition change rate and observation rate. How to model the patient personalization in the model is also well worth exploring.

ACKNOWLEDGMENTS

This work is supported by National Research Foundation, Prime Ministers Office, Singapore under its Competitive Research Programme (CRP Award No. NRF-CRP8-2011-08). We would like to thank Professor H. V. Jagadish and Assistant Professor Wei Wang for the discussion and useful advice that help to improve this work. We would also like to thank the anonymous referees for their valuable comments and helpful suggestions.

REFERENCES

[1] Matthew James Beal. 2003. *Variational algorithms for approximate Bayesian inference*. University of London United Kingdom.
[2] Rich Caruana. 1998. Multitask learning. In *Learning to learn*. Springer, 95–133.
[3] Zhengping Che, David Kale, Wenzhe Li, Mohammad Taha Bahadori, and Yan Liu. 2015. Deep Computational Phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 507–516.
[4] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2016. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *arXiv preprint arXiv:1606.01865* (2016).
[5] Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. 2015. Distilling Knowledge from Deep Networks with Applications to Healthcare Domain. *arXiv preprint arXiv:1512.03542* (2015).
[6] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).
[7] Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)* (1977), 1–38.
[8] Simon Duchesne, Anna Caroli, Cristina Geroldi, D Louis Collins, and Giovanni B Frisoni. 2009. Relating one-year cognitive change in mild cognitive impairment to baseline MRI features. *NeuroImage* 47, 4 (2009), 1363–1370.
[9] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.
[10] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[11] George Hripcsak, David J Albers, and Adler Perotte. 2015. Parameterizing time in electronic health record studies. *Journal of the American Medical Informatics Association* (2015), ocu051.
[12] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3 (2016).
[13] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.
[14] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzell. 2016. Learning to diagnose with LSTM recurrent neural networks. *International Conference on Learning Representations* (2016).
[15] Zachary C Lipton, David C Kale, and Randall Wetzel. 2016. Modeling Missing Data in Clinical Time Series with RNNs. In *Machine Learning for Healthcare*.
[16] Liqiang Nie, Luming Zhang, Yi Yang, Meng Wang, Richang Hong, and Tat-Seng Chua. 2015. Beyond Doctors: Future Health Prediction from Multimedia and Multimodal Observations. In *Proceedings of the 23rd ACM International Conference on Multimedia (MM '15)*. ACM, New York, NY, USA, 591–600.
[17] Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. 2016. DeepCare: A Deep Dynamic Memory Model for Predictive Medicine. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 30–41.
[18] Rimma Pivovarov, David J Albers, Jorge L Sepulveda, and Noémie Elhadad. 2014. Identifying and mitigating biases in EHR laboratory tests. *Journal of biomedical informatics* 51 (2014), 24–34.
[19] Lawrence R Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77, 2 (1989), 257–286.
[20] Matthias B Schulze, Kurt Hoffmann, Heiner Boeing, Jakob Linseisen, Sabine Rohrmann, Matthias Möhlig, Andreas FH Pfeiffer, Joachim Spranger, Claus Thamer, Hans-Ulrich Häring, et al. 2007. An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes. *Diabetes care* 30, 3 (2007), 510–515.
[21] Cynthia M Stonnington, Carlton Chu, Stefan Klöppel, Clifford R Jack, John Ashburner, Richard SJ Frackowiak, Alzheimer Disease Neuroimaging Initiative, et al. 2010. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. *Neuroimage* 51, 4 (2010), 1405–1413.
[22] Xiang Wang, David Sontag, and Fei Wang. 2014. Unsupervised Learning of Disease Progression Models. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. ACM, New York, NY, USA, 85–94.
[23] Jiayu Zhou, Jun Liu, Vaibhav A. Narayan, and Jieping Ye. 2012. Modeling Disease Progression via Fused Sparse Group Lasso. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12)*. ACM, New York, NY, USA, 1095–1103.
[24] Jiayu Zhou, Lei Yuan, Jun Liu, and Jieping Ye. 2011. A Multi-task Learning Formulation for Predicting Disease Progression. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*. ACM, New York, NY, USA, 814–822.