Learning Phenotypes and Dynamic Patient Representations via RNN Regularized Collective Non-Negative Tensor Factorization

Kejing Yin,[1] Dong Qian,[1] William K. Cheung,[1] Benjamin C. M. Fung,[2] Jonathan Poon[3]
[1]Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China
[2]School of Information Studies, McGill University, Montreal, Canada
[3]Hong Kong Hospital Authority, Hong Kong SAR, China
{cskjyin, dongqian, william}@comp.hkbu.edu.hk, ben.fung@mcgill.ca, jonathan@ha.org.hk

Abstract

Non-negative Tensor Factorization (NTF) has been shown effective to discover clinically relevant and interpretable phenotypes from Electronic Health Records (EHR). Existing NTF based computational phenotyping models aggregate data over the observation window, resulting in the learned phenotypes being mixtures of disease states appearing at different times. We argue that by separating the clinical events happening at different times in the input tensor, the temporal dynamics and the disease progression within the observation window could be modeled and the learned phenotypes will correspond to more specific disease states. Yet how to construct the tensor for data samples with different temporal lengths and properly capture the temporal relationship specific to each individual data sample remains an open challenge. In this paper, we propose a novel Collective Non-negative Tensor Factorization (CNTF) model where each patient is represented by a temporal tensor, and all of the temporal tensors are factorized collectively with the phenotype definitions being shared across all patients. The proposed CNTF model is also flexible to incorporate non-temporal data modality and RNN-based temporal regularization. We validate the proposed model using MIMIC-III dataset, and the empirical results show that the learned phenotypes are clinically interpretable. Moreover, the proposed CNTF model outperforms the state-of-the-art computational phenotyping models for the mortality prediction task.

Introduction

With the global adoption of Electronic Health Records (EHR) over the past decade, a large amount of clinical data about patients, including diagnoses, laboratory test results, medication prescriptions, *etc.*, were accumulated, providing great opportunities to accelerate clinical research and improve healthcare quality by strategic use of the EHR data (Yadav et al. 2018). However, using the raw EHR data is very challenging due to the inherently complex nature of healthcare and the data recording process, which is reflected by the fact that EHR data are often largely missing, frequently inaccurate and possibly biased (Hripcsak and Albers 2013), making the *true disease states* of patients not directly observable from the data. Therefore, the raw EHR data are often mapped to some clinically relevant and interpretable concepts, or *phenotypes*, that reveal the latent true disease states of patients (Kirby et al. 2016). With the aim of extracting phenotypes without intensive human supervision to scale well in large-scale datasets, a large number of machine learning based *computational phenotyping* models have been proposed (Hripcsak and Albers 2013), among which the Non-negative Tensor Factorization (NTF) has shown effective for this task with its capability of preserving and modeling the high-dimensional interactions (Henderson et al. 2017; Kim et al. 2017; Yin et al. 2018). Given the EHR dataset, the tensor representing the interactions among different data modalities, *e.g.*, lab tests and medications, can be defined and the interpretable phenotypes then can be discovered by factorizing the tensor.

Despite the advances in computational phenotyping using the NTF models, there are still fundamentally challenging issues to be solved. One of the important ones is that the temporal progression of patients is not well considered in general. Most of the NTF based models integrate the data over the observation window to build the input tensor (Ho et al. 2014), making the clinical events happening at different times being mixed in the input tensor. Consequently, the resulting phenotypes would also be mixtures of disease states that appearing at different times in the patient journey, instead of describing one distinct and specific disease state. It will be particularly undesirable for circumstances where disease states are evolving. In Intensive Care Units (ICU), for example, patients rapidly progress from one disease state to another. Given the data accumulated over the ICU stay, it would be extremely difficult to recover the disease states which are in fact appearing at different time points. As an empirical evidence, it has been observed in (Yin et al. 2018) that the disease "other disease of lung" appears in almost every phenotype. This is very likely due to the fact that many patients in ICU would finally develop to acute respiratory failure and thus dominates the learned phenotypes. Therefore, we believe that separating the events happening at different time in the input tensor would yield more distinct and specific disease states. However, modeling the temporal dynamic using the NTF model is not straightforward. Firstly, the length of the observation window for different patients may differ from each other, making it difficult to align the patient records. Although different heuristics can be used,

for example, downsampling or zero-padding, loss of information and introduction of bias would be resulted at the same time. Secondly, by simply adding time as a dimension, the global temporal relationship would be captured as a part of the phenotype definition, making the phenotypes difficult to be interpreted. We further elaborate this point after presenting the necessary preliminaries and model formulation in the proposed model section.

In this paper, we propose a novel Collective Non-negative Tensor Factorization (CNTF) model to tackle the aforementioned challenges by representing the EHR data using a collection of temporal tensors with different temporal lengths, instead of using one single tensor for all patients. Each of the temporal tensors corresponds to one individual patient, and the sizes of all dimensions other than time dimension are consistent for all patients. An additional static tensor model is also incorporated to allow the integration of non-temporal data modalities. An RNN-based regularization is further introduced to model the temporal dependency of the evolution of patients' disease states. We evaluate the proposed model on the MIMIC-III dataset (Johnson et al. 2016). The empirical results show that the disease states appearing at different times throughout the patient journey can be separated, which cannot be easily done by the existing models. The learned phenotypes also demonstrate better predictive power at the early stage of the hospital stay when compared with the baselines. To the best of our knowledge, this is the first work on computational phenotyping from varying-length temporal EHR data with the modality interactions being preserved.

Related Work

Non-negative tensor factorization (NTF) has been intensively studied, and great efforts have been made to apply NTF models to the computational phenotyping task with different data distribution assumptions and additional constraints. Ho et al. (2014) proposed an NTF-based computational phenotyping model. It was then extended by adding a bias tensor to infer the population-wise baseline characteristics (Ho, Ghosh, and Sun 2014), and by incorporating pairwise constraints (Wang et al. 2015) and similarity constraints (Henderson et al. 2017) for promoting the diversity of the learned phenotypes. In addition, other information including domain knowledge (Wang et al. 2015), clustering structure (Kim et al. 2017), label information (Yang et al. 2017), and the diagnosis-medication correspondence (Yin et al. 2018) was taken into consideration and incorporated into the NTF framework.

The aforementioned studies accumulate the clinical events over the observation period to construct the input tensor, without modeling the disease progression within the observation window. In fact, modeling the temporal relationship based on matrix factorization or tensor factorization has attracted increasing attention. Xiong et al. (2010) proposed a temporal collaborative filtering method based on the Bayesian probabilistic tensor factorization framework, where the time factor is assumed to be dependent on their immediate predecessor to capture the smooth global evolution trend. Similarly, an auto-regressive temporal regularization (Yu, Rao, and Dhillon 2016) was incorporated into the matrix factorization model to learn the temporal dependency for better prediction.

However, these models assume that all the data items are of the same temporal length and can be naturally aligned, which unfortunately is not applicable in the computational phenotyping context where the length of patient records varies significantly and cannot be aligned naturally due to the extremely diverse possibilities for disease state progression. The work most related to ours is the SPARTan model proposed in (Perros et al. 2017), where the time dimension is taken into account by forming an irregular tensor with the phenotypes being inferred by the PARAFAC2 decomposition. While targeting the same problem, our proposed model is essentially different and has several advantages over SPARTan. First, SPARTan constructs a matrix (*i.e.* a slice of a tensor) for each patient with items from different modalities concatenated to one axis, while we construct a tensor for each patient with the interactions among modalities being preserved. Second, SPARTan only imposes non-negativity constraints on the phenotype definitions, leaving the patient representations possibly being negative. This hurts the interpretability of the model as a patient will then be represented by some phenotypes which can "cancel" out each other. In our proposed model, both the phenotype definitions and the patient representations are constrained to be non-negative, resulting the "parts of the object" being captured by the phenotypes (Lee and Seung 1999). Third, incorporating side information, *e.g.*, modality without time information, to the SPARTan model is not straightforward, but our proposed model offers the flexibility to integrate such information. Fourth, SPARTan does not show how to capture the temporal dependency of the disease state progression, while we introduce the RNN-based regularization to model the temporal dependency.

Recurrent Neural Network (RNN) has been shown powerful in modeling sequential data and time series. Recently, various studies applied RNN model to analyze multivariable clinical time series (Che et al. 2018) and clinical event sequences (Choi et al. 2016a; 2016b), where its effectiveness has been repeatedly validated. As summarized in (Purushotham et al. 2018), deep learning models, including RNN, outperform all other models consistently for predicting various targets, *e.g.* mortality, length-of-stay, *etc.*, especially when the raw clinical time series is utilized.

Notations and Preliminaries

In this paper, we denote tensors by calligraphic letters ($\mathcal{X}$), matrices by capital letters ($\mathbf{X}$), vectors by boldface lowercase letters ($\mathbf{x}$) and scalars by lowercase letters (x). We use the superscripts with parentheses to index the elements in a collection. For instance, $\mathcal{X}^{(p)}$ denotes the p^{th} tensor from a collection of tensors $\{\mathcal{X}^{(1)}, \ldots, \mathcal{X}^{(N)}\}$.

CP Factorization. The CP factorization (Kolda and Bader 2009) of a tensor approximates the K^{th} order target tensor with the sum of component rank-one tensors, where a rank-one tensor is defined as the outer product of K vectors. Each of the component rank-one tensors is interpreted as one la-

tent factor. For example, the CP factorization of a 3^{rd} order tensor is defined as follows:

$$\mathcal{X} \approx \sum_{r=1}^{R} \mathbf{u}_r^{(1)} \circ \mathbf{u}_r^{(2)} \circ \mathbf{u}_r^{(3)} = [\![\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \mathbf{U}^{(3)}]\!], \quad (1)$$

where R is the number of rank-one tensors.

Poisson Non-negative CP Factorization. To enhance the interpretability of the CP factorization model, the Poisson non-negative CP factorization (Chi and Kolda 2012) further assumes that the input tensor follows a Poisson distribution parameterized by the reconstruction from its CP factors, and the non-negativity constraint is imposed on the factor matrices, resulting the following optimization problem:

$$\begin{aligned} \underset{\mathbf{U}^{(n)}}{\arg\min} \quad & f(\mathcal{M}) \equiv \sum_{\mathbf{i}} m_{\mathbf{i}} - x_{\mathbf{i}} \log m_{\mathbf{i}} \\ \text{subject to} \quad & \mathcal{M} = [\![\lambda; \mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(K)}]\!] \\ & \mathbf{U}^{(k)} \geq \mathbf{0}, \text{ for } k = 1, \ldots, K, \\ & \|\mathbf{u}_r^{(k)}\|_1 = 1 \ \ \forall r \ \forall k. \end{aligned} \quad (2)$$

Proposed Model

In this section, we describe the framework of our proposed model to jointly learn the static phenotype definitions and the patient-specific dynamic representation. We start from the collective non-negative tensor factorization (CNTF), which models each patient with a temporal tensor to avoid aligning patients with different temporal length. Then we demonstrate the flexibility of the proposed basic model where non-temporal data modalities without time stamps can also be readily incorporated. Finally, we introduce an RNN-based regularization to better model the temporal relationship. The overview of the proposed framework is illustrated in Fig. 1.

Collective Non-Negative Tensor Factorization

Given a collection of patient records with K modalites that are recorded with time stamps, we aim to simultaneously discover the static phenotype definitions describing the true disease states and the dynamic representation of the patients revealing the dynamic changes of the disease states of the patients throughout the observation window. The length of the observation window for each individual patient may differ from each other. For instance, if the observation window is a hospital stay, it is not feasible to construct a single tensor for all patients as most of the existing models do due to the inconsistency of the time dimension. Instead, we construct a $(K+1)^{th}$ order interaction tensor for each patient, resulting a collection of N_p temporal tensors, *i.e.* $\mathbb{X} = \{\mathcal{X}^{(p)} | \mathcal{X}^{(p)} \in \mathbb{R}^{T_p \times I_1 \times \cdots \times I_K}, \text{ for } p = 1, \ldots, N_p\}$, where N_p is the number of patients, $\mathcal{X}^{(p)}$ is the temporal tensor for the p^{th} patient, T_p is the temporal length of the p^{th} patient's records and I_k is the size of the k^{th} dimension. Without loss of generality, we assume $K = 2$ with the dimensions being *lab tests* and *medications* respectively for simplifying the notations. As presented in the previous section, the non-negative

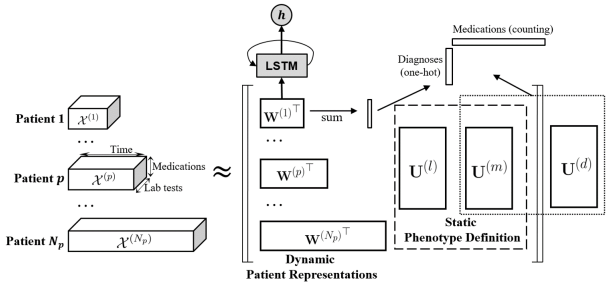

Figure 1: The framework of the proposed model. A 3^{rd} order *time-labtest-medication* tensor is constructed for each patient, and all of the temporal tensors are factorized with the phenotype definitions being shared across all the patients. RNN-based regularization is introduced to model the time dependency of the dynamic patient representations. Another tensor model called HITF (Yin et al. 2018) is also incorporated to allow non-temporal modalities to be utilized.

CP factorization of the 3^{rd} order temporal tensor yields three latent factor matrices $\mathbf{W}^{(p)} \in \mathbb{R}^{T_p \times R}$, $\mathbf{U}^{(l)} \in \mathbb{R}^{I_l \times R}$ and $\mathbf{U}^{(m)} \in \mathbb{R}^{I_m \times R}$, where R is the number of phenotypes, I_l is the number of lab tests and I_m is the number of medications. We refer to the latter two factor matrices as the phenotype definitions. In order to discover phenotypes that account for all patients rather than a single patient, we introduce the hard constraint that $\mathbf{U}^{(k)}$ has to be shared across all patients for all k. The first factor matrix $\mathbf{W}^{(p)}$ is referred to as dynamic patient representation because its entry $w_{tr}^{(p)}$ describes how likely the r^{th} phenotype exists at the particular time point t. We may understand this, intuitively, as learning a "dictionary" that describes some potentially clinically meaningful disease states, and concurrently selecting different non-negative combinations of these disease states for different patients at different time points to approximate the input data. The formulation of the CNTF model with lab tests and medications is given in Eq. 3.

$$\begin{aligned} \underset{\mathbf{W}^{(p)}, \mathbf{U}^{(l)}, \mathbf{U}^{(m)}}{\arg\min} \quad & f^{\text{CNTF}} \equiv \sum_{p=1}^{N_p} \frac{1}{T_p} \left(\sum_{ijk} \widehat{x}_{ijk}^{(p)} - x_{ijk}^{(p)} \log \widehat{x}_{ijk}^{(p)} \right) \\ \text{subject to} \quad & \widehat{\mathcal{X}}^{(p)} = [\![\mathbf{W}^{(p)}, \mathbf{U}^{(l)}, \mathbf{U}^{(m)}]\!] \ \ \forall p \\ & \mathbf{W}^{(p)} \geq \mathbf{0} \ \ \forall p, \\ & \mathbf{U}^{(l)} \geq \mathbf{0}, \mathbf{U}^{(m)} \geq \mathbf{0}, \end{aligned} \quad (3)$$

where the loss function is given by the weighted sum of that of factorizing each individual temporal tensor. To prevent the total loss being dominated by samples with very long temporal lengths, the individual loss of each sample is weighted by the reciprocal of its temporal length.

The advantages of the proposed schema are twofold.

- **Avoiding temporal resampling or padding.** As emphasized earlier, constructing a 4^{th} order tensor for all patients would require the size of time dimension to be consistent. Downsampling the longer data or zero-padding the shorter data have to be performed, causing raise of bias or loss of information. Neither one is desirable.

- **Revealing patient-specific dynamic patterns.** Even if a 4^{th} order tensor can be constructed, factorizing the tensor with time dimension yields temporal factors that account for the *global* evolution across all data samples (Xiong et al. 2010). Although this could be advantageous under some particular scenarios, we believe it would be preferable to make the temporal factor *specific* for each patient since the disease progression could be very distinct for different individuals, even with the same diagnoses.

To illustrate the second point in more detail, let us assume that a 4^{th} order tensor $\mathcal{X}'$ with size of $N_p \times T_p \times I_l \times I_m$ can be constructed with its CP factors for the four dimensions being $\mathbf{W}'$, $\mathbf{U}^{(t)}$, $\mathbf{U}^{(l)}$ and $\mathbf{U}^{(m)}$ respectively. Then we have $x'_{pijk} = \sum_{r=1}^{R} w'_{pr} u^{(t)}_{ir} u^{(l)}_{jr} u^{(m)}_{kr}$, where $w'_{pr} u^{(t)}_{ir}$ is the patient loading and can be interpreted as the probability of phenotype r being existent for patient p at time i after proper re-scaling. It clearly follows that the patient loading vector over time is essentially the *globally* shared temporal factors weighted by a patient-specific scalar w'_{pr}, resulting the disease progression for all patients following the same dynamic pattern with different amplitude. To the contrary, it is straightforward that the proposed CNTF model reveals the dynamic patterns specific for each individual patient.

Incorporating Non-temporal Data Modality

It is often the case that some data types do not have time stamps. For example, in MIMIC-III dataset (Johnson et al. 2016), the diagnosis codes are generated upon patient discharge for the billing purpose. Thus the time of making diagnosis is not available. Yet the diagnosis information is very useful for discovering clinically meaningful phenotypes. We integrate the diagnosis by adopting the Hidden Interaction Tensor Factorization (HITF) model proposed in (Yin et al. 2018). The HITF model is derived based on the accumulation of diagnoses and medications over the observation window. It takes a *patient-by-medication* counting matrix and a *patient-by-diagnosis* binary matrix as input, and computes the CP factorization of the hidden tensor describing the interactions among the medications and diagnoses. For the p^{th} patient, we sum up the patient representation $\mathbf{W}^{(p)}$ along the time dimension as the representation for the whole observation window. The input then would be a medication vector $\mathbf{m}$ indicating what and how many medications are prescribed to the p^{th} patient and a binary diagnosis vector $\mathbf{d}$ indicating the diagnoses assigned to the patient. We rewrite the formulation of the HITF model for an individual patient as follows:

$$\begin{aligned} \underset{\mathbf{W}^{(p)},\mathbf{U}^{(d)},\mathbf{U}^{(m)}}{\arg\min} \quad & f_p^{\text{HITF}} \equiv \sum_i \hat{d}_i^{(p)} - d_i^{(p)} \log(e^{\hat{d}_i^{(p)}} - 1) + \\ & \qquad \sum_j \hat{m}_j^{(p)} - m_j^{(p)} \log \hat{m}_j^{(p)} \\ \text{subject to} \quad & \hat{\mathbf{d}}^{(p)} = \mathbf{e}^\top \mathbf{W}^{(p)} \operatorname{diag}(\mathbf{e}^\top \mathbf{U}^{(m)}) \mathbf{U}^{(d)\top} \\ & \hat{\mathbf{m}}^{(p)} = \mathbf{e}^\top \mathbf{W}^{(p)} \operatorname{diag}(\mathbf{e}^\top \mathbf{U}^{(d)}) \mathbf{U}^{(m)\top} \\ & \mathbf{W}^{(p)} \geq \mathbf{0}, \mathbf{U}^{(m)} \geq \mathbf{0}, \mathbf{U}^{(d)} \geq \mathbf{0}. \end{aligned} \tag{4}$$

RNN-based Temporal Regularization

Although the temporal relationship can be captured by $\mathbf{W}^{(p)}$ as described earlier, the temporal dependency of the disease state over time is not explicitly modeled, implying each time point being treated independently. However, the independence assumption is not appropriate here for the time dimension, as it does not take into account the ordering of the clinical events, which is inherently important for medical applications. In order to model the temporal dependency, we propose to make use of an RNN which is recently predominant for time series and sequential data analysis. With the dynamic patient representations being learned, we may regard each $\mathbf{W}^{(p)}$ as a multi-variable time series with each variable describing the progression of the existence of the corresponding phenotype for patient p. Given the time series prior to time t, *i.e.* $\mathbf{w}_1, \ldots, \mathbf{w}_{t-1}$ where we omit the superscript (p) and subscript r denoting the patient and phenotype respectively, we use the RNN model to predict $\mathbf{w}_t$ and minimize the Mean Square Error (MSE) between the real and predicted value. The regularization term is written as:

$$\mathcal{R}(\mathbf{W}^{(p)}) = \frac{1}{T_p} \sum_{t=2}^{T_p} \left\| g(\mathbf{w}_{t-1}) - \mathbf{w}_t \right\|_2^2, \tag{5}$$

where $g(\mathbf{w}_{t-1})$ is the prediction output given by the RNN model. In this work, we use a two-layer LSTM network (Hochreiter and Schmidhuber 1997) with 200 hidden units as the RNN model.

As a regularization, the RNN model is jointly learned with the CNTF model. Intuitively, the RNN model captures the temporal dependency with its hidden units, and then the patient representation $\mathbf{W}^{(p)}$ is updated so that the recovery error of the CNTF model and the temporal predictive MSE loss together is minimized, enforcing the patient representation being mostly consistent with the regularity captured by the LSTM network as well as recovering the temporal tensor.

Learning Algorithms

The final loss function is given by the weighted sum of the CNTF loss, the HITF loss and the temporal regularization loss as follows:

$$\ell = \alpha_1 f^{\text{CNTF}} + \alpha_2 \sum_{p=1}^{N_p} f_p^{\text{HITF}} + \beta \sum_{p=1}^{N_p} \mathcal{R}(\mathbf{W}^{(p)}), \tag{6}$$

where the variables $\mathbf{W}^{(p)}\ \forall p$, $\mathbf{U}^{(l)}$ and $\mathbf{U}^{(m)}$ have to satisfy the non-negativity constraints. To ease the parameter tuning, α_1 is fixed to one throughout all the experiments.

The medication vector used in HITF model is the accumulation of the entire hospital visit. Thus, intuitively, emphasizing the HITF loss too much (large α_2) would possibly hinder the medications used at different disease stages being well separated, while a too small α_2 could fail to capture the correspondence between the diagnoses and the lab tests. Therefore, choosing a suitable α_2 is very crucial.

We adopt the block coordinate descent optimization framework and mini-batch projected gradient descent to solve the problem. In each mini-batch, we first sample m data points $\{\mathcal{X}^{(i)}|i \in \mathcal{L}\}$ with the $\mathcal{L}$ being the data point indices. Then, we update $\mathbf{U}^{(l)}$ and $\mathbf{U}^{(m)}$ in turn with all other variables fixed, followed by updating $\mathbf{W}^{(i)}$ $\forall i \in \mathcal{L}$. Lastly, we feed $\mathbf{W}^{(i)}$ $\forall i$ as input to the LSTM network and update it using the standard back-propagation. The optimization procedure is summarized in Algorithm 1.

Algorithm 1: Optimization Framework for Solving LSTM Regularized CNTF Model

Input : *time-labtest-medication* tensor collection: $\{\mathcal{X}^{(p)}|\mathcal{X}^{(p)} \in \mathbb{R}^{T_p \times I_l \times I_m}, p = 1, \ldots, N_p\}$, medication vectors: $\{\mathbf{m}^{(p)}, p = 1, \ldots, N_p\}$, diagnosis vectors: $\{\mathbf{d}^{(p)}, p = 1, \ldots, N_p\}$, model parameters: α_1, α_2 and β.

Output: patient representations: $\mathbf{W}^{(p)}$ $\forall p$, phenotype definitions: $\mathbf{U}^{(l)}$, $\mathbf{U}^{(m)}$ and $\mathbf{U}^{(d)}$.

```
initialization;
for each epoch do
    for each mini-batch do
        sample mini-batch of m tensors and vectors from
          input with indices L;
        for X ∈ {U^(l), U^(m), U^(d)} do
            update X by descending its stochastic gradient;
            non-negative projection by X ← max(0, X);
        end
        for i ∈ L do
            update W^(i) by descending its stochastic
              gradient;
            non-negative projection by
              W^(i) ← max(0, W^(i));
        end
        update LSTM model by back-propagation;
    end
end
```

The time complexity of the CNTF model remains the same with formulating as a 4^{th} order tensor factorization problem, but in practice the CNTF model is more efficient because solving for $\mathbf{W}^{(i)}$ is independent of each other with all other variables fixed, and the gradient *w.r.t.* $\mathbf{U}^{(l)}$, $\mathbf{U}^{(m)}$ and $\mathbf{U}^{(d)}$ can be computed by summing up the gradient of each individual data sample, thus allowing them to be easily parallelized.

Experiments and Results

We conduct the experiments on a real-world Intensive Care Units (ICU) dataset, MIMIC-III, where the quality of the inferred phenotypes is evaluated. Furthermore, we use the inferred phenotypes as features for the mortality prediction task and evaluate the classification accuracy.

Data Set

Medical Information Mart for Intensive Care (MIMIC-III) (Johnson et al. 2016) is a large-scale, open-source and de-identified ICU dataset, containing records related to over forty thousand patients who stayed in the ICU at Beth Israel Deaconess Medical Center between 2001 and 2012. In this paper, we focus on the medication prescriptions, of which the prescription date and duration dates are recorded, and the laboratory test results with time stamps recorded. Since many laboratory tests are requested and performed repeatedly in ICU, we only use the abnormal laboratory test events to avoid the frequent normal laboratory results dominating the input tensor. We construct a 3^{rd} order *time-labtest-medication* binary interaction tensor $\mathcal{X}^{(p)}$ for patient p by setting the tensor entry $x_{tij}^{(p)}$ to be one if the abnormal labtest event i and the medication event j co-occur at time t. The time resolution is one day. We extract a subset of MIMIC-III dataset containing $4,590$ adult patients with length-of-stay longer than 7 days, and 50% of them deceased in the hospital. We also exclude the base type drugs, *e.g.* D5W, and use the top 300 most frequent medications. The diagnosis codes are generated upon patient discharge by reviewing the clinical notes during the hospital stay, and thus the exact time of the diagnoses being assigned is not available. We group the diagnoses by the first three digits of their ICD-9 codes and use the top 300 most frequent diagnoses.

Phenotypes

The primary task of computational phenotyping is to derive clinically meaningful and interpretable phenotypes that correspond to some true disease states. Thus, we first evaluate the quality of the learned phenotypes. In order to include diagnoses in the phenotypes to enhance the interpretability, the HITF model is incorporated as described in Eq. 4, and the weighting α_2 is set to 0.05. The number of phenotypes is set to 50. The RNN regularization is switched on with weight β set to 10.

Table 1 shows three phenotypes derived by our proposed model. It can be seen that the inferred phenotypes correspond to different disease states in ICU, which is endorsed by a medical expert. Phenotype 1 corresponds to the diagnosis, Chronic Kidney Disease (CKD) and the identified abnormal laboratory tests, especially the RBC (Red Blood Cells) in urine, blood osmolality and protein/creatinine ratio in urine. In the clinical context, the disease state CKD is indeed associated with elevated RBC in urine due to renal tubular necrosis, elevated blood osmolality due to electrolyte retention in the vascular system, and elevated protein loss in the urine leading to an abnormal protein/creatinine ratio. Phenotype 9 corresponds to the diagnosis Other Disease of the Lung and abnormal laboratory tests pO2, pCO2, pH of the arterial blood gas. Again, this correlates well with the clinical context, where reduced oxygen levels and pH, and elevated carbon dioxide levels all indicate the presence of acute respiratory failure (which is classified under the "other disease of lung" in the ICD-9 coding system).

We compare our results against the Rubik (Wang et al. 2015) model, which is one of the state-of-the-art computational phenotyping algorithms. We accumulate the observation over the hospital stay and construct a single tensor with four dimensions, *i.e. patient-labtest-medication-diagnosis*.

Phenotype 1	Phenotype 4	Phenotype 9
Chronic kidney disease (CKD) (0.536)	Other forms of chronic ischemic heart disease (0.507) Cardiac dysrhythmias (0.372) Essential hypertension (0.024)	Other diseases of lung (0.876)
RBC (Urine) (0.200)	Hematocrit (Blood) (0.072)	pO2 (Blood Gas) (0.253)
Osmolality, Measured (Blood) (0.117)	Red Blood Cells (Blood) (0.071)	pCO2 (Blood Gas) (0.237)
Protein/Creatinine Ratio (Urine) (0.069)	Hemoglobin (Blood) (0.070)	pH (Blood Gas) (0.215)
Hydromorphone (0.336)	Acetaminophen (0.188)	Acetaminophen (0.113)
Phenylephrine (0.038)	Metoclopramide (0.102)	Insulin (0.099)
Aspirin (0.033)	Insulin Human Regular (0.070)	Bisacodyl (0.089)

Table 1: Three examples of the learned phenotypes. The rows correspond to diagnoses, abnormal laboratory results and medications respectively, where the numbers between parentheses are the weightings. Due to space limitation, only the first three items are listed.

Phenotype 1	Phenotype 2	Phenotype 3
Other diseases of lung (0.045) Septicemia (0.040) Certain adverse effects not elsewhere classified (0.039)	Other diseases of lung (0.040) Acute kidney failure (0.036) Certain adverse effects not elsewhere classified (0.032)	Acute kidney failure (0.039) Other diseases of lung (0.037) Cardiac dysrhythmias (0.033)
Glucose(Blood) (0.019)	Hematocrit(Blood) (0.017)	Glucose(Blood) (0.018)
Red Blood Cells(Blood) (0.019)	Red Blood Cells(Blood) (0.017)	Hematocrit(Blood) (0.018)
Hematocrit(Blood) (0.019)	Glucose(Blood) (0.017)	Red Blood Cells(Blood) (0.018)
Vancomycin (0.017)	Vancomycin (0.013)	Vancomycin (0.015)
Insulin (0.015)	Potassium Chloride (0.013)	Potassium Chloride (0.014)
Potassium Chloride (0.015)	Pantoprazole Sodium (0.012)	Heparin (0.014)

Table 2: Three examples of the phenotypes derived by the Rubik model. The rows correspond to diagnoses, abnormal laboratory results and medications respectively. Due to space limitation, only the first three items are listed.

Table 2 shows the phenotypes derived by the Rubik model, where we can see that the weightings of the clinical items within each phenotype are widely distributed, instead of concentrating on some specific items. The inferred phenotypes all correspond to some complex, critical and possibly end-stage disease states, including the diagnoses of septicemia and acute kidney failure, and medication of vancomycin which is often used in ICU for treatment of life-threatening infections by Gram-positive bacteria that are unresponsive to other antibiotics. Moreover, the identified abnormal laboratory tests are very general, which do not specifically relate to either the diagnoses or the medications.

The comparison between the phenotypes derived by our proposed CNTF model and the Rubik model reveals that it is extremely difficult to separate the disease states appearing at different stages of the patient journey given the input tensor being accumulated over the observation window. With our proposed CNTF model; however, the different disease states occurring at different time points could be discovered, reflected by the fact that the chronic diseases, such as CKD, can be captured with meaningful combinations of medications and abnormal laboratory tests. Therefore, we conclude that our proposed CNTF model infers significantly more interpretable and clinically meaningful phenotypes than the baseline.

	Sparsity	Similarity
Rubik	0.79	0.90
CNTF	**0.96**	**0.43**

Table 3: Sparsity and similarity of phenotypes derived by the proposed CNTF model and the baseline Rubik model.

Sparsity and Similarity

Sparsity and similarity are two commonly used proxy metrics for measuring the interpretability of the derived phenotypes quantitatively (Kim et al. 2017; Yin et al. 2018). The sparsity is defined by the ratio of zero elements in the phenotype definition matrices, and the similarity is defined as the average cosine similarity score given by:

$$\text{Similarity Score} = \frac{\sum_k \sum_{r_1}^{R} \sum_{r_2 > r_1}^{R} \left\{ \cos(\mathbf{U}^{(k)}_{:r_1}, \mathbf{U}^{(k)}_{:r_2}) \right\}}{R(R-1)}, \tag{7}$$

where R is the number of phenotypes. Table 3 shows the sparsity and the similarity of the phenotypes inferred by our proposed CNTF model and that of the baseline Rubik model. It is evident that CNTF derives much more sparse and distinct phenotypes compared with the Rubik model.

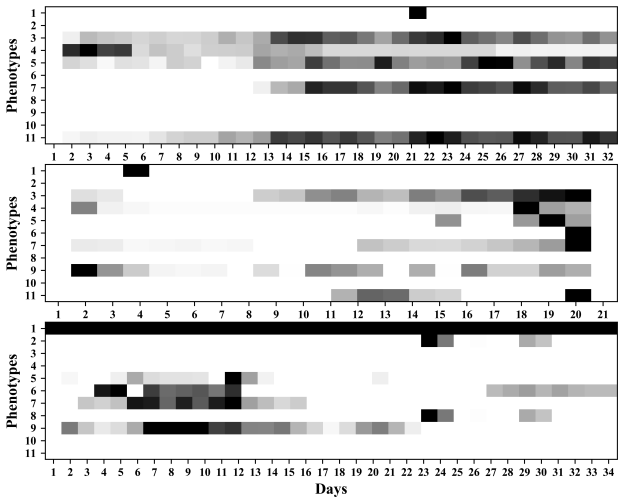

Figure 2: Visualization of three examples of the dynamic patient representations. Each row corresponds to a phenotype, and the grey level indicates the weighting of the phenotype at different time points (normalized by the maximum value of each row). The definitions of phenotype 1, 4 and 9 are given in Table 1, and that of the remaining phenotypes are given in the supplemental material.

Interpretation of the Dynamic Patient Representations

As described earlier, the dynamic patient representation $\mathbf{W}^{(p)}$ indicates the evolution of the disease states over the observation window. With meaningful phenotypes being inferred, we anticipate that the patient representations are also highly interpretable. Fig. 2 shows the visualization of three examples of the dynamic patient representations learned together with the phenotype definitions. Each sub-figure corresponds to one individual patient, where each row within each sub-figure corresponds to one particular phenotype, and the grey level of each cell indicates the strength of the corresponding phenotype being present at that time. We normalize the values by the maximum value of each row for better visual effect.

We presented the visualization to a medical expert for qualitative evaluation. According to the expert, the learned patient representations are highly interpretable. The first example patient has phenotype 4, which corresponds to the disease "Chronic Heart Disease", with high values in the first several days and decreasing in the remaining of the hospital stay. This suggests that the lab tests and medications are related to this disease entity only during the initial few days of the ICU stay. This patient's data then goes on to demonstrate high values for phenotypes 3, 5, 7 and 11, which correspond to Other Disease of the Lung, Cardiac Dysrhythmias, Acute Kidney Failure, and Cardiac Dysrhymias with Heart Failure, respectively. Essentially, the data describe a clinical scenario in which the patient is admitted with a problem related to an existing condition (chronic heart disease) which is treated unsuccessfully, so the patient deteriorates and develops multiple organ failure (lung, heart, kidney failure). Indeed, closer review of the clinical textual documentation of this patient shows that the aforementioned scenario does closely correlate with what actually occurred.

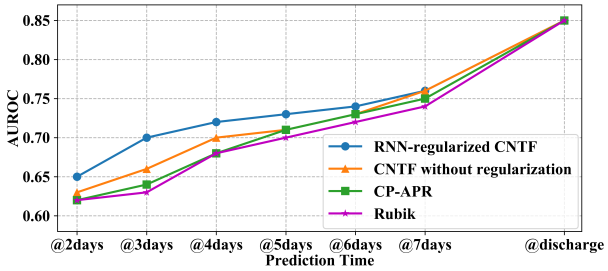

Figure 3: Prediction accuracy of in-hospital mortality at different time

Mortality Prediction Task

We further evaluate the derived phenotypes by performing an in-hospital mortality prediction task using the derived phenotypes as features. We split the data into training set and test set with a proportion of $8:2$. The phenotypes are derived based on the training set, which is totally unsupervised. Then we fix the learned phenotype definitions and project the test set onto the learned phenotypes to obtain the patient representation for the test set. Finally we use a lasso regularized logistic regression to perform the binary classification. We measure the AUROC on different days in an accumulated manner, *e.g.*, the AUROC value for the second day is obtained by considering the patient representations within the first two days. The HITF model is switched off by setting α_2 in Eq. 6 to zero for this task since the diagnoses codes are only available after discharge. We avoid using diagnosis codes for making predictions prior to discharge to ensure fair evaluation.

We compare our proposed CNTF model with two baselines, Rubik and CP-APR, where the latter is a commonly used CP-APR factorization model. The baseline models do not take into account the time dimension, and only accumulate the data over the observation window to infer the phenotypes. In particular, they accumulate the data over the window prior to the prediction, *e.g.*, three days, and project them to the learned phenotypes to obtain the patient representation for the period prior to the prediction. Then, prediction is carried out. The results are shown in Fig. 3. We can see that the proposed CNTF model outperforms all the baselines for making predictions prior to discharge. Without the temporal regularization, the CNTF achieves marginal improvement compared with the baselines, and after adding the RNN-based regularization, the performance further improves significantly. Using the CNTF with the RNN regularization, we can achieve AUROC of 0.70 for the third day, which can only be achieved by the baselines after five days. The significant improvement of the prediction performance is very likely due to the fact that the phenotypes derived by CNTF can well represent the patients at times across the whole hospital stay, while the phenotypes derived by the

baselines cannot. This also validates that CNTF can infer phenotypes that correspond to more specific disease states, rather than mixtures of different disease states. Finally at discharge, all models achieve AUROC of 0.85, which is not surprising since the patients can be well represented by the baseline phenotype given the data accumulated over the whole hospital stay.

Conclusion

In this paper, we present a novel Collective Non-negative Tensor Factorization (CNTF) model to simultaneously learn the dynamic patient representations that are specific for each individual patient, and the phenotype definitions that are shared across all the patients. The proposed model takes into account the varying length of the patient records by forming a temporal tensor for each patient, the non-temporal data modalities by incorporating the HITF model, and the temporal dependency of the disease states by introducing an RNN-based regularization.

The experimental results demonstrate that the phenotypes inferred by CNTF are clinically meaningful and interpretable, and correspond to different specific disease states that occur at different times of the patient journey, which cannot be easily obtained by the baseline model with the input tensor being accumulation over the observation window. Moreover, this is also validated by the significant predictive performance boost in the early stage of the hospital admission. For future research directions, we will focus on utilizing data with different temporal resolutions to discover more clinically relevant phenotypes.

Acknowledgments

This research is partially supported by General Research Fund 12202117 from the Research Grants Council of Hong Kong.

References

Che, Z.; Purushotham, S.; Cho, K.; Sontag, D.; and Liu, Y. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports* 8(1):6085.

Chi, E. C., and Kolda, T. G. 2012. On tensors, sparsity, and nonnegative factorizations. *SIAM Journal on Matrix Analysis and Applications* 33(4):1272–1299.

Choi, E.; Bahadori, M. T.; Schuetz, A.; Stewart, W. F.; and Sun, J. 2016a. Doctor AI: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, 301–318.

Choi, E.; Bahadori, M. T.; Sun, J.; Kulas, J.; Schuetz, A.; and Stewart, W. 2016b. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, 3504–3512.

Henderson, J.; Ho, J. C.; Kho, A. N.; Denny, J. C.; Malin, B. A.; Sun, J.; and Ghosh, J. 2017. Granite: Diversified, sparse tensor factorization for electronic health record-based phenotyping. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, 214–223. IEEE.

Ho, J. C.; Ghosh, J.; Steinhubl, S. R.; Stewart, W. F.; Denny, J. C.; Malin, B. A.; and Sun, J. 2014. Limestone: High-throughput candidate phenotype generation via tensor factorization. *Journal of Biomedical Informatics* 52:199–211.

Ho, J. C.; Ghosh, J.; and Sun, J. 2014. Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 115–124. ACM.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.

Hripcsak, G., and Albers, D. J. 2013. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association* 20(1):117–121.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Li-wei, H. L.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data* 3:160035.

Kim, Y.; El-Kareh, R.; Sun, J.; Yu, H.; and Jiang, X. 2017. Discriminative and distinct phenotyping by constrained tensor factorization. *Scientific Reports* 7(1):1114.

Kirby, J. C.; Speltz, P.; Rasmussen, L. V.; Basford, M.; Gottesman, O.; Peissig, P. L.; Pacheco, J. A.; Tromp, G.; Pathak, J.; Carrell, D. S.; et al. 2016. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *Journal of the American Medical Informatics Association* 23(6):1046–1052.

Kolda, T. G., and Bader, B. W. 2009. Tensor decompositions and applications. *SIAM Review* 51(3):455–500.

Lee, D. D., and Seung, H. S. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755):788.

Perros, I.; Papalexakis, E. E.; Wang, F.; Vuduc, R.; Searles, E.; Thompson, M.; and Sun, J. 2017. SPARTan: Scalable PARAFAC2 for large & sparse data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 375–384. ACM.

Purushotham, S.; Meng, C.; Che, Z.; and Liu, Y. 2018. Benchmarking deep learning models on large healthcare datasets. *Journal of Biomedical Informatics*.

Wang, Y.; Chen, R.; Ghosh, J.; Denny, J. C.; Kho, A.; Chen, Y.; Malin, B. A.; and Sun, J. 2015. Rubik: Knowledge guided tensor factorization and completion for health data analytics. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1265–1274. ACM.

Xiong, L.; Chen, X.; Huang, T.-K.; Schneider, J.; and Carbonell, J. G. 2010. Temporal collaborative filtering with Bayesian probabilistic tensor factorization. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, 211–222. SIAM.

Yadav, P.; Steinbach, M.; Kumar, V.; and Simon, G. 2018. Mining electronic health records (EHRs): a survey. *ACM Computing Surveys (CSUR)* 50(6):85.

Yang, K.; Li, X.; Liu, H.; Mei, J.; Xie, G.; Zhao, J.; Xie, B.; and Wang, F. 2017. TaGiTeD: Predictive task guided tensor decomposition for representation learning from electronic health records. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*.

Yin, K.; Cheung, W. K.; Liu, Y.; Fung, B. C. M.; and Poon, J. 2018. Joint learning of phenotypes and diagnosis-medication correspondence via hidden interaction tensor factorization. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 3627–3633.

Yu, H.-F.; Rao, N.; and Dhillon, I. S. 2016. Temporal regularized matrix factorization for high-dimensional time series prediction. In *Advances in Neural Information Processing Systems*, 847–855.