RESEARCH ARTICLE

Readmission prediction via deep contextual embedding of clinical concepts

Cao Xiao[1☯], **Tengfei Ma**[2☯], **Adji B. Dieng**[3], **David M. Blei**[3], **Fei Wang**[4*]

1 AI for Healthcare, IBM Research, Cambridge, MA, United States of America, **2** IBM T.J. Watson Research Center, Yorktown Heights, NY, United States of America, **3** Department of Computer Science, Columbia University, New York, NY, United States of America, **4** Weill Cornell Medical School, Cornell University, New York, NY, United States of America

☯ These authors contributed equally to this work.
* few2001@med.cornell.edu

Abstract

Objective

Hospital readmission costs a lot of money every year. Many hospital readmissions are avoidable, and excessive hospital readmissions could also be harmful to the patients. Accurate prediction of hospital readmission can effectively help reduce the readmission risk. However, the complex relationship between readmission and potential risk factors makes readmission prediction a difficult task. The main goal of this paper is to explore deep learning models to distill such complex relationships and make accurate predictions.

Materials and methods

We propose CONTENT, a deep model that predicts hospital readmissions via learning interpretable patient representations by capturing both local and global contexts from patient Electronic Health Records (EHR) through a hybrid Topic Recurrent Neural Network (TopicRNN) model. The experiment was conducted using the EHR of a real world Congestive Heart Failure (CHF) cohort of 5,393 patients.

Results

The proposed model outperforms state-of-the-art methods in readmission prediction (e.g. 0.6103 ± 0.0130 vs. second best 0.5998 ± 0.0124 in terms of ROC-AUC). The derived patient representations were further utilized for patient phenotyping. The learned phenotypes provide more precise understanding of readmission risks.

Discussion

Embedding both local and global context in patient representation not only improves prediction performance, but also brings interpretable insights of understanding readmission risks for heterogeneous chronic clinical conditions.

OPEN ACCESS

Citation: Xiao C, Ma T, Dieng AB, Blei DM, Wang F (2018) Readmission prediction via deep contextual embedding of clinical concepts. PLoS ONE 13(4): e0195024. https://doi.org/10.1371/journal.pone.0195024

Editor: Chenping Hou, National University of Defense Technology, CHINA

Received: August 23, 2017

Accepted: March 15, 2018

Published: April 9, 2018

Data Availability Statement: All relevant data are within the paper, its Supporting Information files, and GitHub. Code and additional data can be accessed at https://github.com/danicaxiao/CONTENT.

Funding: This work was supported by National Science Foundation IIS-1650723 and IIS-1716432 to FW. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Conclusion

This is the first of its kind model that integrates the power of both conventional deep neural network and the probabilistic generative models for highly interpretable deep patient representation learning. Experimental results and case studies demonstrate the improved performance and interpretability of the model.

Introduction

A hospital readmission is defined as the admission to a hospital within a short amount of time after discharge, where 30-day is typically considered a clinically meaningful time window [1]. Excessive hospital readmissions disrupt the normality of patients' lives and have negative impacts on the healthcare systems [2]. For example, in the US, it has been reported by the Medicare Payment Advisory Committee that 17.6% of hospital-admitted patients were readmitted within 30 days of discharge, which accounted for $17:9 billion Medicare spending per year, while 76% of them are potentially avoidable [1]. To curb hospital readmission rates, the Patient Protection and Affordable Care Act was set up to penalize hospitals with excessive readmission at a minimum of 3% of their Medicare reimbursement. Despite the efforts, it is estimated that the scrutiny of readmission rates will continue to grow over the next few years.

To prevent excessive readmissions, procedures such as patient follow-ups and educations have been implemented, which could be costly for individual patient. Therefore, targeted follow-ups that focus on patients with high risks of readmissions are preferred. This raises the demand for assessing patient readmission risks and consequently brings the readmission prediction to the forefront of healthcare research. Accurate prediction of hospital readmission is difficult because of its complex entanglements with the patients' health conditions, especially the chronic ones. In recent years, there have been some research on hospital readmission prediction from patient Electronic Health Records (EHRs) [1–6]. There are many challenges for working with EHR such as its incompleteness, noisiness, heterogeneity, etc. [7], and the existing research typically needs to rely on appropriate feature engineering [4, 8], whose optimality is difficult to justify from both computational and clinical perspectives.

In order to solve the challenges, we seek for deep learning models to perform readmission predictions. Deep learning models are well known for their end-to-end learning capabilities so we do not need to worry about the feature engineering part [9, 10]. Moreover, deep learning models are proved to be very powerful at distilling the complicated relationships hidden in the data and thus demonstrate good prediction performance [10, 11]. In this paper, we develop CONTENT, which is a deep learning model that transforms patients' complicated event structures in their EHR into deep clinical concept embedding, which can be viewed as a novel form of patient representation encoding the patient clinical conditions from both long and short terms. We draw the analogy between EHR modeling and natural language models [12] to consider the short-term dependencies among medical events in EHRs as local context of a patient journey and long-term effect as global context. Such contexts impact the latent relations between the clinical variables (e.g. diagnoses, procedures, medications, etc.) and the target variable (i.e., readmission). We design a hybrid deep learning model structure that combines topic modelling [13] and Recurrent Neural Network (RNN) [14] to distill the complex knowledge hidden in those contexts and perform accurate readmission prediction.

It is worthwhile to highlight the following aspects of the proposed CONTENT model.

- The proposed model explores both the global and local contexts within the patient journey from his/her EHRs. The global context (the general conditions of the patient, such as those chronic diseases, comorbidities, etc.) is captured by topic models and local context (the short term disease progressions) is captured by RNN. In this way, we can better capture the heterogeneities across different patient individuals and make the model more precise. Empirical results also show the joint modeling could achieve better overall performance evaluated on the readmission prediction tasks.
- Because of the incorporation of the global context, the resultant model is more interpretable comparing to simple RNN models. Our model will produce a context vector for each patient, which characterizes his/her overall condition.

Background

Predictive modeling and deep learning

Most of the existing works on predicting 30-day hospital readmissions were developed with administrative claims with certain components from EHR such as vital signs and lab tests [1–3, 5]. Those events are typically aggregated over a certain period of time (a.k.a. observation window) with some simple feature transformation [4, 8], and then fed into a predictor such as logistic regression or random forest for the prediction task [1, 6].

One limitation of those conventional approaches is that they cannot take the time information into account. The temporalities of the events in patient records are crucial because they can potentially suggest the progression pattern of the patient conditions. Recently, researchers have been exploring deep learning models, such as Convolutional Neural Network (CNN) [15, 16] and Recurrent Neural Network (RNN) [17–20] to capture the complex temporal relationships among the medical events. For example, in [15], the authors proposed a multilayered convolutional neural nets (CNNs) to extract complex patient representations that capture convoluted relations among various clinical events. In [16], each patient's EHR is represented as a temporal matrix with time on one dimension and medical events on the other dimension, and a four-layer CNN model is built for extracting representations. An RNN model was adopted for predicting the onset risk for heart failure patients from their EHRs [17]. A temporal Long Short Term Memory (LSTM, which is a variant of RNN) model is proposed to capture the progression patterns for Parkinson's disease [18].

These existing works typically construct a unique model for the entire patient cohort. Because of the high heterogeneity of the disease conditions across patient individuals and the complexity of the dependencies of hospital readmission and the medical events within patient EHRs, it would be very difficult to learn a single model that can capture all those complexities with a limited number of patients. The proposed model in this paper assigns a global patient-specific context vector for each patient, and the prediction for the patient is dependent on both the context vector and an RNN. In this way, we can model the patients more precisely.

RNN and contextual RNN

An RNN is a fully connected neural network with recurrent connections in its hidden layer [10]. They take the input at current time step t along with the hidden state at time step t − 1 to compute the current hidden state. Mathematically, an RNN defines the conditional probability of each input w_t given all of the previous inputs $w_{1:\,t-1}$ through a hidden state h_t via a softmax function:

$$p(w_t|w_{1:t-1}) \triangleq p(w_t|\boldsymbol{h}_t)$$

$$\boldsymbol{h}(t) = f(\boldsymbol{h}_{t-1}, w_{t-1})$$

The function *f*(·) can either be a standard RNN cell or a more complex cell such as gated recurrent (GRU) unit [21] or long short-term memory (LSTM) unit [22]. In this paper, we choose the GRU cell for CONTENT model because it can achieve similar effects as LSTM with a much simpler structure. More concretely, GRU can overcome the vanishing gradient problem as well as capture the effect of long-term dependencies with a sophisticated gating mechanism. Given input $\boldsymbol{x}_t$, the function *GRU*(·) updates hidden states as follows.

$$\begin{aligned}
\boldsymbol{z}_t &= \sigma(\boldsymbol{U}_z\boldsymbol{x}_t + \boldsymbol{W}_z\boldsymbol{h}_{t-1}) \\
\boldsymbol{r}_t &= \sigma(\boldsymbol{U}_r\boldsymbol{x}_t + \boldsymbol{W}_r\boldsymbol{h}_{t-1}) \\
\tilde{\boldsymbol{h}}_t &= tanh(\boldsymbol{U}_h\boldsymbol{x}_t + \boldsymbol{r}_t \odot \boldsymbol{W}_h\boldsymbol{h}_{t-1}) \\
\boldsymbol{h}_t &= \boldsymbol{z}_t \odot \boldsymbol{h}_{t-1} + (1 - \boldsymbol{z}_t) \odot \tilde{\boldsymbol{h}}_t
\end{aligned}$$

where $\sigma(\cdot)$ denotes the sigmoid function and $\odot$ denotes the element wise multiplication; $\boldsymbol{x}_t$ is the input at time step *t*, $\boldsymbol{h}_{t-1}$ is the previous hidden state. $\boldsymbol{U}_z$ and $\boldsymbol{W}_z$ are weight matrices for update gate $\boldsymbol{z}_t$, and $\boldsymbol{U}_r$ and $\boldsymbol{W}_r$ are weight matrices for the reset gate r_t. We drop the biases here for simplicity of notation. In this formulation, the update gate selects whether the hidden state is updated with a new hidden state $\tilde{\boldsymbol{h}}_t$. The reset gate $\boldsymbol{r}_t$ decides whether the previous hidden state $\boldsymbol{h}_{t-1}$ is ignored [21].

Although in principle RNN-based models can "remember" arbitrarily long span history if provided enough capacity, in practice such large-scale neural networks can easily encounter difficulties during optimization or overfitting [23, 24]. Thus, several contextual recurrent neural network models were proposed to explicitly model long span context to improve learning [7, 9, 13, 25–27]. In language models, since much of the long span context comes from semantic coherence, and the topic models [13] can be used to capture global semantic coherency. Therefore, the recently proposed TopicRNN model [26] uses topic models in a recognition network to directly capture long-range semantic dependencies (i.e. global context) via latent topics. These latent topics are then used as additional bias to the output layer of an RNN-based model. In this study, CONTENT is an extension of the contextual RNN model, particularly the TopicRNN model, with hierarchical inputs ("hospital visits" and "clinical events") and sequential binary outputs (indication of readmission) at the "visit" level in EHR data. In addition, the CONTENT does not model stop words.

Materials and method

Data description

In this work, we conducted the experiment using data from a real world EHR repository of Congestive Heart Failure (CHF) cohort including 5,393 patients. The input data includes disease, lab test, and medication codes, all binary encoded indicating their occurrence or absence. The CHF cohort is constructed by clinical experts according to the following criteria: 1) ICD-9 diagnosis of heart failure appeared in the EHR for at least two outpatient encounters, indicating consistency in clinical assessment, and 2) At least one medication was prescribed with an associated ICD-9 diagnosis of heart failure. In addition, the diagnosis date was defined as its first appearance in the record. These criteria have also been previously validated as part of Geisinger Clinical involvement in a pay-for-performance pilot study conducted by Centers for Medicare and Medicaid Services (CMS) [28]. More details could be found in [29]. A sample of EHR record segments is illustrated in Fig 1. In addition to the CHF dataset, we also evaluated

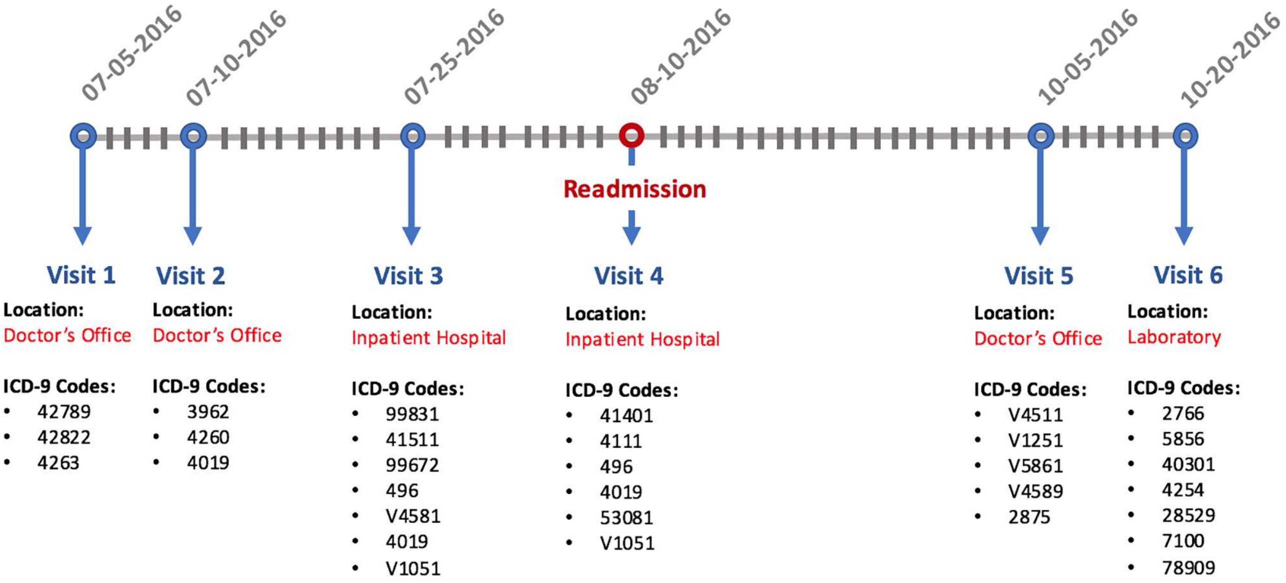


Fig 1. An example segment of EHR records, where visits could occur to different locations. Patients who are re-admitted to "inpatient hospital" within 30 days of their releases from "inpatient hospital" are considered readmissions.

https://doi.org/10.1371/journal.pone.0195024.g001

based on a synthetic EHR data simulated from a de-identified real world patient dataset. The synthetic data is generated as follows: for each original patient record we randomly sample 30% to 50% of the visits in that record and drop the un-sampled visits. After subsampling, we permute patient index. Next, for each new patient record, we randomly combine it with another new record, with the event time of the second patient record being aligned to the first one. We consider such combined record as one synthetic patient record. Following this approach, we generated 3000 synthetic patients, of which 2000 are used in model training, 500 for validation, and 500 for testing. The synthetic data will serve as a benchmark for reproducing experimental results in this paper. However, since they cannot faithfully reflect real patient conditions, the performance comparison will more rely on the real world CHF data. We will also only discuss the learned patient patterns based on the results from the real world data. The basic statistics for both datasets are summarized in Table 1.

The CONTENT model

We formalize the CONTENT model in this section. Denote C as the number of medical events in the EHR data and $\{c_1, \cdots, c_C\}$ as the set of medical events. Each patient $\boldsymbol{p}$ makes T_p visits $\boldsymbol{V}_1, \cdots, \boldsymbol{V}_{T_p}$, where the visit $\boldsymbol{V}_t$ at time t can be represented using a subset of medical events.

Table 1. Basic statistics of CHF and synthetic datasets.

Dataset	Congestive Heart Failure	Synthetic EHR Data
# patients	5, 393	3, 000
# visits	455, 106	239, 936
# events	1, 306, 685	685, 482
Avg. # of visits per patient	84.4	79.98
Avg. # of events per patient	242.3	228.49
# of unique event codes	618	618

https://doi.org/10.1371/journal.pone.0195024.t001

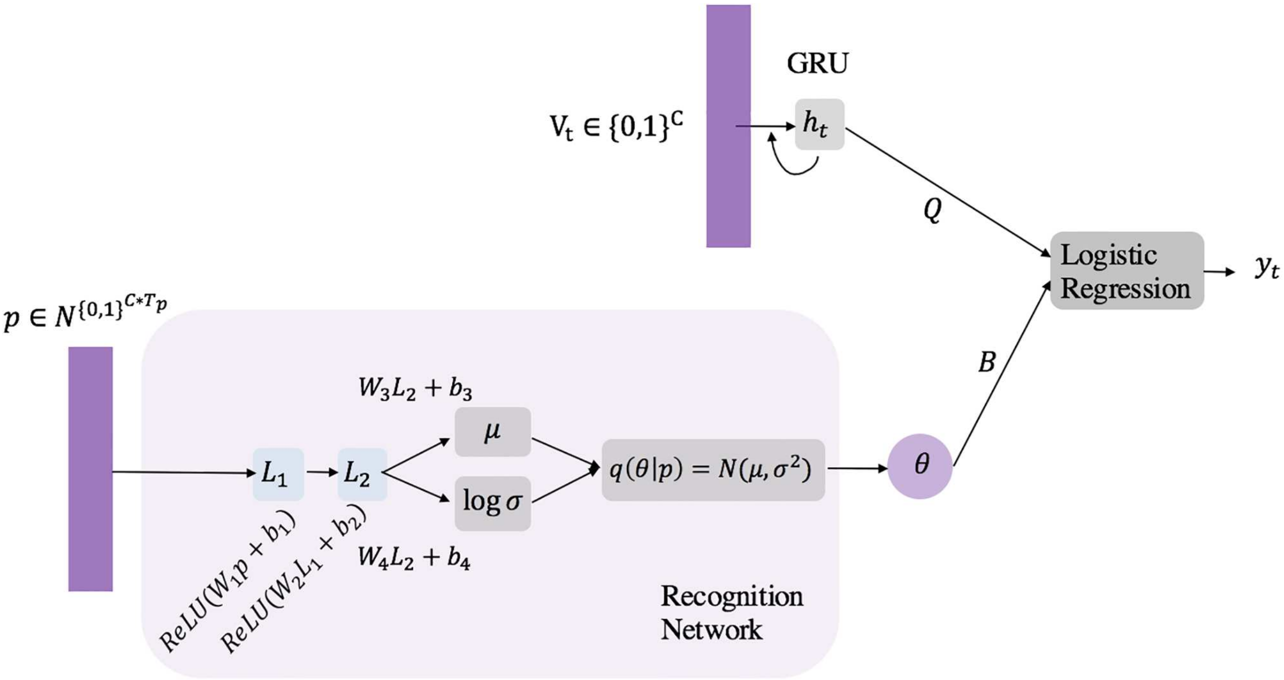


Fig 2. The CONTENT model.

https://doi.org/10.1371/journal.pone.0195024.g002

Given such a structure, it is easy to draw an analogy between data in our model and language models: the set of patients can be considered as the document corpus, the EHRs of each patient can be regarded as a separate document, and each visit of a specific patient can be viewed as a paragraph in a document. Thus representing a patient as a sequence of visits is just as representing a document as a sequence of paragraphs. The difference is that in our case all events within the same visit are treated as simultaneous events. With such an analogy, the CONTENT model can be similarly constructed as a language model. We denote $\boldsymbol{y} = \{y^1, \cdots, y^N\}$ as the observed patient hospital admission indicators, $\boldsymbol{h} = \{\boldsymbol{h}^1, \cdots, \boldsymbol{h}^N\}$ as the collection of RNN hidden states for all patients with $\boldsymbol{h}^p = \{\boldsymbol{h}_1^p, \ldots, \boldsymbol{h}_{V_p}^p\}$ being the RNN hidden state sequence for patient $\boldsymbol{p}$. Θ is the collection of all model parameters, and θ is the hidden variable which represents the context vector. The hospital readmission prediction will be made based on the combination of the patient context vector and the hidden state of an RNN model. Fig 2 provides an illustration of the CONTENT model.

The CONTENT model is essentially a generative model. Its generative process is described as follows: for a particular patient $\boldsymbol{p}$ with visits $\boldsymbol{V}_{1:T_p}$,

1. Draw patient context vector $\boldsymbol{\theta} \sim N(0, \boldsymbol{I})$.
2. For the tth visit,
   a. Computer hidden state $\boldsymbol{h}_t = GRU(\boldsymbol{V}_{t-1}, \boldsymbol{h}_{t-1}; \boldsymbol{W}_v, \boldsymbol{W}_h)$,
   b. Compute logit score $\boldsymbol{z}_t = \boldsymbol{Q}^t \boldsymbol{h}_t + \boldsymbol{B}_t^T \boldsymbol{\theta}, \boldsymbol{B}_t = \frac{1}{M_t} \sum \boldsymbol{b}_m$ and b_m is the latent topic vector for medical code m in this visit; M_t is the number of codes in the visit.
   c. Compute readmission indicator $y_t \sim \sigma(z_t)$.

Assume that the dimension of the latent word embedding is *H*, and the dimension of topics is *N*. The parameters of the model include the word embedding matrix $\boldsymbol{W}_v \in \mathbb{R}^{C\times H}$, where *C* is the number of distinct words (medical events). $\boldsymbol{W}_h$ is the RNN parameter set $\{\boldsymbol{U}_z, \boldsymbol{W}_z, \boldsymbol{z}_t, \boldsymbol{U}_r, \boldsymbol{W}_r, \boldsymbol{r}_t\}$ in the $GRU(\cdot)$ functions as defined in previous section. Here we also have $\boldsymbol{Q} \in \mathbb{R}^H$ and $\boldsymbol{b}_m \in \mathbb{R}^N$. Following the general auto-encoding variational Bayes model and the TopicRNN, we also use the multivariate Gaussian prior for $\boldsymbol{\theta}$ to make it easier for inference. The context vector $\boldsymbol{\theta} \in \mathbb{R}^N$ encodes the patient's contextual information, which can be regarded as the different clinical subtypes of the hospital readmission task.

Model inference

To make predictions, ideally we need to maximize the log marginal likelihood:

$$\log p(\boldsymbol{y}|\boldsymbol{h}, \boldsymbol{\Theta}) = \log \int p(\boldsymbol{y}|\boldsymbol{h}, \boldsymbol{\Theta}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

However, directly optimizing it is intractable [30], so we adopt approximate variational inference techniques [30] to approximate it. Let $q(\boldsymbol{\theta})$ be the variational distribution that approximates the intractable posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$. The log marginal likelihood could be rewritten as

$$\log p(\boldsymbol{y}|\boldsymbol{h}, \boldsymbol{\Theta}) = D_{KL}(q(\boldsymbol{\theta})|||p(\boldsymbol{\theta}|\boldsymbol{y})) + ELBO.$$

Here, the first term is the Kullback-Leibler (KL) divergence that measures the distance between the approximate distribution $q(\boldsymbol{\theta})$ and the true posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$. The second term is the evidence lower bound (ELBO) [31] with the following form:

$$ELBO = E_{q(\boldsymbol{\theta})}[\log p(\boldsymbol{y}|\boldsymbol{h}, \boldsymbol{\theta}, \boldsymbol{\Theta}) + \log p(\boldsymbol{\theta}) - \log q(\boldsymbol{\theta})] \leq \log p(\boldsymbol{y}|\boldsymbol{h}, \boldsymbol{\Theta})$$

where the ELBO is the variational objective function to be optimized. It is a lower bound to the marginal log likelihood by positivity of the KL divergence. It therefore constitutes a principled objective for optimizing the log marginal likelihood.

Following TopicRNN [26] and the recent techniques in deep generative models [30], we formulate $q(\boldsymbol{\theta})$ as an inference network using a feed-forward neural network. The inference network takes the patient representation matrix $\boldsymbol{p}$ as the input, and then project it onto a lower dimensional subspace using a multilayer perceptron (MLP) as formulated below.

$$\begin{aligned}
\boldsymbol{r}_1 &= ReLU(\boldsymbol{W}_{r_1}\boldsymbol{p} + \boldsymbol{b}_{r_1}) \\
\boldsymbol{r}_2 &= ReLU(\boldsymbol{W}_{r_2}\boldsymbol{r}_1 + \boldsymbol{b}_{r_2}) \\
\boldsymbol{\mu}(\boldsymbol{p}) &= \boldsymbol{W}_{\mu}\boldsymbol{r}_2 + \boldsymbol{b}_{\mu} \\
\log \boldsymbol{\sigma}(\boldsymbol{p}) &= \boldsymbol{W}_{\sigma}\boldsymbol{r}_2 + \boldsymbol{b}_{\sigma} \\
q(\boldsymbol{\theta}|\boldsymbol{p}) &= N(\boldsymbol{\mu}(\boldsymbol{p}), diag(\boldsymbol{\sigma}^2(\boldsymbol{p}))) \\
\boldsymbol{\theta} &\sim q(\boldsymbol{\theta}|\boldsymbol{p}).
\end{aligned}$$

For each time step, we update the RNN hidden state at time stamp *t*, $\boldsymbol{h}_t$, to model the sequence of visits for the patient. We do so by representing a visit at time *t* as a binary vector $\boldsymbol{V}_t \in \{0, 1\}^C$, where the *i*-th entry is 1 only if $c_i \in \boldsymbol{V}_t$ with c_i being the *i*-th distinct medical event in the dictionary. For the T_p-th visit, the inference network only takes the patient matrix from previous visits of the patient, $p[0{:}\ T_p]$ of size T_p, as the input. Thus, the patient matrix $\boldsymbol{p}$ is of dimension $C \times T_p$ where T_p is the number of the visits of the patient. The hidden state is then updated following the GRU update rule: $\boldsymbol{h}_t = GRU(\boldsymbol{V}_{t-1}, \boldsymbol{h}_{t-1}; \boldsymbol{W}_v, \boldsymbol{W}_h)$. The hospital readmission indicator y_t at time step t is then computed via logistic regression using both the hidden state of the RNN $\boldsymbol{h}_t$ and the patient-specific context vector $\boldsymbol{\theta} : \boldsymbol{y}_t = \sigma(\boldsymbol{Q}^T \boldsymbol{h}_t + \boldsymbol{B}_t^T \boldsymbol{\theta})$. Note this approach is highly scalable since it does not have the bottleneck of computing the normalization constant of the softmax function as is the case in language models.

The ELBO depends on all parameters of the model, including the weight matrices of the recognition network and the recurrent neural network. The learning procedure for the CONTENT model is to estimate the optimal values of those parameters by using stochastic gradient descent (Adam) [32] with back-propagation through time.

Clinical concept and patient embedding

The projection matrix $\boldsymbol{W}_v$ can be thought of as a matrix that embeds the clinical concepts (i.e., medical events) into the low dimensional space. The context vector $\boldsymbol{\theta}$ sampled from the recognition network serves as a distributed representation of the patient's medical history. Then we can represent each patient as the concatenation of the context vector and the final hidden state vector of the RNN. In the empirical studies, we will demonstrate their representation power by clustering patients using these vector representations.

Evaluation strategy

We assess the performance of the proposed CONTENT model on the task of CHF patient readmission prediction. Specifically, we predict whether a CHF patient who is currently in hospital will be re-admitted as "in hospital" within 30 days of his or her release from the current "in hospital" episode. Since the task is a binary classification, we choose the area under the receiver operating characteristic curve (ROC-AUC), the area under the precision-recall curve (PR-AUC), and the accuracy (ACC) as three measures. A model with higher ROC-AUC or PR-AUC is considered a better model. Advantages of AUCs as metrics are that they do not require choosing a threshold for assigning labels to scores and that they are independent of class bias in the test set.

Model implementation

The proposed model is implemented using Theano 8.2 [33]. Code can be found in https://github.com/danicaxiao/CONTENT. RNN was implemented as a Gated Recurrent Unit (GRU). The word embedding sequences are used as inputs, and a logistic regression is applied over the hidden layer. The hyper-parameters of CONTENT and baselines are set as follows: 1) for word embedding via word2vec [25], we get word vectors of 100 dimensions. 2) the size of hidden layers of RNN is 200. Training is done through Adam at learning rate 0.001 with shuffled mini-batches of batch size 1. For model comparison, we split the data into training (4000 patients), validation (700 patients), and testing (693 patients). We train the model using the training data, optimize the parameters on validation data, and compare model performance using the out-of-sample testing strategy on testing data. The experiment was repeated 10 times and we report the average performance along with the standard deviations.

Results

Performance comparison of readmission predictions

Table 2 compares the prediction performance of the proposed model with several state-of-the-art baselines. The proposed CONTENT model outperforms baselines on all metrics. It is due to CONTENT incorporates both local and global contextual information, especially for the diseases that have heterogeneous manifestations such as CHF. The GRU+word2vec predicts

Table 2. Performance comparison on CHF data. CONTENT outperforms Word2vec+LR, Med2vec+LR, GRU, GRU+Word2Vec, and RETAIN on different performance metrics.

Method	PR-AUC	ROC-AUC	ACC
Word2vec+LR	0.3445±0.0204	0.5360±0.0246	0.6828±0.0120
Med2vec+LR	0.3836±0.0149	0.5937±0.0120	0.6915±0.0095
GRU	0.3862±0.0136	0.5998±0.0124	0.6856±0.0082
GRU+Word2Vec	0.3430±0.0157	0.5616±0.0157	0.6731±0.0091
RETAIN	0.3720±0.0148	0.5707±0.0140	0.6814±0.0111
CONTENT	0.3894±0.0153	0.6103±0.0130	0.6934±0.0090

https://doi.org/10.1371/journal.pone.0195024.t002

worse than the basic GRU model. This may be due to low-dimensional concept embedding via word2vec blurs the boundary of some heterogeneous subtypes and thus causes wrong predictions. In addition, the RETAIN [34] also predicts worse than a basic GRU model. Although RETAIN adopts a sophisticated attention mechanism to set more weights on events that are considered more important, their attention strategy is not relevant to the prediction task (e.g. readmission prediction) since the attention weights in [34] are generated from only the hidden states of the GRU, while the task-related context could be ignored by this model. For Med2vec [35], we did not use the demographic information in the original paper in order to keep a fair comparison. The Med2vec takes advantage of the hierarchical information of the EHR data, and thus is a better representation method than word2vec and gaining better results.

In Table 3 we compare the prediction performance based on a set of synthetic data generated from a real generic patient cohort. During data generation, for each raw sequence of events, we dropped randomly sampled 30%−50% events, perturbed the time information for each visits, combined it with another subsampled sequence of events. The generation procedure effectively introduced lots of missing information, noise and anomaly. Results show that the proposed CONTENT model again outperforms most baselines due to it models patient representations and predict readmissions not only based on the RNN states but also on the topics. As the topics are exchangeable and globally modeled as a context, the CONTENT would be less impacted by some missing visits, noise and perturbed time information. However, when comparing with RETAIN, the proposed model gains much better PR-AUC since the precision is much higher, but slightly worse ROC-AUC since the attention model in RETAIN improves prediction accuracy in general.

Clustering of patient patterns

As we explained in the model inference section, to gain understanding of the learned patient representations, we concatenate the topical context vector $\boldsymbol{\theta}$ and the final hidden state of RNN as the patient-specific vector representation. These vectors are then used to cluster the CHF

Table 3. Performance comparison on synthetic data. CONTENT outperforms Word2vec+LR, Med2vec+LR, GRU, GRU+Word2Vec, and RETAIN on different performance metrics.

Method	PR-AUC	ROC-AUC	ACC
Word2vec+LR	0.5155±0.0021	0.6040±0.0188	0.6229±0.0179
Med2vec+LR	0.5906±0.0057	0.6884±0.0044	0.7170±0.0087
GRU	0.5929±0.0100	0.6881±0.0048	0.7141±0.0040
GRU+Word2Vec	0.5907±0.0174	0.6836±0.0031	0.7117±0.0045
RETAIN	0.5525±0.0005	0.6927±0.0001	0.7310±0.0001
CONTENT	0.6011±0.0191	0.6886±0.0074	0.7170±0.0069

https://doi.org/10.1371/journal.pone.0195024.t003

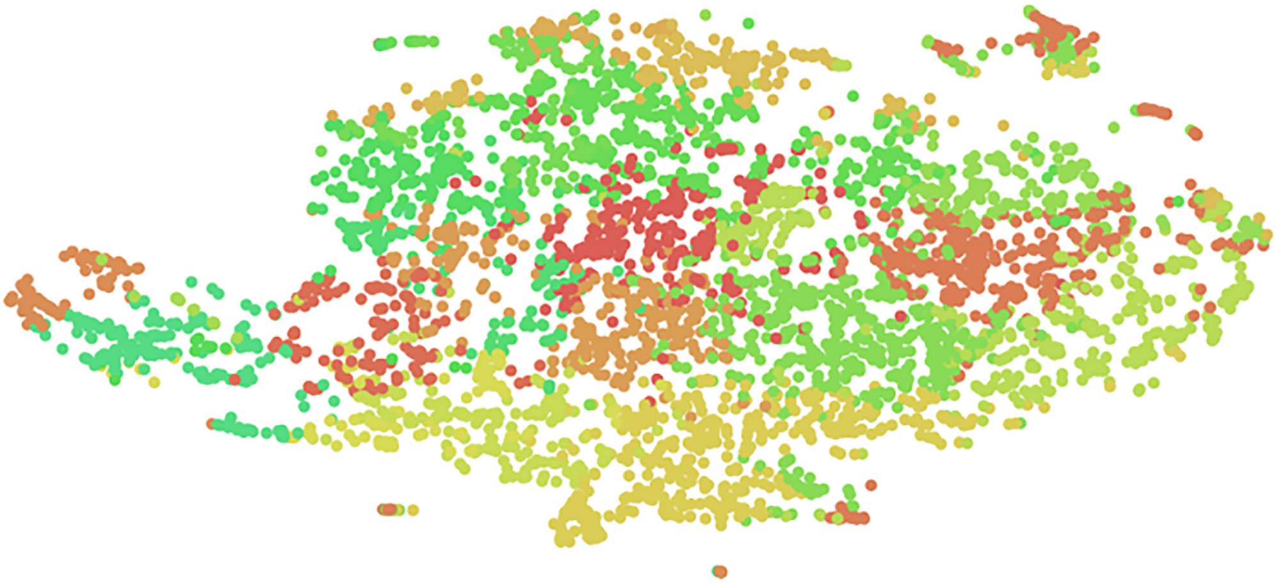


Fig 3. Clustering of patient representations.

https://doi.org/10.1371/journal.pone.0195024.g003

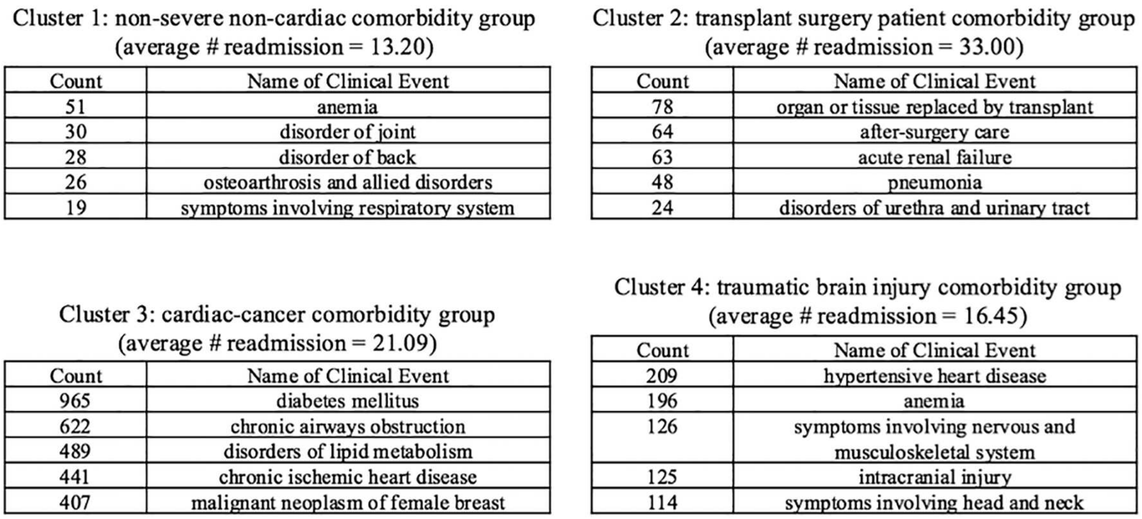
cohort into patient subgroups with more homogeneous latent patterns. To be specific, we apply k-means algorithm and set k = 20 to generate 20 subgroups. The clustering result is plotted in Fig 3.

To take a closer look at the learned subgroups, we pick 4 clusters out of the 20 clusters. To quantitatively evaluate their differences, we calculated the average number of readmission for each cluster. In addition, we also make qualitative evaluation by analyzing the top clinical events ranked by their counts in the cluster. Note that we omit the top three common events shared by all CHF patients, including 1) essential hypertension: a major risk factor of CHF, 2) cardiac dysrhythmia: a condition about irregular heart rhythm or abnormal heart rate, and if long-term impending, could indicate higher likelihood of CHF-related hospital readmission, and 3) heart failure. These events demonstrate the commonalities of the CHF condition manifestations. We omit them in order to focus more on the cluster-specific clinical patterns. The results are listed in Fig 4.

Combine the top clinical patterns and the average count of readmissions, we find that the clusters may represent different CHF comorbidity subgroups where comorbidity conditions serve as context and would impact risks of readmissions. We studied literature and derived the most likely interpretations for the clusters as explained below.

Cluster 1 (in orange) probably relates to the comorbidity group where patients have non-severe non-cardiac conditions. For example, the top event anemia is known to be a common condition among the non-cardiac comorbidity group, with a prevalence ranging from 4% to 55% [36]. In addition, other top events, e.g. the disorders of back, disorders of joint, and osteoarthritismainly occur among senior people in the non-cardiac CHF comorbidity group as discussed in [37]. Moreover, the concomitant symptoms affecting respiratory system were also discussed in literature and considered very common to CHF patients [37]. Due to these comorbidity conditions are not critical, under such context this cluster has low readmission on average.

As a contrast, most of the top conditions in Cluster 3 (in green) are cardiac comorbidities that directly relate to the presence of CHF. In addition, many patients in this cluster were



Cluster 1: non-severe non-cardiac comorbidity group (average # readmission = 13.20)

Count	Name of Clinical Event
51	anemia
30	disorder of joint
28	disorder of back
26	osteoarthrosis and allied disorders
19	symptoms involving respiratory system

Cluster 2: transplant surgery patient comorbidity group (average # readmission = 33.00)

Count	Name of Clinical Event
78	organ or tissue replaced by transplant
64	after-surgery care
63	acute renal failure
48	pneumonia
24	disorders of urethra and urinary tract

Cluster 3: cardiac-cancer comorbidity group (average # readmission = 21.09)

Count	Name of Clinical Event
965	diabetes mellitus
622	chronic airways obstruction
489	disorders of lipid metabolism
441	chronic ischemic heart disease
407	malignant neoplasm of female breast

Cluster 4: traumatic brain injury comorbidity group (average # readmission = 16.45)

Count	Name of Clinical Event
209	hypertensive heart disease
196	anemia
126	symptoms involving nervous and musculoskeletal system
125	intracranial injury
114	symptoms involving head and neck

Fig 4. Top clinical events for selected clusters.

https://doi.org/10.1371/journal.pone.0195024.g004

diagnosed as "malignant neoplasm of female breast (breast cancer)". Literature indicates the comorbid of CHF would become a risk factor for poor outcomes for breast cancer, adversely impact the cancer treatments [38], and thus could lead to more frequent hospital readmissions.

In addition, we find Cluster 4 (in yellow) quite interesting as many CHF patients have the following events "intracranial injury" and "symptoms involving head and neck". We suspect that for this group, the readmission might be due to the injuries rather than CHF itself. While the injuries could also be caused due to CHF related conditions, for example, the comorbid vision problem (e.g. cataracts) of CHF, the comorbid hypertensive heart disease, or the prevalence of various neurological disorders such as the event "symptoms involving nervous and musculoskeletal system" indicates [39].

Last, we also find Cluster 2 (in red) quite special as many patients have received organ or tissue transplant surgeries. It is reasonable to believe transplant surgeries relate to high risks of readmission for CHF patients. For CHF patients received cardiac transplantation, they were often considered at advanced stage with severe dysfunctions. This can be inferred from contaminant acute diseases, such as pneumonia and acute renal failure, as well as disorders of urethra and urinary tract, a common after surgery disorder. Moreover, CHF could be onset after transplant surgeries, e.g. the comorbid CHF after hematopoietic cell transplantation [40].

Discussion

This study presents CONTENT, a deep model that learns distributed patient representation from the EHR data and performs prediction for the 30-day readmissions. The CONTENT incorporates global context by capturing latent topics via a recognition network and uses global context as additional bias to the output layer of an RNN model, so that the RNN can focus its modeling capacity on the local context. With this design, the proposed model achieves more accurate prediction results than the state-of-the-art baselines.

The importance of learning both local and global context from analyzing the learned clusters are two-fold: 1) although CHF patients share many commonalities, e.g. having hypertension and cardiac dysrhythmia, their risks of readmission can vary due to different local or global contextual information. For example, patients recently received heart or cell transplantation surgery have high risks of readmission, thus need more frequent follow-ups after last discharge. Another example, if patients only have non-severe non-cardiac comorbid conditions, their risks of readmission would be lower than other groups. 2) although the readmission prediction is based on CHF cohort, the patients may be readmitted due to some other reasons, which could become confounding factors here. For example, patients are readmitted due to comorbid traumatic brain injury, which can be induced by CHF comorbid conditions or other reasons.

What is worth mentioning is in this work we did not explicitly perform any missing value imputation for the input clinical events. The observation that time intervals between two visits are irregular is a very common phenomenon in healthcare and clinical setting. Since most EHR data are not missing at random (NMAR) [41–43], it is challenging for existing imputation methods to be used on EHR data. Our CONTENT model not explicitly aims for solving the problem of missing values. However, it does implicitly decrease the impact brought by missing values by capturing the global context using topics. In future work we can also employ dropout techniques in the model. The dropout technique is essentially equivalent to randomly removing some visits or codes. So the final model will be more robust to missing visits.

To summarize, the CONTENT model not only learns more accurate patient representation and thus leads to better prediction performance, but also generates interpretable representations that could be used to cluster patients into more homogeneous patient subgroups. Analyzing the top clinical features in each subgroup provide interpretability and help us gain better understanding of CHF comorbidity and various reasons and risks of 30-day readmissions for CHF patients.

The limitation of this work is that we only use clinical events as original input features. However, unlike previous models that are designed to only take sequence of events as input features, the proposed model can be extended to extract better global context from generic form of inputs, e.g. patient profile or other side information. This will be one future direction of extension.

Conclusion

In this paper, we propose CONTENT, an end-to-end deep sequential predictive model that embeds local and global contextual information via RNN and a topic model based recognition network, respectively. We evaluated the model with hospital readmission prediction task on a cohort of CHF patients. CONTENT outperforms baseline and can also explicitly generates interpretable subgroups to improve understanding of heterogeneous readmission risks among CHF cohort. Future work includes applying the model to different cohorts to show the generality of the approach. We will also include side information for generating better global context.

Supporting information

S1 Data. The simulation data used in the experiment. S1_Data.txt is the simulation data used in the experiment and the implementation code can be found at https://github.com/danicaxiao/CONTENT.
(ZIP)

Acknowledgments

The work of Fei Wang is supported by NSF IIS-1650723 and IIS-1716432.

Author Contributions

Conceptualization: Cao Xiao, Adji B. Dieng, David M. Blei, Fei Wang.

Data curation: Cao Xiao.

Funding acquisition: Fei Wang.

Investigation: Cao Xiao, Tengfei Ma, Fei Wang.

Methodology: Cao Xiao, Tengfei Ma, Adji B. Dieng, Fei Wang.

Project administration: Fei Wang.

Resources: Fei Wang.

Supervision: David M. Blei, Fei Wang.

Validation: Cao Xiao.

Writing – original draft: Cao Xiao, Tengfei Ma.

Writing – review & editing: Fei Wang.

References

1. Basu Roy S, Teredesai A, Zolfaghar K, Liu R, Hazel D, Newman S, et al. Dynamic Hierarchical Classification for Patient Risk-of-Readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD'15. New York, NY, USA: ACM; 2015. p. 1691–1700. Available from: http://doi.acm.org/10.1145/2783258.2788585.
2. McIlvennan C, Eapen Z, Allen L. Hospital Readmissions Reduction Program. Circulation. 2015; 131 (20). https://doi.org/10.1161/CIRCULATIONAHA.114.010270 PMID: 25986448
3. K Z. Predicting Risk-of-Readmission for Congestive Heart Failure Patients: A Multi-Layer Approach. IEEE Trans on Big Data. 2013;.
4. Mathias J, Agrawal A, Feinglass J, Cooper A, Baker D, Choudhary A. Development of a 5 year life expectancy index in older adults using predictive mining of electronic health record data. Journal of American Medical Informatics Association. 2013; 20(e1). https://doi.org/10.1136/amiajnl-2012-001360
5. Tran T, Luo W, Phung D, Gupta S, Rana S, Kennedy RL, et al. A framework for feature extraction from hospital medical data with applications in risk prediction. BMC Bioinformatics. 2014; 15(1):425. https://doi.org/10.1186/s12859-014-0425-8 PMID: 25547173
6. Hammill B, Curtis L, Fonarow G, Heidenreich P, Yancy C, Peterson E, et al. Incremental value of clinical data beyond claims data in predicting 30-day outcomes after heart failure hospitalization. Circ Cardiovasc Qual Outcomes. 2011; 4(1). https://doi.org/10.1161/CIRCOUTCOMES.110.954693
7. Hripcsak G, Albers D. Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association. 2013; 20:117–121. https://doi.org/10.1136/amiajnl-2012-001145 PMID: 22955496
8. B C, J W, S W, et al. Systematic review: Impact of health information technology on quality, efficiency, and costs of medical care. Annals of Internal Medicine. 2006; 144(10):742–752. https://doi.org/10.7326/0003-4819-144-10-200605160-00125
9. Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence. 2013; 35(8):1798–1828. https://doi.org/10.1109/TPAMI.2013.50 PMID: 23787338
10. Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016.
11. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521(Jan):436–444. https://doi.org/10.1038/nature14539 PMID: 26017442
12. Bengio Y, Ducharme R, Vincent P, Janvin C. A Neural Probabilistic Language Model. J Mach Learn Res. 2003; 3:1137–1155.
13. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of machine Learning research. 2003; 3 (Jan):993–1022.
14. Elman JL. Finding structure in time. COGNITIVE SCIENCE. 1990; 14(2):179–211. https://doi.org/10.1207/s15516709cog1402_1

15. Nguyen P, Tran T, Wickramasinghe N, Venkatesh S. Deepr: A Convolutional Net for Medical Records. IEEE journal of biomedical and health informatics. 2017; 21 1:22–30. https://doi.org/10.1109/JBHI.2016.2633963 PMID: 27913366

16. Cheng Y, Wang F, Zhang P, Hu J. Risk Prediction with Electronic Health Records: A Deep Learning Approach. In: Proceedings of the 2016 SIAM International Conference on Data Mining, Miami, Florida, USA, May 5-7, 2016; 2016. p. 432–440. Available from: http://dx.doi.org/10.1137/1.9781611974348.49.

17. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. Journal of the American Medical Informatics Association. 2016; p. ocw112. https://doi.org/10.1093/jamia/ocw112

18. Baytas IM, Xiao C, Zhang X, Wang F, Jain AK, Zhou J. Patient Subtyping via Time-Aware LSTM Networks. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD'17; 2017. p. 65–74.

19. Che C, Xiao C, Liang J, Jin B, Zhou J, Wang F. An RNN Architecture with Dynamic Temporal Matching for Personalized Predictions of Parkinson's Disease. In: SIAM International Conference on Data Mining; 2017.

20. Ma T, Xiao C, Wang F. Health-ATM: A Deep Architecture for Multifaceted Patient Health Record Representation and Risk Prediction. In: SIAM International Conference on Data Mining; 2018.

21. Cho K, van Merriënboer B, Gülçehre Ç, Bahdanau D, Bougares F, Schwenk H, et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics; 2014. p. 1724–1734. Available from: http://www.aclweb.org/anthology/D14-1179.

22. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput. 1997; 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735 PMID: 9377276

23. Pascanu R, Mikolov T, Bengio Y. On the Difficulty of Training Recurrent Neural Networks. In: Proceedings of the 30th International Conference on International Conference on Machine Learning—Volume 28. ICML'13. JMLR.org; 2013. p. III–1310–III–1318. Available from: http://dl.acm.org/citation.cfm?id=3042817.3043083.

24. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Mach Learn Res. 2014; 15(1):1929–1958.

25. Mikolov T, Zweig G. Context Dependent Recurrent Neural Network Language Model; 2012.

26. Dieng AB, Wang C, Gao J, Paisley J. TopicRNN: A Recurrent Neural Network with Long-Range Semantic Dependency. International Conference On Learning Representations. 2017;.

27. Arisoy E, Saraclar B, Roark B, Shafran I. Discriminative language modeling with linguistic and statistically derived features. Audio, Speech, and Language Processing, IEEE Transactions on. 2012; 20 (2):540–550.

28. M P, P B, H R, et al. Bnp-guided vs symptom-guided heart failure therapy: The trial of intensified vs standard medical therapy in elderly patients with congestive heart failure (time-chf) randomized trial. JAMA. 2009; 301(4):383–392. https://doi.org/10.1001/jama.2009.2

29. Wu J, Roy J, Stewart W. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med Care. 2010; 48(6).

30. Kingma DP, Welling M. Auto-Encoding Variational Bayes. CoRR. 2013;abs/1312.6114.

31. Blei DM, Kucukelbir A, McAuliffe JD. Variational Inference: A Review for Statisticians. Journal of the American Statistical Association. 2017; 112(518):859–877. https://doi.org/10.1080/01621459.2017.1285773

32. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. CoRR. 2014;abs/1412.6980.

33. Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. arXiv e-prints. 2016;abs/1605.02688.

34. Choi E, Bahadori MT, Sun J, Kulas J, Schuetz A, Stewart W. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. In: Advances in Neural Information Processing Systems; 2016. p. 3504–3512.

35. Choi E, Bahadori MT, Searles E, Coffey C, Sun J. Multi-layer Representation Learning for Medical Concepts. arXiv preprint arXiv:160205568. 2016;.

36. Katz SD. Mechanisms and Treatment of Anemia in Chronic Heart Failure. Congestive Heart Failure. 2004; 10(5):243–247. https://doi.org/10.1111/j.1527-5299.2004.03298.x PMID: 15470302

37. Lang CC, Mancini DM. Non-cardiac comorbidities in chronic heart failure. Heart. 2007; 93(6):665–671. https://doi.org/10.1136/hrt.2005.068296 PMID: 16488925

38. Robert I, Griffiths ML, Gleeson JMV, Danese MD. Impact of Undetected Comorbidity on Treatment and Outcomes of Breast Cancer. International Journal of Breast Cancer. 2014;.

39. Thompson H, Dikmen S, Temkin N. Prevalence of Comorbidity and its Association with Traumatic Brain Injury and Outcomes in Older Adults. Research in gerontological nursing. Research in gerontological nursing. 2012; 5(1):17–24. https://doi.org/10.3928/19404921-20111206-02 PMID: 22165997

40. Armenian S, Sun C, Francisco L. Late Congestive Heart Failure After Hematopoietic Cell Transplantation. Journal of Clinical Oncology. 2008; 26(34):5537–5543. https://doi.org/10.1200/JCO.2008.17.7428 PMID: 18809605

41. Lin JH, Haug P. Exploiting Missing Clinical Data in Bayesian Network Modeling for Predicting Medical Problems. Journal of Biomedical Informatics. 2008; 41(1):1–14. https://doi.org/10.1016/j.jbi.2007.06.001 PMID: 17625974

42. Pivovarov R, Albers DJ, Sepulveda JL, Elhadad N. Identifying and Mitigating Biases in EHR Laboratory Tests.” Journal of Biomedical Informatics. Journal of Biomedical Informatics. 2014; 51:24–34. https://doi.org/10.1016/j.jbi.2014.03.016 PMID: 24727481

43. Little RJA, Rubin DB. Statistical Analysis with Missing Data. John Wiley & Sons.; 2014.