

# DeepCare: A Deep Dynamic Memory Model for Predictive Medicine

Trang Pham, Truyen Tran, Dinh Phung and Svetha Venkatesh

February 2, 2016

## Abstract

Personalized predictive medicine necessitates the modeling of patient illness and care processes, which inherently have long-term temporal dependencies. Healthcare observations, recorded in electronic medical records, are episodic and irregular in time. We introduce DeepCare, an *end-to-end* deep dynamic neural network that reads medical records, stores previous illness history, infers current illness states and predicts future medical outcomes. At the data level, DeepCare represents care episodes as vectors in space, models patient health state trajectories through explicit memory of historical records. Built on Long Short-Term Memory (LSTM), DeepCare introduces time parameterizations to handle irregular timed events by moderating the forgetting and consolidation of memory cells. DeepCare also incorporates medical interventions that change the course of illness and shape future medical risk. Moving up to the health state level, historical and present health states are then aggregated through multiscale temporal pooling, before passing through a neural network that estimates future outcomes. We demonstrate the efficacy of DeepCare for disease progression modeling, intervention recommendation, and future risk prediction. On two important cohorts with heavy social and economic burden – diabetes and mental health – the results show improved modeling and risk prediction accuracy.

## 1 Introduction

When a patient is admitted to hospital, there are two commonly asked questions: “what is happening?” and “what happens next?” The first question refers to the diagnosis of the illness, the second is about prediction of future medical risk [1]. While there are a wide array of diagnostic tools to answer the first question, the technologies are much less advanced in answering the second [2]. Traditionally, this prognostic question may be answered by experienced clinicians who have seen many patients, or by clinical prediction models with well-defined and rigorously collected risk factors. But this is expensive and of limited availability. Modern electronic medical records (EMRs) promise to offer a fast and cheap alternative. An EMR typically contains the history of hospital encounters, diagnoses and interventions, lab tests and clinical narratives. The wide adoption of EMRs has led to intensified research in building predictive models from this rich data source in the past few years [3, 4, 5].

Answering to prognostic inquiries necessitates modeling patient-level temporal healthcare processes. An effective modeling must address four open challenges: (i) *Long-term dependencies in healthcare*: the future illness and care may depend critically on historical illness and interventions. For example, the onset of diabetes at middle age remains a risk factor for the rest of the life; cancers may recur after years; and a previous surgery may prevent certain future interventions. (ii) *Representation*

*of admission*: an admission episode consists of a variable-size discrete set containing diagnoses and interventions. (iii) *Episodic recording and irregular timing*: medical records vary greatly in length, are inherently episodic in nature and irregular in time [6]. The data is episodic because it is only recorded when the patient visits hospital and is undergone an episode of care. The episode is often tightly packed in a short period, typically ranging from a day to two weeks. The timing of arrivals is largely random. (iv) *Confounding interactions between disease progression and intervention*: medical records are a mixture of the course of illness, the developmental and the intervening processes. In addition to addressing these four challenges, a predictive system should be *end-to-end* and *generic* so that it can be deployed on different hospital implementations of EMRs. An end-to-end system requires minimal or no feature engineering to read medical records, infer present illness states and predict future outcomes.

Existing methods are poor in handling such complexity. They inadequately capture variable length [4] and ignore the long-term dependencies [7, 8]. Temporal models based on Markovian assumption are unable to model temporal irregularity and have no memory, and thus can completely forget previous major illness given an irrelevant episode [9]. Deep learning, which has recently revolutionized cognitive fields such as speech recognition, vision and computational linguistics, holds a great potential in constructing end-to-end systems [10]. However, its promise to healthcare has not been realized [11, 12, 13].

To this end, we introduce DeepCare, an *end-to-end* deep dynamic memory neural network that addresses the four challenges. DeepCare is built on Long Short-Term Memory (LSTM) [14, 15], a recurrent neural network equipped with *memory cells* to store experiences. At each time-step, the LSTM reads an input, updates the memory cell, and returns an output. Memory is maintained through a *forget gate* that moderates the passing of memory from one time step to another, and is updated by seeing new input at each time step. The output is determined by the memory and moderated by an *output gate*. In DeepCare, the LSTM models the illness trajectory and healthcare processes of a patient encapsulated in a time-stamped sequence of admissions. The inputs to the LSTM are information extracted from admissions. The outputs represent illness states at the time of admission. *Memory maintenance enables capturing of long-term dependencies*, thus addressing the first challenge. In fact, this capacity has made LSTM an ideal model for a variety of sequential domains [16, 15, 17]. No LSTM has been used in healthcare, however – one major difficulty would be the lack of handling of set inputs, irregular timing and interventions.

Addressing these three drawbacks, DeepCare modifies LSTM in several ways. For representing admission, which is a set of discrete elements such as diagnoses and interventions, the solution is to embed these elements into continuous vector spaces. Vectors of the same type are then pooled into a single vector. Type-specific pooled vectors are then concatenated to represent an admission. In that way, *variable-size admissions are embedded in to continuous distributed vector space*. The admission vectors then serve as input features for the LSTM. As the embedding is learnt from data, the model does not rely on manual feature engineering.

For *irregular timing*, the *forget gate* is extended to be a function of *irregular time gap* between consecutive time steps. We introduce two new forgetting mechanisms: monotonic decay and full time-parameterization. The decay mimics the natural forgetting when learning a new concept in human. The parameterization accounts for more complex dynamics of different diseases over time. The resulting model is sparse in time and efficient to compute since only observed records are incorporated, regardless of the irregular time spacing. Finally, in DeepCare *the confounding interaction between disease progression and interventions is modeled as follows*. Interventions influence the output gate of current illness states and the forget gate that moderates memory carried into the future. As a result, the illness states (the output) are moderated by past and current interventions.

Once illness states are outputted by the LSTM layer, they are aggregated through a new time-decayed multiscale pooling strategy. This allows further handling of time-modulated memory. Finally at the top layer, pooled illness states are passed through a neural network for estimating future prognosis. In short, computation steps in DeepCare can be summarized as

$$P(y | \mathbf{u}_{1:n}) = P(\text{nnet}_y(\text{pool}\{\text{LSTM}(\mathbf{u}_{1:n})\})) \quad (1)$$

where  $\mathbf{u}_{1:n}$  is the input sequence of admission observations,  $y$  is the outcome of interest (e.g., readmission),  $\text{nnet}_y$  denotes estimate of the neural network with respect to outcome  $y$ , and  $P$  is probabilistic model of outcomes. Overall, DeepCare is an end-to-end prediction model that relies on no manual feature engineering, is capable of reading generic medical records, memorizing a long history, inferring illness states and predicting the future risk.

We demonstrate our DeepCare on answering a crucial part of the holy grail question “what happens next?”. In particular, we demonstrate our model on *disease progression*, *intervention recommendation* and *future risk prediction*. Disease progression refers to the next disease occurrence given the medical history. Intervention recommendation is about predicting a subset of treatment procedures for the current diagnoses. Future risk may involve readmission or mortality within a predefined period after discharge. We note in passing that the forecasting of future events may be considerably harder than the traditional notion of classification (e.g., objects/documents categorization) due to inherent uncertainty in unseen interleaved events. Our experiments are demonstrated on two datasets of very different nature – diabetes (a well-defined chronic condition) and mental health (a diverse mixture of many acute and chronic conditions). The cohorts were collected from a large regional hospital in the period of 2002 to 2013. We show that DeepCare outperforms state-of-the-art baseline classification methods.

To summarize, through introducing DeepCare, we make four modeling contributions: (i) handling long-term dependencies in healthcare; (ii) a novel representation of variable-size admission as fixed-size continuous vectors; (iii) modeling episodic recording and irregular timing; and (iv) capturing confounding interactions between disease progression and intervention. We also contribute to the healthcare analytics practice by demonstrating the effectiveness of DeepCare on disease progression, intervention recommendation and medical risk prediction. Finally, we wish to emphasize that although DeepCare is designed as predictive model targeted to healthcare, DeepCare can be applied to other temporal domains with similar data characteristics (i.e., long-term dependencies, discrete set inputs, irregular timing and confounding interventions).

The paper is organized as follows. Section 2 provides background for Electronic Medical Records, sequential and deep learning for healthcare. Section 3 presents preliminaries for DeepCare model: Recurrent neural networks, LSTM and learning word representation. DeepCare is described in Section 4 while the experiments and results are reported in Section 5. Finally, Section 6 discusses further and concludes the paper.

## 2 Background

### 2.1 Electronic medical records (EMRs)

An electronic medical record (EMR) is a digital version of patients health information. A wide range of information can be stored in EMRs, such as detailed records of symptoms, data from monitoring devices, clinicians’ observations [18]. EMR systems store data accurately, decrease the risk of data replication and the risk of data lost. EMRs are now widely adopted in developed countries and are

increasingly present in the rest of the world. It is expected that EMRs in hospital help improve treatment quality and reduce healthcare costs [19].

A typical EMR contains information about a sequence of admissions for a patient. There are two types of admission methods: planned (routine) and unplanned (emergency). Unplanned admission refers to transfer from the emergency department. EMRs typically store admitted time, discharge time, lab tests, diagnoses, procedures, medications and clinical narratives. Diagnoses, procedures and medications stored in EMRs are typically coded in standardized formats. Diagnoses are represented using WHO’s ICD (International Classification of Diseases) coding schemes<sup>1</sup>. For example, E10 encodes Type 1 diabetes mellitus, E11 encodes Type 2 diabetes mellitus while F32 indicates depressive episode. The procedures are typically coded in CPT (Current Procedural Terminology) or ICHI (International Classification of Health Interventions) schemes<sup>2</sup>. Medication names can be mapped into the ATC (Anatomical Therapeutic Chemical) scheme<sup>3</sup>.

The wide adoption of EMRs has led to calls for meaningful use [3, 20]. One of the most important uses is building predictive models [3, 21, 6, 4, 5]. Like most applications of machine learning, the bottleneck here is manual feature engineering due to the complexity of the data [22][21]. Our DeepCare solves this problem by building an end-to-end system where features are learnt automatically from data.

## 2.2 Sequential models for healthcare

Although healthcare is inherently episodic in nature, it has been well-recognized that modeling the entire illness trajectory is important [23][24]. Nursing illness trajectory model was popularized by Strauss and Corbin [25], but the model is qualitative and imprecise in time [26]. Thus its predictive power is very limited.

Electronic medical records (EMRs) offer the quantitative alternative with precise timing of events. However, EMRs are complex – they reflect the interleaving between the illness processes and care processes. The timing is irregular – patients only visit hospital when the illness is beyond a certain threshold, even though the illness may have been present long before the visit. Existing work that handles such irregularities includes interval-based extraction [4], but this method is rather coarse and does not explicitly model the illness dynamics.

Capturing disease progression has been of great interest [27, 28], and much effort has been spent on Markov models [7, 29] and dynamic Bayesian networks [30]. However, healthcare is inherently non-Markovian due to the long-term dependencies. For example, a routine admission with irrelevant medical information would destroy the effect of severe illness [9], especially for chronic conditions. Irregular timing and interventions have not been adequately modeled. Irregular-time Bayesian networks [31] offer a promise, but its power has yet to be demonstrated. Further, assuming discrete states are inefficient since the information pathway has only  $\log(K)$  bits for  $K$  states. Our work assumes distributed and continuous states, thus offering much larger state space.

## 2.3 Deep learning for healthcare

Deep learning is currently at the center of a new revolution in making sense of a large volume of data. It has achieved great successes in cognitive domains such as speech, vision and NLP [10]. To date, deep learning approach to healthcare has been an unrealized promise, except for several very recent works

---

<sup>1</sup><http://apps.who.int/classifications/icd10/browse/2016/en>

<sup>2</sup><http://www.who.int/classifications/ichi/en/>

<sup>3</sup>[http://www.whocc.no/atc\\_ddd\\_index/](http://www.whocc.no/atc_ddd_index/)

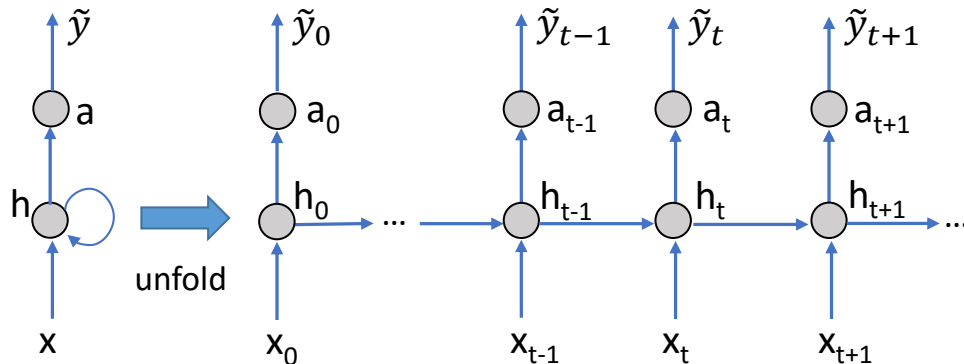


Figure 1: (Left) A typical Recurrent Neural Network and (Right) an RNN unfolded in time. Each RNN unit at time step  $t$  reads input  $\mathbf{x}_t$  and previous hidden state  $\mathbf{h}_{t-1}$ , generates output  $\mathbf{a}_t$  and predicts the label  $\tilde{y}_t$ .

[11, 12], where irregular timing is not property modeled. We observe that is a considerable similarity between NLP and EMR, where diagnoses and interventions play the role of nouns and modifiers, and an EMR is akin to a sentence. A major difference is the presence of precise timing in EMR, as well as the episodic nature. This suggests that it is possible to extend NLP language models to EMR, provided that irregular timing and episodicity are properly handled. Our DeepCare contributes along that line. Going down to the genetic basis of health, a recent work called DeepFind [32] uses convolutional networks to detect regular DNA/RNA motifs. This is unlike DeepCare, where irregular temporal dynamics are modeled.

### 3 Preliminaries

In this section, we briefly review building blocks for DeepCare, which will be described fully in Sec. 4.

#### 3.1 Recurrent neural network

A Recurrent Neural Network (RNN) is a neural network repeated over time. In particular, an RNN allows self-loop connections and shared parameters across different time steps. While a feedforward neural network maps an input vector into an output vector, an RNN maps a sequence into a sequence. Unlike hidden Markov models, where the states are typically discrete and the transitions between states are stochastic, RNNs maintain distributed continuous states with deterministic dynamics. The recurrent connections allow an RNN to memorize previous inputs, and therefore capture longer dependencies than a hidden Markov model does. Since the first version of RNN was introduced in the 1980s [33], many varieties of RNN have been proposed such as Time-Delay Neural Networks [34] and Echo State Network [35]. Here we restrict our discussion to the simple RNN with a single hidden layer as shown in Fig. 1.

## Forward propagation

An RNN unit has three connections: a recurrent connection from the previous hidden state to the current hidden state ( $\mathbf{h}_{t-1} \rightarrow \mathbf{h}_t$ ), an input-to-hidden-state connection ( $\mathbf{x}_t \rightarrow \mathbf{h}_t$ ) and a hidden-state-to-output connection ( $\mathbf{h}_t \rightarrow \mathbf{a}_t$ ). At time step  $t$ , the model reads the input  $\mathbf{x}_t \in \mathbb{R}^M$  and previous hidden state  $\mathbf{h}_{t-1} \in \mathbb{R}^K$  to compute the hidden state  $\mathbf{h}_t$  (Eq. 2). Thus  $\mathbf{h}_t$  summarizes information from all previous inputs  $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_t$ . The output  $\mathbf{a}_t \in \mathbb{R}^k$  (Eq. 3) is generated by a transformation function of  $\mathbf{h}_t$ , where  $k$  is the number of classes in the classification tasks. To predict the label  $\tilde{y}_t$ ,  $\mathbf{a}_t$  is then passed through a probabilistic function  $f_{prob}$  to compute the vector of probabilities  $\mathbf{P} = [P(\tilde{y}_t = 0 | \mathbf{x}_t, \dots, \mathbf{x}_0), \dots, P(\tilde{y}_t = k-1 | \mathbf{x}_t, \dots, \mathbf{x}_0)]$  (Eq. 4), where  $P(0 | \mathbf{x}_t, \dots, \mathbf{x}_0), \dots, P(k-1 | \mathbf{x}_t, \dots, \mathbf{x}_0) \geq 0$  and  $P(0 | \mathbf{x}_t, \dots, \mathbf{x}_0) + \dots + P(k-1 | \mathbf{x}_t, \dots, \mathbf{x}_0) = 1$ . Denote by  $a_t^i$  the element  $i^{th}$  of the vector  $\mathbf{a}_t$ . For two classes,  $f_{prob}$  is normally a logistic sigmoid function:

$$P(\tilde{y}_t = 1 | \mathbf{x}_t, \dots, \mathbf{x}_0) = \text{sigmoid}(a_t^1) = \frac{1}{1 + e^{-a_t^1}}$$

and for multiple classes,  $f_{prob}$  is a softmax function:

$$P(\tilde{y}_t = i | \mathbf{x}_t, \dots, \mathbf{x}_0) = \text{softmax}(a_t^i) = \frac{e^{a_t^i}}{\sum_j e^{a_t^j}}$$

for  $i = 0, \dots, k-1$ .

The weighted matrices  $W \in \mathbb{R}^{K \times M}$ ,  $U \in \mathbb{R}^{K \times K}$  and  $V \in \mathbb{R}^{k \times K}$  and bias vectors  $\mathbf{b}$  and  $\mathbf{c}$  are shared among all time steps. This allows the model to learn with varied length sequences and produce an output at each time step as follows:

$$\mathbf{h}_t = \tanh(\mathbf{b} + W\mathbf{h}_{t-1} + U\mathbf{x}_t) \quad (2)$$

$$\mathbf{a}_t = \mathbf{c} + V\mathbf{h}_t \quad (3)$$

$$P(\tilde{y}_t) = f_{prob}(\mathbf{a}_t) \quad (4)$$

At step 0, there is no previous hidden state,  $\mathbf{h}_0$  is computed as  $\tanh(\mathbf{b} + U\mathbf{x}_0)$ .

The total loss for a sequence  $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_n$  and its corresponding labels  $y_0, y_1, \dots, y_n$ , where  $y_0, y_1, \dots, y_n \in [0, 1, \dots, k-1]$ , would be the sum of the losses over all time steps:

$$L(y | \mathbf{x}) = \sum_{t=0}^n L_t(\tilde{y}_t = y_t | \mathbf{x}_t \dots \mathbf{x}_0) = - \sum_{t=0}^n \log P(\tilde{y}_t = y_t)$$

## Back-propagation

RNNs can be trained to minimize the loss function using gradient descent. The derivatives with respect to the parameters can be determined by the Back-Propagation Through Time algorithm [36]. This algorithm obtains the gradients by the chain rule like the standard back-propagation.

## Challenge of long-term dependencies

Many experiments have shown that gradient based learning algorithms face difficulties in training RNN. This is because the long term dependencies in long input sequences lead to vanishing or exploding

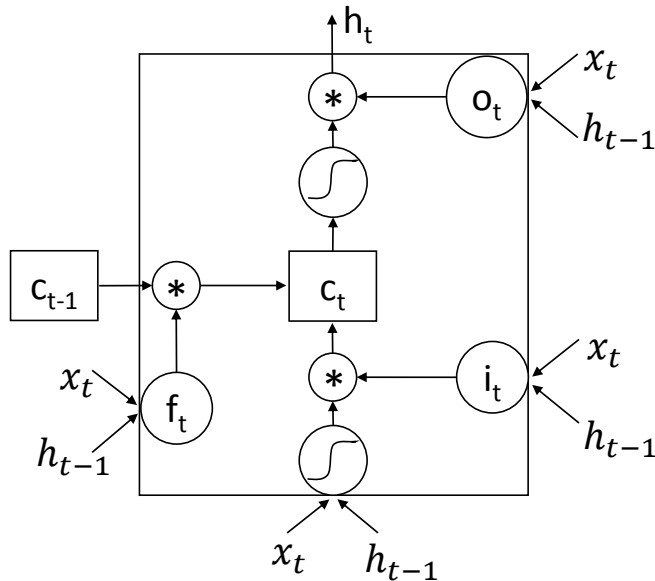


Figure 2: An LSTM unit that reads input  $\mathbf{x}_t$  and previous output state  $\mathbf{h}_{t-1}$  and produces current output state  $\mathbf{h}_t$ . An unit has a memory cell  $\mathbf{c}_t$ , an input gate  $\mathbf{i}_t$ , an output gate  $\mathbf{o}_t$  and a forget gate  $\mathbf{f}_t$ .

gradients [37, 38]. Many approaches have been proposed to solve the problem, such as Leaky Units [39], Nonlinear AutoRegressive models with eXogenous (NARX) [40] and Long-Short Term Memory (LSTM) [14]. Among them, LSTM has proved to be the most effective for handling very long sequences [14, 41], and thus will be chosen as a building block in our DeepCare.

### 3.2 Long-short term memory

This section reviews Long Short-Term Memory (LSTM) [14, 41], a modified version of RNN, to address the problem of long-term dependencies. Central to an LSTM is a linear self-loop memory cell which allows gradients to flow through long sequences. The memory cell is gated to moderate the amount of information flow into or from the cell. LSTMs have been significantly successful in many applications, such as machine translation [17], handwriting recognition [42] and speech recognition [43].

Fig. 2 describes an LSTM unit. Instead of a simple RNN unit, an LSTM unit has a memory cell that has state  $\mathbf{c}_t \in \mathbb{R}^K$  at time  $t$ . The information flowing through the memory cell is controlled by three gates: an input gate, a forget gate and an output gate. The input gate  $\mathbf{i}_t \in \mathbb{R}^K$  controls the input flowing into the cell, the forget gate  $\mathbf{f}_t \in \mathbb{R}^K$  controls the forgetting of the memory cell, and the output gate  $\mathbf{o}_t \in \mathbb{R}^K$  moderates the output flowing from the memory cell. Before describing detailed formulas, we denote the element-wise sigmoid function of a vector by  $\sigma$  and the element-wise product of two vectors by  $*$ .

The three gates are all sigmoidal units which set every element of the gates to a value between 0 and 1:

$$\mathbf{i}_t = \sigma(W_i \mathbf{x}_t + U_i \mathbf{h}_{t-1} + \mathbf{b}_i) \quad (5)$$

$$\mathbf{f}_t = \sigma(W_f \mathbf{x}_t + U_f \mathbf{h}_{t-1} + \mathbf{b}_f) \quad (6)$$

$$\mathbf{o}_t = \sigma(W_o \mathbf{x}_t + U_o \mathbf{h}_{t-1} + \mathbf{b}_o) \quad (7)$$

where  $W_{\{i,f,o\}}$ ,  $U_{\{i,f,o\}}$ ,  $\mathbf{b}_{\{i,f,o\}}$  are parameters. The gates control the amount of information passing through, from full when the gate value is 1, to complete blockage when the value is 0.

At each time step  $t$ , the input features are first computed by passing input  $\mathbf{x}_t \in \mathbb{R}^M$  and the previous hidden state  $\mathbf{h}_{t-1} \in \mathbb{R}^K$  through a squashing tanh function:

$$\mathbf{g}_t = \tanh(W_c \mathbf{x}_t + U_c \mathbf{h}_{t-1} + \mathbf{b}_c) \quad (8)$$

The memory cell is updated through partially forgetting the previous memory cell and reading the moderated input features as follows:

$$\mathbf{c}_t = \mathbf{f}_t * \mathbf{c}_{t-1} + \mathbf{i}_t * \mathbf{g}_t \quad (9)$$

The memory cell sequence is additive, and thus the gradient is also updated in a linear fashion through the chain rule. This effectively prevents the gradient from vanishing or exploding. The memory cell plays a crucial role in memorizing past experiences through the *learnable* forgetting gates  $\mathbf{f}_t$ . If  $\mathbf{f}_t \rightarrow \mathbf{1}$ , all the past memory is preserved, and new memory keeps updated with new inputs. If  $\mathbf{f}_t \rightarrow \mathbf{0}$ , only new experience is updated and the system becomes memoryless.

Finally, a hidden output state  $\mathbf{h}_t$  is computed based on the memory  $\mathbf{c}_t$ , gated by the output gate  $\mathbf{o}_t$  as follows:

$$\mathbf{h}_t = \mathbf{o}_t * \tanh(\mathbf{c}_t) \quad (10)$$

Note that since the system dynamic is deterministic,  $\mathbf{h}_t$  is a function of all previous input:  $\mathbf{h}_t = \text{LSTM}(\mathbf{x}_{1:t})$ . The output states are then used to generate outputs. We subsequently review two output types: sequence labeling and sequence classification.

### LSTM for sequence labeling

The output states  $\mathbf{h}_t$  can be used to generate labels at time  $t$  as follows:

$$P(y_t = l \mid \mathbf{x}_{1:t}) = \text{softmax}(\mathbf{v}_l^\top \mathbf{h}_t) \quad (11)$$

for label specific parameters  $\mathbf{v}_l$ .

### LSTM for sequence classification

LSTMs can be used for sequence classification using a simple mean-pooling strategy over all output states coupled with a differentiable loss function. For example, in the case of binary outcome  $y \in \{0, 1\}$ , we have:

$$P(y = 1 \mid \mathbf{x}_{1:n}) = \text{LR}(\text{pool}\{\text{LSTM}(\mathbf{x}_{1:n})\}) \quad (12)$$

where LR denotes probability estimate of the logistic regression, and  $\text{pool}\{\mathbf{h}_{1:n}\} = \frac{1}{n} \sum_{t=1}^n \mathbf{h}_t$ .



### 3.3 Learning word representation

We use “word” to refer to a discrete element within a larger context (e.g., a word in a document, or a diagnosis in an admission described in Sec. 4.2). Recall that input fed into many machine learning models is often represented as a fix-length feature vector. For text, bag-of-words are commonly used. A word  $w$  is represented by a one-hot vector  $\mathbf{v}_w \in \mathbb{R}^{|V|}$ , where  $\mathbf{v}_w = [v_w^1, \dots, v_w^{|V|}]$  and  $|V|$  is the number of words in the dictionary:  $\mathbf{v}_w = [0, \dots, 0, 1, 0, \dots, 0]$  ( $v_w^i = 1$  if  $w = i$ , which implies  $w$  is the word  $i^{\text{th}}$  in the dictionary, and  $v_w^i = 0$ , otherwise). Under bag-of-words representation, the vector of a sentence  $w_0, \dots, w_n$  is the sum of its word vectors:  $\mathbf{u} = \mathbf{v}_{w_0} + \mathbf{v}_{w_1} + \dots + \mathbf{v}_{w_n}$ . However, the bag-of-words method fails to capture ordering and semantic of the words [44].

A powerful alternative to bag-of-words is to embed words into continuous distributed representation in a vector space of  $M$  dimensions where  $M \ll |V|$  [45]. Every word is map to a unique vector which is a column in a matrix  $E \in \mathbb{R}^{M \times |V|}$ . There are several benefits for word embedding. First, the dimensionality is greatly reduced and does not depend on the appearance of new words. Second, the semantic of a word is represented in a distributed fashion, that is, there are multiple elements that encode the word meaning. Third, manipulation of continuous vectors is much easier with current algebraic tools such as addition and matrix multiplication, as evidenced in recent works [46]. For example, the similarity between two words is simply a cosine between two vectors. More importantly, the embedding matrix  $E$  can be learnt from data.

There are various approaches to learn the embedding matrix  $E$ . The most popular approach is perhaps Continuous Bag-of-Words model [46]. For a word  $w_i$  in a sequence of words, the model uses the words surrounding  $w_i$  to predict  $w_i$ . With an input context size of  $C$ ,  $w_{i-C}, \dots, w_{i-1}, w_{i+1}, \dots, w_{i+C}$  are called context words of  $w_i$ . All the context words are embedded into vectors using embedding matrix  $E$  and then averaged to get the mean vector  $\mathbf{h}$

$$\mathbf{h} = \frac{E^{w_{i-C}} + \dots + E^{w_{i-1}} + E^{w_{i+1}} + \dots + E^{w_{i+C}}}{2C}$$

where  $E^t$  is the column  $t^{\text{th}}$  of the matrix  $E$ . The model then generates the output  $\mathbf{a} = \bar{E}\mathbf{h}$ , where  $\bar{E} \in \mathbb{R}^{|V| \times M}$  and predict the center word  $w_i$  using softmax function

$$P(w_i | w_{i-C}, \dots, w_{i-1}, w_{i+1}, \dots, w_{i+C}) = \text{softmax}(\mathbf{a})$$

The parameters  $E$  and  $\bar{E}$  are learnt by minimizing the loss function

$$L = \frac{1}{T} \sum_{i=1}^T \log P(w_i | w_{i-C}, \dots, w_{i-1}, w_{i+1}, \dots, w_{i+C})$$

through back-propagation using stochastic gradient descent.

Another approach to learn the embedding matrix  $E$  is language modeling with an RNN [47]. More formally, given a sequence of words:  $w_0, w_1, \dots, w_t$ , the objective is maximizing the log probability  $\log P(w_{t+1} | w_t, \dots, w_1, w_0)$ . Each word  $w_i$  in the sequence is embedded into vector  $\mathbf{x}_i = E^{w_i}$  and the sequence  $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_t$  is the input of an RNN. The model only produces the output  $\mathbf{a}_t$  at the step  $t$  (See Sec. 3.1, Eq. 3) and predict the next word using a multiclass classifier with a softmax function

$$P(w_{t+1} | w_t, \dots, w_1, w_0) = \text{softmax}(\mathbf{a}_t)$$

The loss function is  $L = \frac{1}{T} \sum_{t=0}^{T-1} \log p(w_{t+1} | w_t, \dots, w_0)$ . The matrix  $E$  and all the parameters of the RNN model are learnt jointly through back-propagation using gradient descent.

## 4 DeepCare

In this section we present our main contribution named DeepCare for modeling illness trajectories and predicting future outcomes. DeepCare is built upon LSTM to exploit the ability to model long-term dependencies in sequences. We extend LSTM to address the three major challenges: (i) *variable-size discrete inputs*, (ii) *confounding interactions between disease progression and intervention*, and (iii) *irregular timing*.

### 4.1 Model overview

Recall from Sec. 2.1, there are two types of admission methods: planned and unplanned. Let  $m_t$  be the admission method at time step  $t$ , where  $m_t = 1$  indicates unplanned admission and  $m_t = 2$  indicates planned admission. Let  $\Delta t$  be the elapsed time between the current admission and its previous one.

As illustrated in Fig. 3, DeepCare is a deep dynamic neural network that has three main layers. The bottom layer is built on LSTM whose memory cells are modified to handle irregular timing and interventions. More specifically, the input is a sequence of admissions. Each admission  $t$  contains a set of diagnosis codes (which is then formulated as a feature vector  $\mathbf{x}_t \in \mathbb{R}^M$ ), a set of intervention codes (which is further formulated as a feature vector  $\mathbf{p}_t \in \mathbb{R}^M$ ), the admission method  $m_t \in \{1,2\}$  and the elapsed time  $\Delta t \in \mathbb{R}^+$ . Denote by  $\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_n$  the input sequence, where  $\mathbf{u}_t = [\mathbf{x}_t, \mathbf{p}_t, m_t, \Delta t]$ , the LSTM computes the corresponding sequence of distributed illness states  $\mathbf{h}_0, \mathbf{h}_1, \dots, \mathbf{h}_n$ , where  $\mathbf{h}_t \in \mathbb{R}^K$  (See Fig. 4b). The middle layer aggregates illness states through multiscale weighted pooling  $\bar{\mathbf{h}} = \text{pool} \{\mathbf{h}_0, \mathbf{h}_1, \dots, \mathbf{h}_n\}$ , where  $\bar{\mathbf{h}} \in \mathbb{R}^{sK}$  for  $s$  scales.

The top layer is a neural network that takes pooled states and other statistics to estimate the final outcome probability, as summarized in Eq. (1) as

$$P(y | \mathbf{u}_{0:n}) = P(\text{net}_y(\text{pool}\{\text{LSTM}(\mathbf{u}_{0:n})\}))$$

The probability  $P(y | \mathbf{u}_{0:n})$  depends on the nature of outputs and the choice of statistical structure. For example, for binary outcome,  $P(y = 1 | \mathbf{u}_{0:n})$  is a logistic function; for multiclass outcome,  $P(y | \mathbf{u}_{0:n})$  is a softmax function; and for continuous outcome,  $P(y | \mathbf{u}_{0:n})$  is a Gaussian. In what follows, we describe the first two layers in more detail.

### 4.2 Representing variable-size admissions

There are two main types of information recorded in an admission: (i) diagnoses of current condition; and (ii) interventions. Interventions include procedures and medications. Diagnoses, procedures and medications are coded using coding schemes which are described in Sec. 2.1. These schemes are hierarchical and the vocabularies are of tens of thousands in size. Thus for a problem, a suitable coding level should be used for balancing between specificity and robustness.

Our approach is to embed admissions into vectors. Fig. 4a illustrates the embedding method. An admission is a set of a varied number of codes (diagnoses and interventions). Codes are first embedded into vectors, analogous to word embedding described in Sec. 3.3. We then pool all the present diagnosis vectors to derive  $\mathbf{x}_t \in \mathbb{R}^M$ . Likewise, we derive the pooled intervention vector  $\mathbf{p}_t \in \mathbb{R}^M$ . Finally, an admission embedding is a  $2M$ -dim vector  $[\mathbf{x}_t, \mathbf{p}_t]$ .

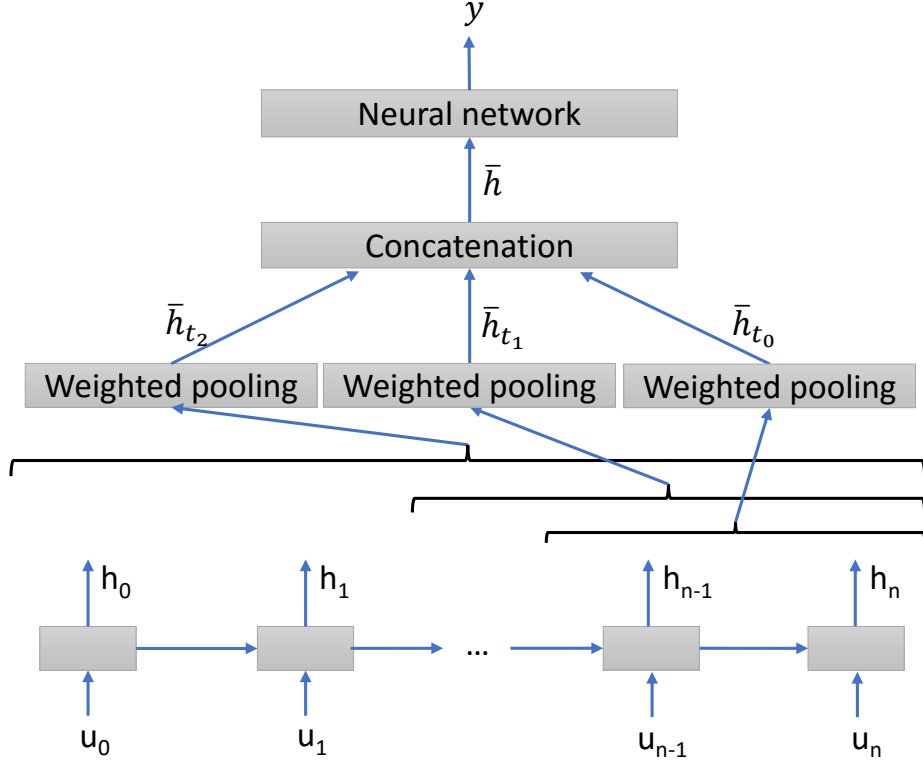


Figure 3: DeepCare architecture. The bottom layer is Long Short-Term Memory [14] with irregular timing and interventions ( see also Fig. 4).

### Pooling

Let  $\mathcal{D}$  be the set of diagnosis codes and  $\mathcal{I}$  be the set of intervention codes. The two sets are indexed from 1 to  $|\mathcal{D}|$  and from 1 to  $|\mathcal{I}|$ , respectively. Denote diagnosis embedding matrix by  $A \in \mathbb{R}^{M \times |\mathcal{D}|}$  and intervention embedding matrix by  $B \in \mathbb{R}^{M \times |\mathcal{I}|}$ . Let  $A^j$  is the  $j^{\text{th}}$  column and  $A_i^j$  is the element at the  $j^{\text{th}}$  column and the  $i^{\text{th}}$  row of the matrix  $A$ . Let  $x_t^i$  be the  $i^{\text{th}}$  element of the vector  $\mathbf{x}_t$  and  $p_t^i$  be the  $i^{\text{th}}$  element of the vector  $\mathbf{p}_t$ . Each admission  $t$  contains  $h$  diagnoses:  $d_1, d_2, \dots, d_h \in \{1, 2, \dots, |\mathcal{D}|\}$  and  $k$  interventions:  $s_1, s_2, \dots, s_k \in \{1, 2, \dots, |\mathcal{I}|\}$ . The admission is pooled by max, sum or mean pooling as follow:

- *Max pooling admission (max adm.)*. The pooling is element-wise as follows:

$$\mathbf{x}_t^i = \max \left( A_i^{d_1}, A_i^{d_2}, \dots, A_i^{d_h} \right)$$

$$\mathbf{p}_t^i = \max \left( B_i^{s_1}, B_i^{s_2}, \dots, B_i^{s_k} \right)$$

for  $i = 1, \dots, M$ . This is analogous to paying selective attention to the element of the highest impact among diagnoses and among interventions. It also resembles the usual coding practice that one diagnosis is picked as the primary reason for admission.

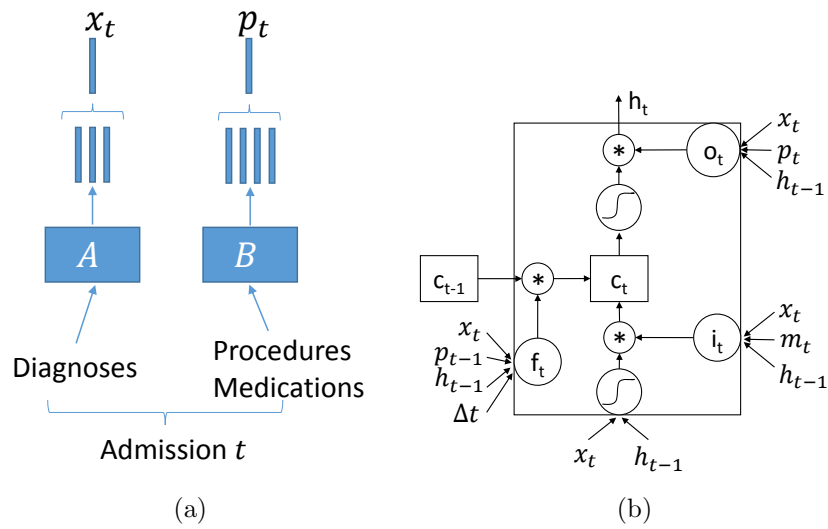


Figure 4: (a) Admission embedding.  $A$  and  $B$  are embedding matrices. Discrete diagnoses and interventions are embedded into 2 vectors  $\mathbf{x}_t$  and  $\mathbf{p}_t$ . (b) Modified LSTM unit as a carrier of illness history. Compared to the original LSTM unit (Fig. 2), the modified unit models times, admission methods, diagnoses and intervention

- *Normalized sum pooling admission (sum adm.)*. In healthcare, risk loosely adds up. A patient with multiple diseases (multiple comorbidities) is more likely to be at risk than those with single condition. We propose the following normalized sum pooling method:

$$\mathbf{x}_t^i = \frac{A_i^{d_1} + A_i^{d_2} + \dots + A_i^{d_h}}{\sqrt{|A_i^{d_1} + A_i^{d_2} + \dots + A_i^{d_h}|}}$$

$$\mathbf{p}_t^i = \frac{B_i^{s_1} + B_i^{s_2} + \dots + B_i^{s_k}}{\sqrt{|B_i^{s_1} + B_i^{s_2} + \dots + B_i^{s_k}|}}$$

for  $i = 1, \dots, M$ . The normalization is to reduce the effect of highly variable length.

- *Mean pooling admission (mean adm.)*. In absence of primary conditions, a mean pooling could be a sensible choice:

$$\mathbf{x}_t = \frac{A^{d_1} + A^{d_2} + \dots + A^{d_h}}{h}$$

$$\mathbf{p}_t = \frac{B^{s_1} + B^{s_2} + \dots + B^{s_k}}{k}$$

### Admission as input

Once admission embedding has been derived, diagnosis embedding is used as input for the LSTM. As interventions are designed to reduce illness, their effect is modeled separately in Sec. 4.3. There are two main types of admission: planned and unplanned. Unplanned admissions refer to transfer from emergency attendances, which typically indicates higher risk. Recall from Eqs. (5,8) that the input gate  $\mathbf{i}$  control how much new information is updated into memory  $\mathbf{c}$ . The gate can be modified to reflect the risk level of admission type as follows:

$$\mathbf{i}_t = \frac{1}{m_t} \sigma(W_i \mathbf{x}_t + U_i \mathbf{h}_{t-1} + \mathbf{b}_i) \quad (13)$$

where  $m_t = 1$  if the admission method is unplanned,  $m_t = 2$  otherwise, and  $\sigma$  is the element-wise sigmoid function of a vector.

### 4.3 Modeling effect of interventions

The intervention vector ( $\mathbf{p}_t$ ) of an admission is modeled as illustrated in Fig. 4b. Since interventions are designed to cure diseases or reduce patient’s illness, the output gate, which controls the illness states, is moderated by the *current* intervention as follows:

$$\mathbf{o}_t = \sigma(W_o \mathbf{x}_t + U_o \mathbf{h}_{t-1} + P_o \mathbf{p}_t + \mathbf{b}_o) \quad (14)$$

where  $P_o$  is the intervention weight matrix for the output gate and  $\mathbf{p}_t$  is intervention at time step  $t$ .

Moreover, interventions may have long-term impacts (e.g., curing disease or introducing toxicity). This suggests the illness forgetting is moderated by *previous* intervention

$$\mathbf{f}_t = \sigma(W_f \mathbf{x}_t + U_f \mathbf{h}_{t-1} + P_f \mathbf{p}_{t-1} + \mathbf{b}_f) \quad (15)$$

where  $\mathbf{p}_{t-1}$  is intervention embedded vector at time step  $t - 1$  and  $P_f$  is the intervention weight matrix for the forget gate.

## 4.4 Capturing time irregularity

When a patient’s history is modeled by LSTM (Sec. 3.2), the memory cell carries the illness history. But this memory needs not be constant as illness states change over time. We introduce two mechanisms of forgetting the memory by modified the forget gate  $\mathbf{f}_t$  in Eq. 15:

### Time decay

There are acute conditions that naturally reduce their effect through time. This suggests a simple decay modeled in the forget gate  $\mathbf{f}_t$ :

$$\mathbf{f}_t \leftarrow d(\Delta_{t-1:t}) \mathbf{f}_t \quad (16)$$

where  $\Delta_{t-1:t}$  is the time passed between step  $t - 1$  and step  $t$ , and  $d(\Delta_{t-1:t}) \in (0, 1]$  is a decay function, i.e., it is monotonically non-increasing in time. One function we found working well is  $d(\Delta_{t-1:t}) = [\log(e + \Delta_{t-1:t})]^{-1}$ , where  $\Delta_{t-1:t}$  is measured in days and  $e \approx 2.718$  is the the base of the natural logarithm.

### Parametric time

Time decay may not capture all conditions, since some conditions can get worse, and others can be chronic. This suggests a more flexible parametric forgetting:

$$\mathbf{f}_t = \sigma \left( W_f \mathbf{x}_t + U_f \mathbf{h}_{t-1} + Q_f \mathbf{q}_{\Delta_{t-1:t}} + P_f \mathbf{p}_{t-1} + \mathbf{b}_f \right) \quad (17)$$

where  $\mathbf{q}_{\Delta_{t-1:t}}$  is a vector derived from the time difference  $\Delta_{t-1:t}$ ,  $Q_f$  is the parametric time weight matrix. For example, we may have:  $\mathbf{q}_{\Delta_{t-1:t}} = \left( \frac{\Delta_{t-1:t}}{60}, \left( \frac{\Delta_{t-1:t}}{180} \right)^2, \left( \frac{\Delta_{t-1:t}}{365} \right)^3 \right)$  to model the third-degree forgetting dynamics.  $\Delta_{t-1:t}$  is measured in days and is divided by 60, 180 and 365 to prevent the vector  $\mathbf{q}_{\Delta_{t-1:t}}$  from large values.

## 4.5 Prognosis through multiscale pooling and recency attention

Once the illness dynamics have been modeled using the memory LSTM, the next step is to aggregate the illness states to infer about the future prognosis (Fig. 3). The simplest way is to use mean-pooling, where  $\bar{\mathbf{h}} = \text{pool} \{ \mathbf{h}_{0:n} \} = \frac{1}{n+1} \sum_{t=0}^n \mathbf{h}_t$ . However, this does not reflect the attention to recency in healthcare. Here we introduce a simple attention scheme that weighs recent events more than old ones:  $\bar{\mathbf{h}} = (\sum_{t=t_0}^n r_t \mathbf{h}_t) / \sum_{t=t_0}^n r_t$ , where

$$r_t = [m_t + \log(1 + \Delta_{t:n})]^{-1}$$

and  $\Delta_{t:n}$  is the elapsed time between the step  $t$  and the current step  $n$ , measured in months;  $m_t = 1$  if emergency admission,  $m_t = 2$  if routine admission. The starting time step  $t_0$  is used to control the length of look-back in the pooling, for example,  $\Delta_{t_0:n} \leq 12$  for one year look-back. Since diseases progress at different rates for different patients, we employ multiple look-backs: 12 months, 24 months, and all available history. Finally, the three pooled illness states are stacked into a vector:  $\bar{\mathbf{h}} = [\bar{\mathbf{h}}_{12}, \bar{\mathbf{h}}_{24}, \bar{\mathbf{h}}_{all}]$  which is then fed to a neural network for inferring about future prognosis.

## 4.6 Model complexity

The number of model parameters are  $M \times |V| + M \times K + K \times K + K \times D$ , which consists of the following components:

### *Parameters in the LSTM layer*

- For admission embedding, we use two embedding matrices  $A$  and  $B$ . We have  $A + B \in \mathbb{R}^{M \times |V|}$
- The input gate:  $W_i \in \mathbb{R}^{M \times K}$ ,  $U_i \in \mathbb{R}^{K \times K}$  and  $b_i \in \mathbb{R}^{K \times 1}$
- The output gate:  $W_o \in \mathbb{R}^{M \times K}$ ,  $U_o \in \mathbb{R}^{K \times K}$ ,  $P_o \in \mathbb{R}^{K \times K}$  and  $b_o \in \mathbb{R}^{K \times 1}$
- The forget gate:  $W_f \in \mathbb{R}^{M \times K}$ ,  $U_f \in \mathbb{R}^{K \times K}$ ,  $P_f \in \mathbb{R}^{K \times K}$  and  $b_f \in \mathbb{R}^{K \times 1}$ . In the case of time decay there are no other parameters and in the case of parametric time, the forget gate has a time weight matrix  $Q_f \in \mathbb{R}^{N_{time} \times K}$  ( $N_{time} = 3$  in our implementation)
- The memory cell:  $W_i \in \mathbb{R}^{M \times K}$ ,  $U_i \in \mathbb{R}^{K \times K}$  and  $b_i \in \mathbb{R}^{K \times 1}$

### *Parameters in the Neural network layer*

- The neural network layer consists of an input-hidden weight matrix  $U_{h1} \in \mathbb{R}^{3K \times D}$ , hidden-output weight matrix  $U_{h2} \in \mathbb{R}^{D \times 2}$  and two bias vectors  $c_1 \in \mathbb{R}^{D \times 1}$  and  $c_2 \in \mathbb{R}^{2 \times 1}$

## 4.7 Learning

Once all the illness states are pooled and stacked into vector  $\bar{\mathbf{h}}$ ,  $\bar{\mathbf{h}}$  is then fed to a neural network with one hidden layer

$$\mathbf{a}_h = \sigma(U_h \bar{\mathbf{h}} + \mathbf{b}_h) \quad (18)$$

$$\mathbf{z}_y = U_y \mathbf{a}_h + \mathbf{b}_y \quad (19)$$

$$P(y | \mathbf{u}_{1:n}) = f_{prob}(\mathbf{z}_y) \quad (20)$$

Learning is carried out through minimizing cross-entropy:  $L = -\log P(y | \mathbf{u}_{0:n})$ . For example, in the case of binary classification,  $y \in \{0, 1\}$ , we use logistic regression to represent  $P(y | \mathbf{u}_{0:n})$ , i.e.  $P(y = 1 | \mathbf{u}_{0:n}) = \sigma(\mathbf{z}_y)$ . The cross-entropy becomes

$$L = -y \log \sigma - (1 - y) \log (1 - \sigma) \quad (21)$$

Despite having a complex structure, DeepCare's loss function is fully differentiable, and thus can be minimized using standard back-propagation. The learning complexity is linear with the number of parameters. See Alg. 1 for an overview of DeepCare forward pass.

---

**Algorithm 1** DeepCare forward pass

---

**Inputs:** Patients’ disease history records

- 1: **for** each step  $t$  **do**
  - 2:    $[\mathbf{x}_t, \mathbf{p}_t] = \text{embedding}(d_1, \dots, d_h, s_1, \dots, s_k)$  (Sec. 4.2)
  - 3:   Compute 3 gates:  $\mathbf{i}_t$  (Eq. 13),  $\mathbf{o}_t$  (Eq. 14),  $\mathbf{f}_t$  (Eq. 16 or Eq. 17)
  - 4:   Compute  $\mathbf{c}_t$  (Eq. 9) and  $\mathbf{h}_t$  (Eq. 10)
  - 5: **end for**
  - 6: Compute  $\bar{\mathbf{h}}$  (Sec. 4.5)
  - 7: Compute  $P(y | \mathbf{u}_{0:n})$  (Eq. 18,19,20)
  - 8: Compute loss function  $L$  (Eq. 21)
- 

## 4.8 Pretraining and regularization

### Pretraining with auxiliary tasks

Pretraining can be done by unsupervised learning on unlabeled data [48, 49]. Pretraining has been proven to be effective because it helps the optimization by initializing weights in a region near a good local minimum [50, 51]. In our work we use auxiliary tasks to pretrain the model for future risk prediction tasks. In our case, auxiliary tasks are predicting diagnoses of the next readmission and predicting interventions of current admission. These tasks play a role in disease progression tracking and intervention recommendation.

We use the bottom layer of DeepCare for training auxiliary tasks. As described in Sec. 4.1, the LSTM layer reads a sequence of admissions  $\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_n$  and computes the corresponding sequence of distributed illness states  $\mathbf{h}_0, \mathbf{h}_1, \dots, \mathbf{h}_n$ . At each step  $t$ ,  $\mathbf{h}_t$  is used to generate labels  $y_t$  by the formula given in Eq. (11) where  $y_t$  can be a set of diagnoses or interventions. After training, the code embedding matrix is then used to initialize the embedding matrix for training the risk prediction tasks. The results of next readmission diagnosis prediction and current admission intervention prediction are reported in Sec. 5.3 and Sec. 5.4.

### Regularization

DeepCare may lead to overfitting because it introduces three more parameter matrices to the sigmoid gates to handle interventions and time. Therefore, we use L2-norm and Dropout to prevent overfitting. L2-norm regularization, also called “weight decay”, is used to prevent weight parameters from extreme values. A constant  $\lambda$  is introduced to control the magnitude of the regularization. Dropout is a regularization method for DNNs. During training, units are deleted with a pre-defined probability  $1 - p$  (dropout ratio) and the remaining parts are trained through back-propagation as usual [52, 53]. This prevents the co-adaptation between units, and therefore prevents overfitting. At the test time, a single neural net is used without dropout and the outgoing weights of a unit that is retained with probability  $p$  during training are multiplied by  $p$ . This combines  $2^k$  ( $k$  is the number of units) shared weight networks into a single neural network at test time. Therefore, dropout is also considered as an ensemble method.

However, the original version of dropout does not work well with RNNs because it may hurt the dependencies in sequential data during training [54, 55]. Thus, dropout in DeepCare is only introduced at input layer and neural network layer:



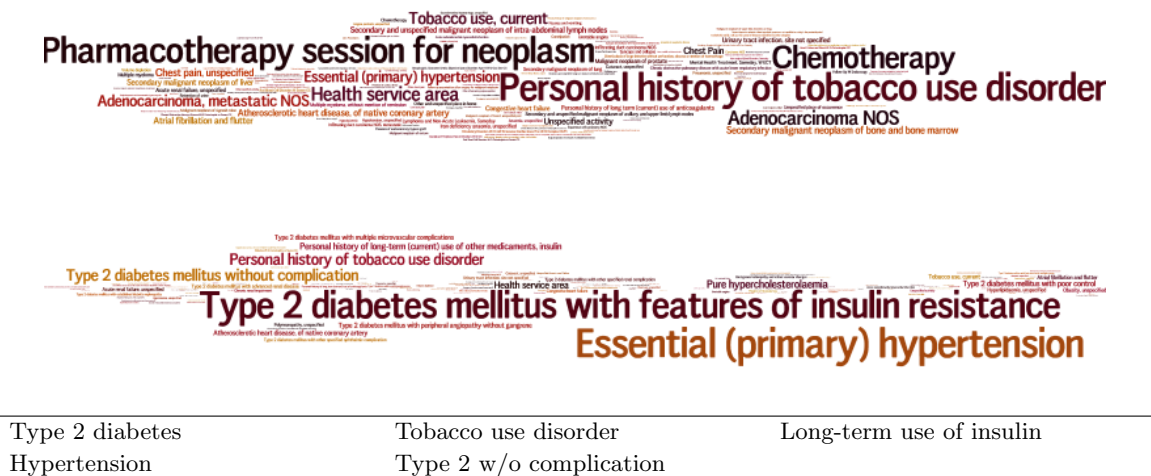
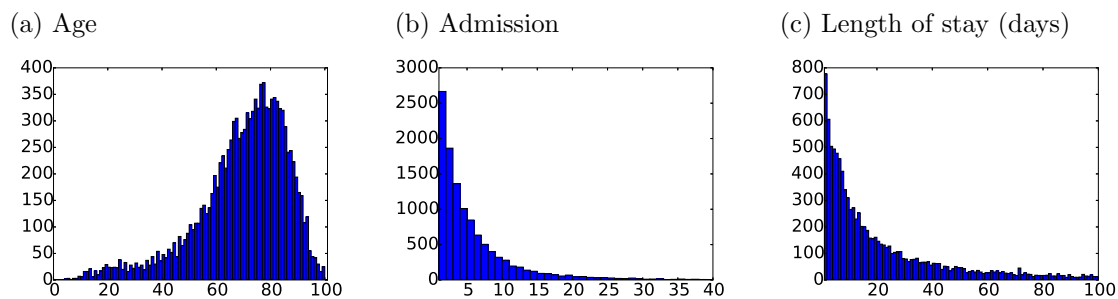


Figure 5: **Top row**: Diabetes cohort statistics (y axis: number of patients; x axis: (a) age, (b) number of admissions, (c) number of days); **Mid row**: Progression from pre-diabetes (upper diag. cloud) to post-diabetes (lower diag. cloud); **Bottom row**: Top diagnoses.

- *Dropout codes*: Before pooling the embedding vectors of diagnoses and interventions in each admission, each of these embedding vectors is deleted with probability  $1 - p_{code}$
- *Dropout input features*: After deriving  $[\mathbf{x}_t, \mathbf{p}_t]$  as described in Sec. 4.2, each value in these two vector is dropped with probability  $1 - p_{feat}$
- *Dropout units in neural network layer*: The pooled state  $z$  as described in Sec. 4.2 is feed as the input of the neural network. Dropout is used at input units with probability  $1 - p_{in}$  and at hidden units with probability  $1 - p_{hid}$ .

## 5 Experiments

We model disease progression, intervention recommendation and future risk prediction in two very diverse cohorts: mental health and diabetes. These diseases differ in causes and progression.

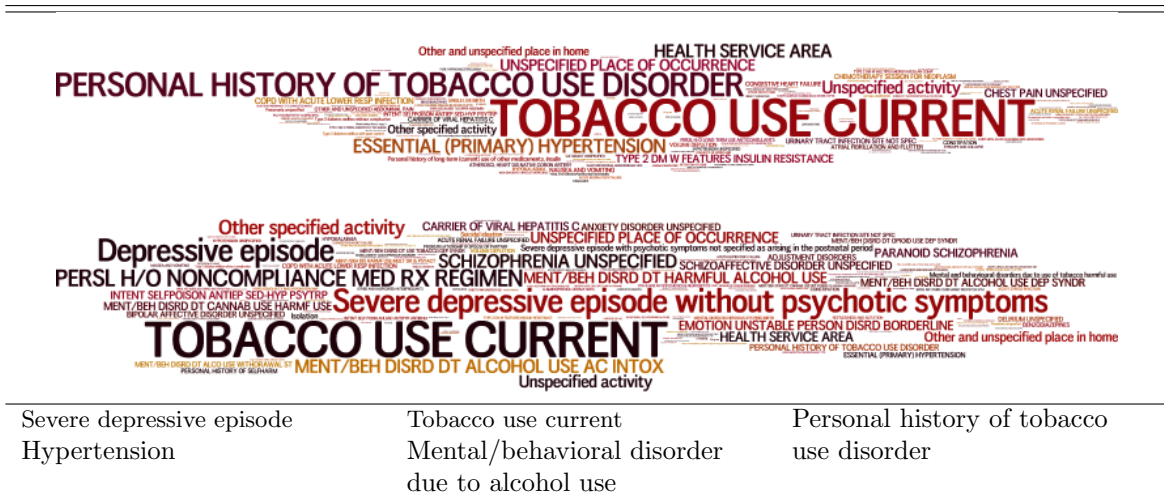
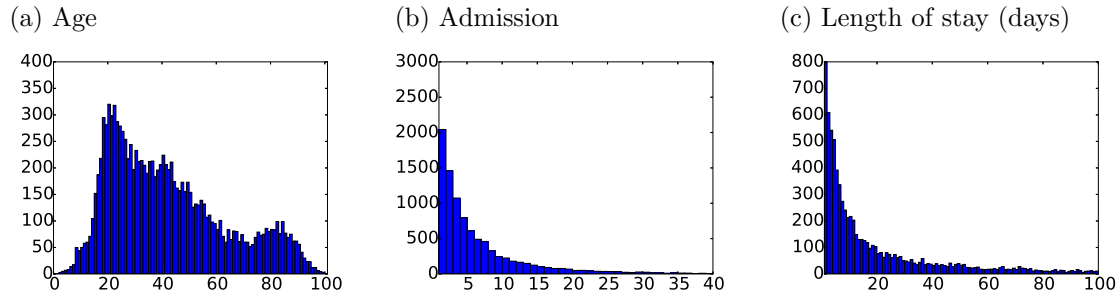


Figure 6: **Top row**: Mental health cohort statistics (y axis: number of patients; x axis: (a) age, (b) number of admissions, (c) number of days); **Mid row**: Progression from pre-mental diseases (upper diag. cloud) to post-mental diseases (lower diag. cloud); **Bottom row**: Top diagnoses.

## 5.1 Data

Data for both cohorts were collected for 12 years (2002-2013) from a large regional Australian hospital. We preprocessed the datasets by removing (i) admissions with incomplete patient information; and (ii) patients with less than 2 admissions. We define the vocabulary as the set of diagnosis, procedure and medication codes. To reduce the vocabulary, we collapse diagnoses that share the first 2 characters into one diagnosis. Likewise, the first digits in the procedure block are used.

The diabetes cohort contained more than 12,000 patients (55.5% males, median age 73). Data statistics are summarized in Fig. 5. After preprocessing, the dataset contained 7,191 patients with 53,208 admissions. The vocabulary consisted of 243 diagnosis, 773 procedure and 353 medication codes. The mental health cohort contains more than 11,000 patients (49.4% males, median age 37). Data statistics are summarized in Fig. 6. After preprocessing, the mental health dataset contained 6,109 patients and 52,049 admissions with the vocabulary of 247 diagnosis, 752 procedure and 319 medication codes. The average age of diabetic patients is much higher than the average age of mental patients (See Fig 5a and Fig 6a).

## 5.2 Implementation

The training, validation and test sets are created by randomly dividing the dataset into three parts of 2/3, 1/6, 1/6 data points, respectively. We vary the embedding and hidden dimensions from 5 to 50 but the results are rather robust. We report best results for disease progression and intervention recommendation tasks with  $M = 30$  and  $K = 40$  and for prediction tasks with  $M = 10$  embedding dimensions and  $K = 20$  hidden units ( $M$  and  $K$  are the number of embedding dimensions and hidden units respectively). Learning is by Stochastic Gradient Descent with the mini-batch of 16 sequences. The learning rate  $\lambda$  is modified as follows. We start with  $\lambda = 0.01$ . When the model cannot find a smaller training cost, we wait  $n_{wait}$  epochs before updating  $\lambda$  as  $\lambda = \lambda/2$ . Initially,  $n_{wait} = 5$ , and is subsequently modified as  $n_{wait} = \min\{15, n_{wait} + 2\}$  for each  $\lambda$  update. Learning is terminated after  $n_{epoch} = 200$  or after learning rate smaller than  $\epsilon = 0.0001$ .

## 5.3 Disease progression

We first verify that the recurrent memory embedded in DeepCare is a realistic model of *disease progression*. The model predicts the next  $n_p$  diagnoses at each discharge using Eq. (11).

For comparison, we implement two baselines: Markov models and plain RNNs. Markov model is a stochastic model used to model changing systems. A Markov model consists of a list of possible states, the possible transitions between those states and the probability of those transitions. The future states depend only on the present state (Markov assumption). The Markov model has memoryless disease transition probabilities  $P(d_t^i | d_{t-1}^j)$  from disease  $d^j$  to  $d^i$  at time  $t$ . Given an admission with disease subset  $D_t$ , the next disease probability is estimated as  $Q(d^i; t) = \frac{1}{|D_t|} \sum_{j \in D_t} P(d_t^i | d_{t-1}^j)$ . Plain RNNs are described in Sec. 3.1.

We use Precision at K (Precision@K) to measure the performance of the models. Precision@K corresponds to the percentage of relevant results in retrieved results. That means if the model predicts  $n_p$  diagnoses of the next readmission and  $n_r$  diagnoses among of them are relevant the model's performance is

$$\text{Precision}@n_p = \frac{n_r}{n_p}$$

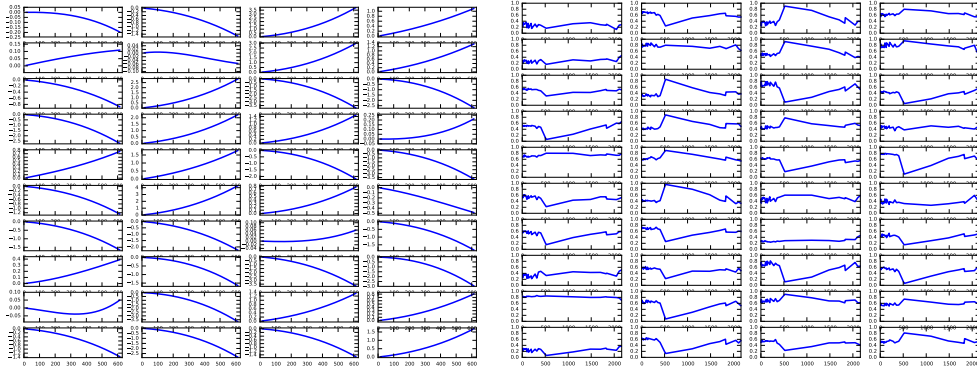


Figure 7: (Left) 40 channels of forgetting due to time elapsed. (Right) The forget gates of a patient in the course of their illness.

Table 1: Precision@ $n_p$  Diagnoses Prediction.

	Diabetes			Mental		
	$n_p = 1$	$n_p = 2$	$n_p = 3$	$n_p = 1$	$n_p = 2$	$n_p = 3$
Markov	55.1	34.1	24.3	9.5	6.4	4.4
Plain RNN	63.9	58.0	52.0	50.7	45.7	39.5
DeepCare (mean adm.)	<b>66.2</b>	<b>59.6</b>	<b>53.7</b>	<b>52.7</b>	<b>46.9</b>	<b>40.2</b>
DeepCare (sum adm.)	65.5	59.3	53.5	51.7	46.2	39.8
DeepCare (max adm.)	66.1	59.2	53.2	51.5	46.7	<b>40.2</b>

### Dynamics of forgetting

Fig. 7(left) plots the contribution of time into the forget gate. The contributions for all 40 states are computed using  $Q_f \mathbf{q}_{\Delta_t}$  as in Eq. (17). There are two distinct patterns: decay and growing. This suggests that the time-based forgetting has a very small dimensionality, and we will under-parameterize time using decay only as in Eq. (16), and over-parameterize time using full parameterization as in Eq. (17). A right balance is interesting to warrant a further investigation. Fig. 7(right) shows the evolution of the forget gates through the course of illness (2000 days) for a patient.

### Diagnoses prediction result

Table 1 reports the Precision@ $n_p$  for different values of  $n_p$ . For diabetes cohort, using plain RNN improves over memoryless Markov model by 8.8% with  $n_p = 1$  and by 27.7% with  $n_{pred} = 3$ . This significant improvement demonstrates the role of modeling the dynamics in sequential data. Modeling irregular timing and interventions in DeepCare gains a further 2% improvement. For mental health cohort, Markov model is failed to predict next diagnoses with only 9.5% for  $n_p = 1$ . Plain RNN gains 50% improvement in Precision@1, while and DeepCare demonstrates a 2% improvement in Precision@1 over RNN.

Table 2: Precision@ $n_p$  intervention prediction

	Diabetes			Mental		
	$n_p = 1$	$n_p = 2$	$n_p = 3$	$n_p = 1$	$n_p = 2$	$n_p = 3$
Markov	35.0	17.6	11.7	20.7	12.2	8.1
Plain RNN	77.7	54.8	43.1	70.4	55.4	43.7
DeepCare (mean adm.)	77.8	54.9	43.3	70.3	55.7	44.1
DeepCare (sum adm.)	<b>78.7</b>	<b>55.5</b>	<b>43.5</b>	<b>71.0</b>	<b>55.8</b>	<b>44.7</b>
DeepCare (max adm.)	78.4	55.1	43.4	70.0	55.2	43.9

Table 3: Effect of pretraining and regularization for unplanned readmission prediction using DeepCare for diabetes dataset. The results are reported in F-score (%)

Approach	Mean adm.	Sum adm.	Max adm.
None	77.8	77.9	78.3
Pretrain	78.3	78.6	<b>78.9</b>
Regularization	<b>79.0</b>	78.7	78.6
Both	78.4	<b>78.9</b>	78.8

## 5.4 Intervention recommendation

Table 2 reports the results of current intervention prediction. For all values of  $n_p$ , RNN consistently outperforms Markov model by a huge margin for both diabetes and mental health cohort. DeepCare with sum-pooling outperforms other models in both diabetes and mental health datasets.

## 5.5 Predicting future risk

Next we demonstrate DeepCare on risk prediction. For each patient, a discharge is randomly chosen as prediction point, from which *unplanned readmission* and *high risk patients* within  $X$  months will be predicted. A patient is in high risk at a particular time  $T$  if he or she have at least three unplanned readmissions within  $X$  months after time  $T$ . We choose  $X = 12$  months for diabetes and  $X = 3$  months for mental health.

For comparison, **baselines** are SVM and Random Forests running on standard non-temporal features engineering using one-hop representation of diagnoses and intervention codes. Then pooling is applied to aggregate over all existing admissions for each patient. Two pooling strategies are tested: *max* and *sum*. Max-pooling is equivalent to the presence-only strategy in [9], and sum-pooling is akin to an uniform convolutional kernel in [4]. This feature engineering strategy is equivalent to zeros-forgetting – any risk factor occurring in the past is memorized.

### Pretraining and Regularization

Table 3 reports the impacts of pretraining and regularization on the results of unplanned readmission prediction in diabetes dataset using DeepCare model. Pretraining and regularization improve the results of all three admission pooling methods. While mean pooling admission is found to perform well with regularization, max pooling produces best results with pretraining and sum pooling produces best results with both approaches.

Table 4: Results of unplanned readmission prediction in F-score (%) within 12 months for diabetes and 3 months for mental health patients (DC is DeepCare, inv. is intervention).

Model	Diabetes	Mental
1. SVM ( <i>max-pooling</i> )	64.0	64.7
2. SVM ( <i>sum-pooling</i> )	66.7	65.9
3. Random Forests ( <i>max-pooling</i> )	68.3	63.7
4. Random Forests ( <i>sum-pooling</i> )	71.4	67.9
5. Plain RNN ( <i>logist. regress.</i> )	75.1	70.5
6. LSTM ( <i>logit. regress.</i> )	75.9	71.7
7. DC ( <i>nnets + mean adm.</i> )	76.5	72.8
8. DC ( [ <i>inv.+time decay</i> ] + <i>recent.multi.pool.+nnets+mean adm.</i> )	77.1	74.5
<b>9. DC ([<i>inv.+param. time</i>] +<i>recent.multi.pool.+nnets+mean adm.</i>)</b>	<b>79.0</b>	<b>74.7</b>

### Unplanned readmission prediction results

Table 4 reports the F-scores of predicting unplanned readmission. For the diabetes cohort, the best baseline (non-temporal) is Random Forests with *sum pooling* has a F-score of 71.4% [Row 4]. Using plain RNN with simple logistic regression improves over best non-temporal methods by a 3.7% difference in 12-months prediction [Row 5, ref: Sec. (3.1,4.2)]. Replacing RNN units by LSTM units gains 4.5% improvement [Row 6, ref: Sec. 3.2]. Moving to deep models by using a neural network as classifier helps with a gain of 5.1% improvement [Row 7, ref: Eq. (1)]. By carefully modeling the irregular timing, interventions and recency+multiscale pooling, we gain 5.7% improvement [Row 8, ref: Secs. (4.4–4.5)]. Finally, with parametric time we arrive at 79.0% F-score, a 7.6% improvement over the best baselines [Row 9, ref: Secs. (4.4)].

For the mental health dataset, the best non-temporal baseline is *sum-pooling* Random Forest with result of 67.9%. Plain RNN and LSTM with logistic regression layer gain 2.6% and 3.8% improvements, respectively. The best model is DeepCare with parametric time with a gap of 6.8% improvement compared to *sum-pooling* Random Forest.

### High risk prediction results

In this part, we report the performance of DeepCare on high risk patient prediction task. Figure 8 reports the F-score of high risk prediction. RNN improves the best non-temporal model (*sum-pooling* SVM) over 10% F-score for both two cohorts. Max-pooling DeepCare best performs in diabetes dataset with nearly 60% F-score, while sum-pooling DeepCare wins in mental health cohort with 50.0% F-score.

## 6 Discussion and Conclusion

### 6.1 Discussion

DeepCare was partly inspired by human memory [56]. There are three kinds of related memory: semantic, episodic and working memory. *Semantic memory* stores the meaning of concepts and their relations. *Episodic memory* refers to the storage of experiences triggered by an event, for example, wedding or earthquake. *Working* memory is a system of temporarily loading and processing information as part of complex cognitive tasks.

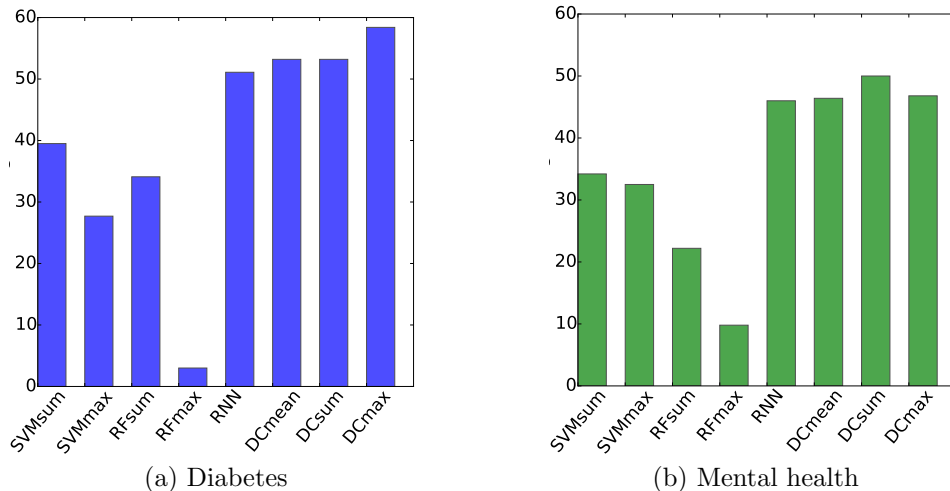


Figure 8: Result of high risk prediction in F-score (%) within 12 months for diabetes (a) and 3 months for mental health (b). DC is DeepCare. Mean, sum, max are 3 admission pooling methods

DeepCare makes use of embedding to represent the semantics of diagnoses, interventions and admissions. In theory, this embedding can be estimated independently of the task at hand. Our previous work learns diagnosis and patient embedding [12] using nonnegative restricted Boltzmann machines [57] and known semantic relations and temporal relations [58]. This method uses global contexts, unlike DeepCare, where only local contexts (e.g., next admission) are considered.

The memory cells in DeepCare are used to store, update, forget and manipulate illness experiences over time-stamped episodes. The inferred experiences are then pooled to reason about the current illness states and the future prognosis. Like human memory, healthcare risk also has a recency effect, that is, more recent events contribute more into the future risk. In DeepCare, two recency mechanisms are used. First, through forgetting, recent events in DeepCare tend to contribute more to the current illness states. Second, multiscale pooling as in Sec. 4.5 has weights decayed over time.

DeepCare can be implemented on existing EMR systems. For that more extensive evaluations on a variety of cohorts, sites and outcomes will be necessary. This offers opportunities for domain adaptations through parameter sharing among multiple cohorts and hospitals. Modeling-wise, DeepCare can also be extended to predict a sequence of outcomes at specific timing, in the same spirit as the sequence to sequence mapping in [17]. Future work also includes more flexibility in time parameterization such as using radial basis expansion and splines. Further, DeepCare is generic so it can be applied to not only medical data but also other kinds of sequential data which contain long-term dependencies, sequence of sets, irregular time and interventions.

## 6.2 Conclusion

In this paper we have introduced DeepCare, an *end-to-end* deep dynamic memory neural network for personalized healthcare. It frees model designers from manual feature extraction. DeepCare reads medical records, memorizes illness trajectories and care processes, estimates the present illness states, and predicts the future risk. Our framework models disease progression, supports intervention recom-

mentation, and provides prognosis from electronic medical records. To achieve precision and predictive power, DeepCare extends the classic Long Short-Term Memory by (i) embedding variable-size discrete admissions into vector space, (ii) parameterizing time to enable irregular timing, (iii) incorporating interventions to reflect their targeted influence in the course of illness and disease progression; (iv) using multiscale pooling over time; and finally (v) augmenting a neural network to infer about future outcomes. We have demonstrated DeepCare on predicting next disease stages, recommending interventions, and estimating unplanned readmission among diabetic and mental health patients. The results are competitive against current state-of-the-arts. DeepCare opens up a new principled approach to predictive medicine.

## References

- [1] E. W. Steyerberg, *Clinical prediction models: a practical approach to development, validation, and updating*. Springer, 2009.
- [2] R. Snyderman and R. S. Williams, “Prospective medicine: the next health care transformation,” *Academic Medicine*, vol. 78, no. 11, pp. 1079–1084, 2003.
- [3] P. B. Jensen, L. J. Jensen, and S. Brunak, “Mining electronic health records: towards better research applications and clinical care,” *Nature Reviews Genetics*, vol. 13, no. 6, pp. 395–405, 2012.
- [4] T. Tran, W. Luo, D. Phung, S. Gupta, S. Rana, R. L. Kennedy, A. Larkins, and S. Venkatesh, “A framework for feature extraction from hospital medical data with applications in risk prediction,” *BMC bioinformatics*, vol. 15, no. 1, p. 6596, 2014.
- [5] T. Tran, D. Phung, W. Luo, and S. Venkatesh, “Stabilized sparse ordinal regression for medical risk stratification,” *Knowledge and Information Systems*, 2014, doi: 10.1007/s10115-014-0740-4.
- [6] T. Tran, D. Phung, W. Luo, R. Harvey, M. Berk, and S. Venkatesh, “An integrated framework for suicide risk prediction,” in *KDD’13*, 2013. [Online]. Available: 2013/conferences/tran.et.al.kdd13.pdf
- [7] C. H. Jackson, L. D. Sharples, S. G. Thompson, S. W. Duffy, and E. Couto, “Multistate Markov models for disease progression with classification error,” *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 52, no. 2, pp. 193–209, 2003.
- [8] H. Lu, D. Zeng, and H. Chen, “Prospective Infectious Disease Outbreak Detection Using Markov Switching Models,” *IEEE Transactions on Knowledge and Data Engineering*, 2009.
- [9] O. Arandjelović, “Discovering hospital admission patterns using models learnt from electronic hospital records,” *Bioinformatics*, p. btv508, 2015.
- [10] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [11] Z. Liang, G. Zhang, J. X. Huang, and Q. V. Hu, “Deep learning for healthcare decision making with EMRs,” in *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*. IEEE, 2014, pp. 556–559.



- [12] T. Tran, T. D. Nguyen, D. Phung, and S. Venkatesh, “Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM),” *Journal of biomedical informatics*, vol. 54, pp. 96–105, 2015.
- [13] J. Futoma, J. Morris, and J. Lucas, “A comparison of models for predicting early hospital readmissions,” *Journal of biomedical informatics*, vol. 56, pp. 229–238, 2015.
- [14] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [15] A. Graves, “Generating sequences with recurrent neural networks,” *arXiv preprint arXiv:1308.0850*, 2013.
- [16] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, “A novel connectionist system for unconstrained handwriting recognition,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 5, pp. 855–868, 2009.
- [17] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Advances in Neural Information Processing Systems*, 2014, pp. 3104–3112.
- [18] C. Paxton, A. Niculescu-Mizil, and S. Saria, “Developing predictive models using electronic medical records: challenges and pitfalls,” in *AMIA Annual Symposium Proceedings*, vol. 2013. American Medical Informatics Association, 2013, p. 1109.
- [19] P. Groves, B. Kayyali, D. Knott, and S. Van Kuiken, “The ‘big data’ revolution in healthcare,” *McKinsey Quarterly*, 2013.
- [20] N. G. Weiskopf, G. Hripcsak, S. Swaminathan, and C. Weng, “Defining and measuring completeness of electronic health records for secondary use,” *Journal of biomedical informatics*, vol. 46, no. 5, pp. 830–836, 2013.
- [21] J. S. Mathias, A. Agrawal, J. Feinglass, A. J. Cooper, D. W. Baker, and A. Choudhary, “Development of a 5 year life expectancy index in older adults using predictive mining of electronic health record data,” *Journal of the American Medical Informatics Association*, vol. 20, no. e1, pp. e118–e124, 2013.
- [22] G. Hripcsak and D. J. Albers, “Next-generation phenotyping of electronic health records,” *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 117–121, 2013.
- [23] B. B. Granger, D. Moser, B. Germino, J. Harrell, and I. Ekman, “Caring for patients with chronic heart failure: The trajectory model,” *European Journal of Cardiovascular Nursing*, vol. 5, no. 3, pp. 222–227, 2006.
- [24] Z. Huang, W. Dong, H. Duan, and H. Li, “Similarity measure between patient traces for clinical pathway analysis: problem, method, and applications.” *IEEE journal of biomedical and health informatics*, vol. 18, no. 1, pp. 4–14, 2014.
- [25] J. M. Corbin and A. Strauss, “A nursing model for chronic illness management based upon the trajectory framework,” *Research and Theory for Nursing Practice*, vol. 5, no. 3, pp. 155–174, 1991.

- [26] S. J. Henly, J. F. Wyman, and M. J. Findorff, “Health and illness over time: The trajectory perspective in nursing science,” *Nursing research*, vol. 60, no. 3 Suppl, p. S5, 2011.
- [27] A. B. Jensen, P. L. Moseley, T. I. Oprea, S. G. Ellesøe, R. Eriksson, H. Schmock, P. B. Jensen, L. J. Jensen, and S. Brunak, “Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients,” *Nature communications*, vol. 5, 2014.
- [28] C. Liu, F. Wang, J. Hu, and H. Xiong, “Temporal phenotyping from longitudinal electronic health records: A graph based framework,” in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 705–714.
- [29] X. Wang, D. Sontag, and F. Wang, “Unsupervised learning of disease progression models,” in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 85–94.
- [30] K. Orphanou, A. Stassopoulou, and E. Keravnou, “Temporal abstraction and temporal Bayesian networks in clinical domains: A survey,” *Artificial intelligence in medicine*, vol. 60, no. 3, pp. 133–149, 2014.
- [31] M. Ramati and Y. Shahar, “Irregular-time Bayesian networks,” *UAI*, pp. 484–491, 2010.
- [32] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, “Predicting the sequence specificities of dna-and rna-binding proteins by deep learning,” *Nature biotechnology*, 2015.
- [33] G. DE.Rumelhart and R.J.Williams, “Learning representations by back-propagating errors,” *Nature*, pp. 323–533, 1986.
- [34] K. J. Lang, A. H. Waibel, and G. E. Hinton, “A time-delay neural network architecture for isolated word recognition,” *Neural networks*, vol. 3, no. 1, pp. 23–43, 1990.
- [35] H. Jaeger, “Echo state network,” *Scholarpedia*, vol. 2, no. 9, p. 2330, 2007.
- [36] P. J. Werbos, “Backpropagation through time: what it does and how to do it,” *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.
- [37] Y. Bengio, P. Simard, and P. Frasconi, “Learning long-term dependencies with gradient descent is difficult,” *Neural Networks, IEEE Transactions on*, vol. 5, no. 2, pp. 157–166, 1994.
- [38] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” *arXiv preprint arXiv:1211.5063*, 2012.
- [39] M. C. Mozer, “Induction of multiscale temporal structure,” *Advances in neural information processing systems*, pp. 275–275, 1993.
- [40] T. Lin, B. G. Horne, P. Tiño, and C. L. Giles, “Learning long-term dependencies in narx recurrent neural networks,” *Neural Networks, IEEE Transactions on*, vol. 7, no. 6, pp. 1329–1338, 1996.
- [41] F. A. Gers, J. Schmidhuber, and F. Cummins, “Learning to forget: Continual prediction with lstm,” *Neural computation*, vol. 12, no. 10, pp. 2451–2471, 2000.
- [42] A. Graves, M. Liwicki, H. Bunke, J. Schmidhuber, and S. Fernández, “Unconstrained on-line handwriting recognition with recurrent neural networks,” in *Advances in Neural Information Processing Systems*, 2008, pp. 577–584.

- [43] A. Graves, A.-r. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6645–6649.
- [44] Q. V. Le and T. Mikolov, “Distributed representations of sentences and documents,” *ICML*, 2014.
- [45] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, “A neural probabilistic language model,” *The Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.
- [46] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.
- [47] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, “Recurrent neural network based language model.” in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, 2010, pp. 1045–1048.
- [48] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.
- [49] A. M. Dai and Q. V. Le, “Semi-supervised sequence learning,” *arXiv preprint arXiv:1511.01432*, 2015.
- [50] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle *et al.*, “Greedy layer-wise training of deep networks,” *Advances in neural information processing systems*, vol. 19, p. 153, 2007.
- [51] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, “Why does unsupervised pre-training help deep learning?” *The Journal of Machine Learning Research*, vol. 11, pp. 625–660, 2010.
- [52] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.
- [53] P. Baldi and P. J. Sadowski, “Understanding dropout,” in *Advances in Neural Information Processing Systems*, 2013, pp. 2814–2822.
- [54] J. Bayer, C. Osendorfer, D. Korhammer, N. Chen, S. Urban, and P. van der Smagt, “On fast dropout and its applicability to recurrent networks,” *arXiv preprint arXiv:1311.0701*, 2013.
- [55] W. Zaremba, I. Sutskever, and O. Vinyals, “Recurrent neural network regularization,” *arXiv preprint arXiv:1409.2329*, 2014.
- [56] A. Baddeley, “Working memory,” *Science*, vol. 255, no. 5044, pp. 556–559, 1992.
- [57] T. Nguyen, T. Tran, D. Phung, and S. Venkatesh, “Learning Parts-based Representations with Nonnegative Restricted Boltzmann Machine,” in *Proc. of 5th Asian Conference on Machine Learning (ACML)*, Canberra, Australia, Nov 2013.
- [58] T. D. Nguyen, T. Tran, D. Phung, and S. Venkatesh, “Graph-induced restricted boltzmann machines for document modeling,” *Information Sciences*, vol. 328, pp. 60–75, 2016.