Deepr: A Convolutional Net for Medical Records

Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, Svetha Venkatesh

***Abstract*—Feaure engineering remains a major bottleneck when creating predictive systems from electronic medical records. At present, an important missing element is detecting predictive *regular clinical motifs* from *irregular episodic records*. We present Deepr (short for Deep record), a new *end-to-end* deep learning system that learns to extract features from medical records and predicts future risk automatically. Deepr transforms a record into a sequence of discrete elements separated by coded time gaps and hospital transfers. On top of the sequence is a convolutional neural net that detects and combines predictive local clinical motifs to stratify the risk. Deepr permits transparent inspection and visualization of its inner working. We validate Deepr on hospital data to predict unplanned readmission after discharge. Deepr achieves superior accuracy compared to traditional techniques, detects meaningful clinical motifs, and uncovers the underlying structure of the disease and intervention space.**

I. Introduction

A major theme in modern medicine is *prospective healthcare*, which refers to the capability to estimate the future medical risks for individuals. These risks can include readmission after discharge, the onset of specific diseases, and worsening from a condition [42]. Such capability would facilitate timely prevention or intervention for maximum health impact, and provide a major step toward personalized medicine. An important data resource in aiding this process are electronic medical records [20]. Electronic medical records (EMRs) contain a wealth of patient information over time. Central to EMR-driven risk prediction is *patient representation*, also known as feature engineering. Representing an EMR amounts to extracting relevant historical signals to form a feature vector.

However, feature extraction in EMR is challenging [44]. An EMR typically consists of a sequence of time-stamped visit episodes, each of which has a subset of coded diagnoses, a subset of procedures, lab tests and textual narratives. The data is *irregular at patient level*. EMR is episodic – events are only recorded when patients visit clinics, and the time gap between two visits is largely random. Representing irregular timing poses a major challenge. EMR varies greatly in length – young patients usually have just one visit for an acute condition, but old patients with chronic conditions may have hundreds of visits. At the same time, the data is *regular at local episode level*. Diseases tend to form clusters (comorbidity) [41] and the disease progression may be dictated by the underlying biological processes [49]. Likewise treatments may follow a certain protocol or best practice guideline [17], and there are well-defined disease-treatment interactions [39]. These regularities can be thought as clinical motifs. Thus an effective EMR representation should be able to identify *regular clinical motifs* out of *irregular data*.

Existing EMR-driven predictive work often relies on high-dimensional sparse feature representation, where features are engineered to capture certain regularities of the data [11], [20] This feature engineering practice is effort intensive and non-adaptive to varying medical records systems. Automated feature representation based on bag-of-words (BoW) is scalable, but it breaks collocation relations between words and ignores the temporal nature of the EMR, thus it fails to properly address the aforementioned challenges.

In this work we present a new prediction framework called Deepr that does not require manual feature engineering. The technology is based on *deep learning,* a new revolutionary approach that aims to build a *multilayered neural learning system* like a brain [25]. When fed with a large amount of raw data, the system learns to recognize patterns with little help from domain experts. Deep learning now powers speech recognition in Google Voice, self-driving cars at Google and Baidu, question answering system at IBM (Watson), and smart assistants at Facebook. It already has a great impact on hundreds of millions (if not billions) people. But healthcare has largely been ignored. We hypothesize that a key to apply deep learning for healthcare patient representation which requires a proper handling of the irregular nature of episodes mentioned above [37]. Deepr fills the gap by offering an *end-to-end* technology that *learns* to represent patients from scratch. It reads medical records, learns the local patterns, adapts to irregular timing, and predicts personalized risk.

The architecture of Deepr is multilayered and is inspired by recent convolutional neural nets (CNNs) in natural languages [9], [21], [25], [30], [51]. The most crucial operation occurs at the bottom level where Deepr transforms an EMR into a "sentence" of multiple phrases separated by special "words" that represent time gap. Each phrase is an visit episode. As with syntactical grammars and collocation patterns in NLP, there might exist "health grammars" and "clinical patterns" in healthcare. Health grammars refer to latent biological and environmental laws that dictate the global evolution of one's health *over time,* e.g., probable progression from "diabetes type II" to "renal failure". To handle irregular timing, time gaps and transfers are treated as special words. With this representation, an EMR is transformed into a sentence of variable length that retains all important events. The other layers of Deepr constitute a CNN, which is similar to those in [9], [21], [51]. First, words are embedded into a continuous vector space. Next, words in sentence are passed through a convolution operation which detects local motifs. Local motifs are then pooled to form a global feature vector, which is passed into a classifier, which predicts the future risk. All components are learned at the same time from data: the data signals are passed from the data to the output, and the training signals are propagated back from the labels to the motif detectors. Hence Deepr is *end-to-end*.

We validate Deepr on a large database of 300K patients collected from a hospital chain in Australia. We focus on

predicting **unplanned readmission within 6 months** after discharge. Compared to existing bag-of-words representation, Deepr demonstrates a superior accuracy as well as the capacity to learn predictive clinical motifs, and to uncover the underlying structure of the space of diseases and interventions.

To summarize, we claim the following contributions:

- A novel representation of irregular-time EMR as a sentence with time gaps and transfers as special words.
- A novel deep learning architecture called Deepr that (i) uncovers the structure of the disease/treatment space, (ii) discovers clinical motifs, (iii) predicts future risk and (iv) explains the prediction by identifying motifs with strong responses in each record. The system is end-to-end, and its inner working can be inspected and visualized, allowing interpretability and transparency.
- An evaluation of these claimed capabilities on a large-scale dataset of 300K patients.

II. Background

a) Medical records: An electronic medical record (EMR) contains information about patient demographics and a sequence of hospital visits for a patient. Admission information may include admission time, discharge time, lab tests, diagnoses, procedures, medications and clinical narratives. Diagnoses, procedures and medications are discrete entities. For example, diagnoses may be represented using ICD-10 coding schemes[1]. For example, in ICD-10, E10 refers to Type 1 diabetes mellitus, E11 to Type 2 diabetes mellitus. The procedures are typically coded in CPT (Current Procedural Terminology) or ICHI (International Classification of Health Interventions) schemes [2]. One of the most important secondary uses of EMR is building predictive models [20], [31], [44], [46].

Most existing prediction methods on EMRs either rely on manual feature engineering [31] or simplistic extraction [44]. They either ignore long-term dependencies or do not adequately capture variable length [2], [31], [44]. Neither are they able to model temporal irregularity [18], [29], [44], [49]. Capturing disease progression has been of great interest [19], [29], and much effort has been spent on Markov models [14], [18], [49]. As Markov processes are memoryless, Markov models forget severe conditions of the past when it sees an admission due to common cold. This is undesirable. A proper modeling, therefore, must be non-Markovian and able to capture long-term dependencies.

b) Deep learning: Deep learning is an approach in machine learning, aiming at producing *end-to-end* systems that learn from raw data and perform desired tasks without manual feature engineering. The current wave of deep learning was initiated by the seminal work of [15] in 2006, but deep learning has been developed for decades [40]. Over the past few years, deep learning has broken records in cognitive domains such as vision, speech and natural languages [25]. Current deep learning is mostly based on multilayered neural networks [40]. All the networks share a common unit – the neuron – which is a simple computational device that applies a nonlinear transform to a linear function of inputs: i.e., $f(x) = \sigma\left(b + \sum_i w_i x_i\right)$. Almost all networks thus far are trained using back-propagation [50], thus enable end-to-end learning.

There are three main deep neural architectures in practice: *feedforward*, *recurrent* and *convolutional*. Feedforward nets (FFN) pass unstructured information from one end to the other, usually from an input to an output, hence they act as a universal function approximator [16]. Recurrent nets (RNN) model dynamics over time (and space) using self-replicated units. They maintain some degree of memory, and thus have potential to capture long-term dependencies. RNNs are powerful computational machines – they can approximate any program [27]. Convolutional nets (CNN) exploit the repeated local motifs across time and space, and thus are translation-invariant – the capacity often seen in human visual cortex [24]. Local motifs are small piece of data, usually of pre-defined sizes, e.g., a batch of pixels, or a n-gram of words. CNN is often equipped with pooling operations to reduce the resolution and enlarge the motifs.

III. Deepr: A Deep Net for Medical Records

In this section, we describe our deep neural net named Deepr (short for Deep net for medical Record) for representing Electronic Medical Records (EMR) and predicting the future risk.

A. Deepr Overview

Deepr is a multilayered architecture based on convolutional neural nets (CNNs). The information flow is summarized in Fig. 1. At the bottom level, Deepr *sequences* the EMR into a "sentence", or equivalently, a sequence of "words". Each word represents a discrete object or event such as diagnosis, procedure, or any derived object such as time-interval or hospital transfer. The next layer *embeds* words into an Euclidean space. On top of the embedding layer is a CNN that reads a small chunk of words in a sliding window to identify local motifs. The local motifs are transformed by Rectified Linear Unit (*ReLU*), which is a nonlinear function. All the transformed motifs are then *max-pooled* across the sentence to derive an EMR-level feature vector. Finally, a *linear classifier* is placed at the top layer for prediction. The entire architecture of Deepr can be summarized as a function $f(r)$ for record r:

$$f(r) \leftarrow \text{Class}\left(\text{Pool}\left\{\text{ReLU}\left(\text{Conv}\left[\text{Embed}\left\{\text{Seq}\left(r\right)\right\}\right]\right)\right\}\right) \quad (1)$$

The CNN plays a crucial role as it detects *clinical motifs* that are predictive. Clinical motifs are co-occurrences of diseases (also known as comorbidity), disease progression, patterns of disease/treatment, and patterns of collocating treatments [21]. However, as CNN is supervised it requires labels, which may not always be available (e.g., new patients with short history). A possible enhancement is through pretraining the embedding layer through a powerful tool known as *word2vec* [34]. As word2vec is unsupervised and relies on local collocation patterns, clinical motifs can be pre-detected, and then further refined through CNN with supervising signals.

[1] http://apps.who.int/classifications/icd10/browse/2016/en

[2] http://www.who.int/classifications/ichi/en/

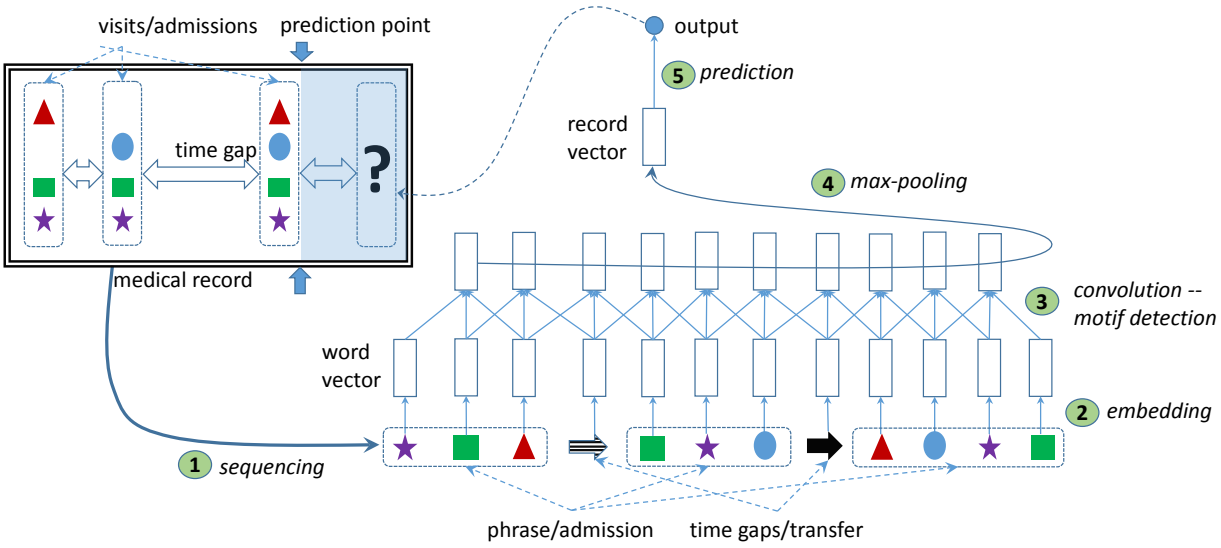

Figure 1. Overview of `Deepr` for predicting future risk from medical record. Top-left box depicts an example of medical record with multiple visits, each of which has multiple coded objects (diagnosis & procedure). The future risk is unknown (question mark (?)). *Steps from-left-to-right*: (1) Medical record is sequenced into phrases separated by coded time-gaps/transfers; then *from-bottom-to-top*: (2) Words are embedded into continuous vectors, (3) local word vectors are convoluted to detect local motifs, (4) max-pooling to derive record-level vector, (5) classifier is applied to predict an output, which is a future event. Best viewed in color.

B. Sequencing EMR

This task refers to transforming an EMR into a sentence, which is essentially a sequence of words. We present here how the words are defined and arranged in the sentence.

Recall that an EMR is a sequence of time-stamped visit episodes. Each episode may contain many pieces of information, but for the purpose of this work, we focus mainly on diagnoses and treatments (which involve clinical procedures and medications). For simplicity, we do not assume perfect timing of each piece, and thus an episode is a finite set of discrete words (diagnoses and treatments). The episode is then sequenced into a phrase. The order of the element in the phrase may follow the pre-defined ordering by the EMR system, for example, primary diagnosis is placed first, followed by secondary diagnoses, followed by procedures. In absence of this information, we may randomize the elements.

Within an episode, occasionally, there are one or more transfers between care providers, for example, separate departments from the same hospital, or between hospitals. In these cases, an admission is a phrase, and an episode is a subset of phrases separated by a transfer event. We create a special word `TRANSFER` for this event. Between two consecutive episodes, there is a time gap, whose duration is generally randomly distributed. We discretize the time gap into five intervals, measured in months: (0-1], (1-3], (3-6], (6-12], 12+. Each interval is assigned a unique identifier, which is treated as a word. For example, `0-1m` is a word for the (0-1] interval gap. With these treatments, an EMR is a sentence of phrases separated by words for transfers or time gaps. The phrases are ordered by their natural time-stamps. For robustness, infrequent words are coded as `RAREWORD`.

The following is an example of a sentence, where diagnoses are in ICD-10 format (a character followed by digits), and procedures are in digits:

```
1910 Z83 911 1008 D12 K31 1-3m R94 RAREWORD H53
Y83 M62 Y92 E87 T81 RAREWORD RAREWORD 1893 D12
S14 738 1910 1916 Z83 0-1m T91 RAREWORD Y83 Y92
K91 M10 E86 6-12m K31 1008 1910 Z13 Z83.
```

Here the phrases are: [`1910 Z83 911 1008 D12`], [`R94 RAREWORD H53 Y83 M62 Y92 E87 T81 RAREWORD RAREWORD 1893 D12 S14 738 1910 1916 Z83`], [`RAREWORD Y83 Y92 K91 M10 E86`], and [`K31 1008 1910 Z13 Z83`]. The time separators are: [`1-3m`], [`0-1m`], and [`6-12m`]. Note that within each phrase, the ordering of words has been randomized.

C. Convolutional Net

c) Embedding: The first step when applying convolutional nets on a sentence is to represent discrete words as continuous vectors. One way is to use the so-called one-hot coding, that is, each word is a binary vector of all zeros, except for just one position indexed by the word. However, this representation creates a high-dimensional vector, which may lead to overfitting and expensive computation. Alternatively, we can use *word embedding*, which refers to assigning a dense continuous vector to a discrete word. For example, the second word [`Z83`] in the example above may be assigned to 3D vector as (0.1 -2.3 0.5). In practice, we maintain a look-up table indexed by words, i.e., $E(w) \in \mathbb{R}^m$ is the vector for word w. The embedding table E is learnable. Applying word embedding to the sentence yields a sequence of vectors, where the vector at position t is $\boldsymbol{x}_t = E(w_t)$.

d) Convolution: On top of the word embedding layers is a *convolutional layer*. Each convolution operation reads a

sliding window of size $2d+1$ and produces p *filter responses* as follows:

$$\boldsymbol{z}_t = \text{ReLU}\left(\boldsymbol{b} + \sum_{j=-d}^{d} W_j \boldsymbol{x}_{t+j}\right) \quad (2)$$

where $\boldsymbol{z}_t \in \mathbb{R}^p$ is filter response vector at position t, $W_j \in \mathbb{R}^{p\times m}$ is the convolution kernel at relative position j (hence, $W \in \mathbb{R}^{p\times m\times(2d+1)}$), $\boldsymbol{b}$ is bias, and $\text{ReLU}(\boldsymbol{x}) = \max\{\boldsymbol{0}, \boldsymbol{x}\}$ (element-wise). When it is clear from the context, we use "filter" to refer to the learnable device that detects motifs, which are manifestation of filters in real data. The rectified linear function enhances strong signals and eliminates weak ones. The bias $\boldsymbol{b}$ and the kernel tensor W are learnable.

e) Pooling: Once the local filter responses are computed by the convolutional layer, we need to *pool* all the responses to derive a global sentence-level vector. We apply here the max-pooling operator:

$$\bar{\boldsymbol{z}} = \max_t \{\boldsymbol{z}_t\} \quad (3)$$

where the max is element-wise. Thus the pooled vector $\bar{\boldsymbol{z}}$ lives in the same space of $\mathbb{R}^p$ as filters responses $\{\boldsymbol{z}_t\}$. Like the rectifier used in Eq. (2), this max-pooling further enhances strong signals across the words in the sentence.

f) Classifier: The final layer of `Deepr` is a classifier that takes the pooled information and predicts the outcome: $f(r) = \text{classifier}(\bar{\boldsymbol{z}}(r))$ for record r. The main requirement is that the classifiers must allow gradient to propagate down to lower layers. Examples include a linear classifier (e.g., logistic regression) or a non-linear parametric classifier (e.g., neural network).

D. Training

`Deepr` has multiple trainable parameters: embedding matrix, biases, convolution kernels, and classifier-specific parameters. As the number of trainable parameters is often large, it necessitates regularizers such as weight shrinkage (e.g., via ℓ_2 norm) or dropouts [43] . For training we also need to specify a loss function, which depends on the nature of classifiers. For example, for binary outcome (e.g., readmission), logistic classifier is usually trained on cross-entropy loss. Training starts with (random) initialization of parameters which are then refined through back-propagation and stochastic gradient descent (SGD). This requires gradients with respect to trainable parameters. Gradient computation is often tedious and erroneous, but it is now fully automated in modern deep learning frameworks such as Theano [3] and Tensorflow [1]. For SGD, parameters are updated after every mini-batch of records (or sentences). Training is stopped after a pre-defined number of epochs (iterations), or on convergence.

g) Pretraining with word2vec: As mentioned in Sec. III-A, the embedding matrix can be pretrained using *word2vec*. Here we do not need labels, and thus we can exploit a large set of unlabeled data.

E. Model Inspection and Visualization

`Deepr` facilitates intuitive model inspection and visualization for better understanding:

h) Identifying motif responses in a sequence: For each motif detector, the motifs response at position t (e.g., $\boldsymbol{z}_t \in \mathbb{R}^p$) can be used to identify and visualize strong motifs. For size-3 motifs, the response weight to a size-3 sub-sequence $(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t, \boldsymbol{x}_{t+1})$ of a sequence $\boldsymbol{x}$ is the term $\sum_{j=-d}^{d} W_j \boldsymbol{x}_{t+j}$ in Eq. (2), which is the dot product of the sub-sequence and the kernel W.

i) Identifying frequent and strong motifs: Motifs with large responses in sequences are collected. From this collection, we keep frequent motifs representative for each outcome class.

j) Computing word similarity: Through embedding $\boldsymbol{x}_w = E(w)$, word similarity can be computed easily, e.g., through cosine $S(w,v) = \boldsymbol{x}_w^\top \boldsymbol{x}_v \left(\|\boldsymbol{x}_w\| \, \|\boldsymbol{x}_v\|\right)^{-1}$.

k) Visualization of similar patients: Patient vectors from Eq. (3) can be used to compute patient similarity. This enables retrieving patients who have similar history and similar future risk likelihood. This is unlike existing methods that compute only similar history, which does not necessarily guarantee similar future. Further, the similarity is not heuristic, and it does not require a heuristic combination of multiple data types (such as diseases and interventions). Fig. 2, for example, shows the distribution of positive and negative classes, in which patient vectors are projected onto 2D using t-SNE [47]. Patients who have similar history and future will stay close together.

l) Visualization in disease/intervention space: Since words are embedded into vectors, visualization in 2D is through dimensionality reduction tools such as PCA or t-SNE [47].

IV. Implementation

In this section, we document implementation details of `Deepr` on a typical EMR system. For ease of exposition, we assume that diseases are coded in ICD-10 format, but other versions are also applicable with minimal changes.

A. Data and Evaluation

Data was collected from a large private hospital chain in Australia in the period of July 2011 – December 2015. The data is coded according to Australian Coding Standard (ACS). The ACS dictates that diagnosis coding is based on ICD-10-AM[3], an Australian adaptation to WHO's ICD-10 system. Likewise, procedure coding follows ACHI (Australian Classification of Health Interventions). The data consists of 590,546 records (300K unique patients), each corresponds to an admission (defined by an admission time and a discharge time).

The data subset for testing `Deepr` was selected as follows. First we identified 4,993 patients who had at least an unplanned readmission within 6 months from a discharge, regardless of the admitting diagnosis. This constituted the risk group. For each risk case, we then randomly picked a control case from the remaining patients. For each risk/control group, we used 830 patients for model tuning, 830 for testing and the rest for training. A discharge (except for the last one in risk group) is randomly selected as prediction point, from which the future risk will be predicted. See also Fig. 1 for a graphical illustration.

[3]https://www.accd.net.au/Icd10.aspx

B. Implementation Details of Deepr

m) Episode definition: Deepr assumes that episodes are well-defined with an admission time and discharge time. However, it is not always the case due to intra-hospital or inter-hospital transfers. Our implementation links two admissions into an episode if they are separated by less than 12 hours, or by 12-24 hours but with documented transfer.

n) Words: For robustness, only level 3 ICD-10-AM codes are used. For example, *F20.0* (paranoid schizophrenia) would be converted into *F20* (schizophrenia). Similarly, the procedures are converted into procedure blocks. Rare words are those occurring less than 100 times in the database.

o) Word order randomization: For motifs detection, randomization is necessary to generate many potential motifs. We also test a special case where words in a phrase are ordered starting with the primary diagnosis followed by other secondary diagnoses, then by procedures in their natural ordering as defined by the EMR system.

p) Sentence length: For CNN, the sentences are trimmed to keep the last min(100, len(sentence)) words. This is to avoid the effects of some patients who have very long sentences which severely skew the data distribution. In a typical EMR, this is equivalent to accounting for up-to 10 visits per patient, which cover more than 95% of patients.

q) Hyper-parameter tuning: Deepr has a number of hyper-parameters pre-specified by model users: embedding dimension m, kernel window size $2d+1$, motif size, number of motifs n per size, number of epochs, mini-batch size, and other classifier-specific settings. Some hyper-parameters can be found through grid search, which finds the best configuration with respect to the accuracy on the development set.

We searched for the best parameters using the training and development data. Then we used the model with the best parameter to predict the unseen test data. The best parameters settings were $m = 100$, $d = 1$, motif size = 3, 4 and 5, $n = 100$ number of epochs = 10, mini-batch size = 64, ℓ_2 regularization $\lambda = 1.0$.

C. Baselines

We implemented the bag-of-words representation and regularized logistic regression (BoW+LR). LR has a parameter C that helps control overfitting. We searched for the best parameter C using the development data. We used the model with the best parameter to predict the unseen test data. We found the best parameter $C = 0.1$, which is equivalent to a prior Gaussian of mean 0 and standard deviation of 0.333.

V. Results

A. Risk Prediction

We predict unplanned readmission within 6 months after a random index discharge. Table I reports the prediction accuracy for all methods, when trained on data with and without coded time-gaps. Time-gaps coding improves the BoW-based prediction, suggesting the importance of proper sequential handling. However, time-gaps do not affect the accuracy of Deepr. This might be due to the convolution, rectification and max-pooling operations (see Sec. III-C), which pick the most powerful convoluted signals in the sequence. The use of *word2vec* to initialize the embedding matrix also has little contribution toward the accuracy. This could be because *word2vec* looks only for local collocations in both directions (past and future), whereas the prediction in Deepr is more global and of longer time horizon only in the future direction. In either cases with and without *word2vec*, Deepr is superior than the baseline BoW+LR.

Method	*W/o time*	*With time*
BoW + LR	0.727	0.741
Deepr (rand init)	**0.754**	**0.753**
Deepr (*word2vec* init)	**0.750**	**0.756**

Table I
ACCURACY ON 6-MONTH UNPLANNED READMISSION PREDICTION FOLLOWING A RANDOM INDEX DISCHARGE WITH AND WITHOUT TIME-GAPS. RAND INIT REFERS TO RANDOM INITIALIZATION OF THE EMBEDDING MATRIX. *Word2vec* INIT REFERS TO PRETRAINING THE EMBEDDING MATRIX USING THE *word2vec* ALGORITHM [34].

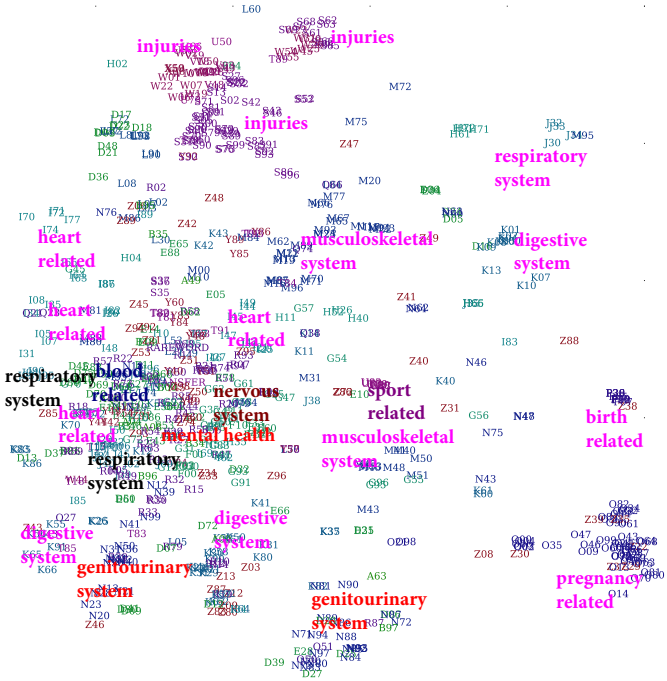

Figure 3. Distribution in the disease space, projected into 2D using t-SNE. Distribution of interventions is omitted for clarity. Best viewed in color.

Fig. (2) shows how Deepr groups similar patients and creates a more linear decision boundary while BoW+LR scatters the patient distribution and has a more complicated decision boundary. Recall that Deepr creates the feature vectors using element-wise max-pooling over all the motifs responses, as in Eq. (3). This demonstrates that the motifs, not just individual words, are important to computing similarity between patients. This also suggests that given a new patient Deepr is better at querying similar patients in the database when future risk is needed.

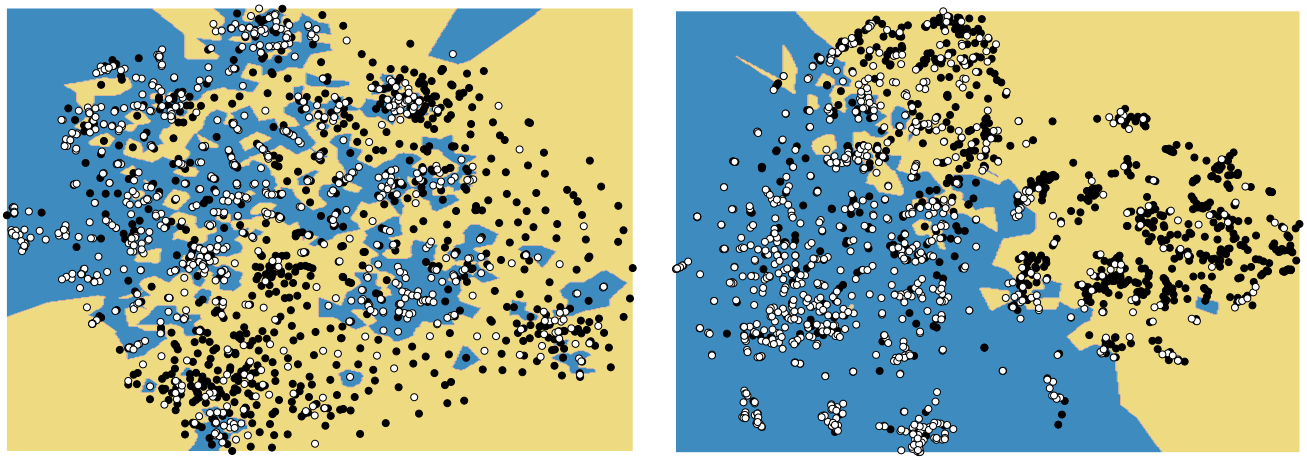

Figure 2. 2D projections of classification on the unseen test set of two methods BoW+LR and Deepr. White points and blue background are negative class, black point and yellow region are positive class. The figure shows Deepr groups similar patients and creates a more linear decision boundary while BoW+LR scatters the patient distribution and has a more complicated decision boundary. The decision boundary is approximated by an exhaustive contouring method, where fine lattice points of the background grid are labeled to the predicted label of their nearest data point, and then the boundary is computed by the contouring algorithm. Best viewed in color.

B. Disease/Procedure Semantics

Recall that Deepr first embeds words into a vector space. This offers a simple but powerful way to uncover and visualize the underlying structure of the word space (see Sec. III-E). Fig. 3 plots the distribution of diseases on 2D. Deepr discovers disease clusters which partly correspond to nodes in the ICD-10 hierarchy. Apart from pregnancy, child birth issues and injuries, the conditions are not totally separately suggesting a complex dependencies in the disease space. The main bock of the disease space has conditions related to heart, blood, metabolic system, respiratory system, nervous system and mental health. A more close examination of most similar conditions to a disease is given in Table II. For example, similar to cesarean section delivery of baby are those related to pregnancy complications (disproportion, failed induction of labor, or diabetes) and corresponding delivery procedures (cesarean section, manipulating fetal presentation, forceps).

We note in passing that we also obtained a similar visualization using only *word2vec* as in [34], which is known to detect hidden semantic relationships between words. Deepr trained on the embedding matrix initialized by *word2vec* did not significantly change the relative positions of words. This suggests that Deepr also captures the semantic relationship between words.

C. Filter Responses and Motifs

While the semantics in the previous sub-section reveal the global relative relation between diseases and procedures, they do not explain local interactions (e.g., motifs). Here we compute the local filter responses per sentence, and from there, a collection of strong and frequent motifs is derived.

Table III shows some sentences with strong responses for Filter 1 and 4 for both risk and no-risk class. It can be seen that the sub-sequences Z85.1163.1910 and 1066.1067.I21 respond strongly for the positive class and contribute to the classification result. The first sub-sequence is about cancer history (Z85), biopsy procedure (1163) and cerebral anesthesia (1910). The other sub-sequence is about heart attack (I21) and kidney-related procedures (1066 and 1067).

Filter ID	Response within a (sub) sentence
1 (readmit)	Z08 . Z85 . 1163 . 1910 . 1089
1 (no-risk)	1744 . 1910 . D24
4 (readmit)	N18 . 668 . I25 . N13 . 1910 . 905 . 1066 . 1067 . I21 . N17 . 607 . 1910 . 1910 . 1067 . 671 . R19
4 (no-risk)	1089 . Z08 . Z85 . Z86 . 6-12m . Z08 . 1910 . Z85 . Z86 . E11 . 1089 . 12-99m . 1910 . Z08

Table III
SOME SENTENCES WITH STRONG RESPONSES FOR FILTERS 1 AND 4. CODE WITH FIRST LETTER IS DIAGNOSIS, CODE WITH ALL NUMBERS IS PROCEDURE, CODE ENDS WITH "M" IS TIME-GAP. THE HEIGHTS OF THE CODES ARE PROPORTIONAL TO THEIR RESPONSE WEIGHTS. THE SUB-SEQUENCE Z85.1163.1910 AND 1066.1067.I21 RESPONSE STRONGLY TO THE POSITIVE CLASS.

From strong and frequent filter responses in all sentences, we derive the list of motifs. Table IV lists the motifs with largest weights and highest frequency of occurrence for code chapter E, I and O. The first motif of Filter 45 shows the pattern that treatment removing toxic substances from the blood co-occurred with care involving dialysis and readmission within 1 month. The second motif in the same row discovers the pattern that type-I diabetes patients involve in education about information and management of diabetes. The third motif in the same row shows type-II diabetes patients readmit within 1-3 months. Filter 26 demonstrates the co-occurrence of diseases related to diabetes. The three motifs show that type-II diabetes patients can have complications such as heart failure, vitamin D deficiency and kidney failure. Filters 10 and 35 show diseases and treatments related to the circulatory system, whereas pregnancy and birth related motifs are shown in Filters 2 and 33 in the last two rows.

Single delivery by cesarean section	*Type 2 diabetes mellitus*	*Atrial fibrillation and flutter*
Diagnoses:	**Diagnoses**:	**Diagnoses**:
Maternal care for disproportion	Personal history of medical treatment	Paroxysmal tachycardia
Placenta praevia	Presence of cardiac/vascular implants	Unspecified kidney failure
Complications of puerperium	Personal history of certain other diseases	Cardiomyopathy
Failed induction of labor	Unspecified diabetes mellitus	Shock, not elsewhere classified
Diabetes mellitus in pregnancy	Problems related to lifestyle	Other conduction disorders
Procedures:	**Procedures:**	**Procedures:**
Cesarean section	Cerebral anesthesia	Insertion or removal procedures on aorta
Medical or surgical induction of labour	Other digital subtraction angiography	Electrophysiological studies [EPS]
Manipulation of fetal presentation	Examination procedures on uterus	Other procedures on atrium
Other procedures associated with delivery	Medical or surgical induction of labour	Coronary artery bypass - other graft
Forceps delivery	Coronary angiography	Coronary artery bypass - saphenous vein

Table II
RETRIEVING TOP 5 SIMILAR DIAGNOSES AND PROCEDURES.

Filter ID	*Motifs*		
45	**0-1m 1060 Z49** Time-gap Haemoperfusion Care involving dialysis	**1916 E10 Z45** Allied health intervention, diabetes education Type 1 diabetes mellitus Adjustment and management of drug delivery or implanted device	**1-3m E11 Z45** Time-gap Type 2 diabetes mellitus Adjustment and management of drug delivery or implanted device
26	**E11 I48 I50** Type 2 diabetes mellitus Atrial fibrillation and flutter Heart failure	**E11 E55 I48** Type 2 diabetes mellitus Vitamin D deficiency Atrial fibrillation and flutter	**E11 I50 N17** Type 2 diabetes mellitus Heart failure Acute kidney failure
10	**1893 I48 K35** Exchange transfusion Atrial fibrillation and flutter Acute appendicitis	**1005 A41 I48** Panendoscopy to ileum with administration of tattooing agent Other sepsis Atrial fibrillation and flutter	**1-3m I48 Z45** Time-gap Atrial fibrillation and flutter Adjustment and management of drug delivery or implanted device
35	**1909 727 I83** Intravenous regional anesthesia Interruption of sapheno-femoral and sapheno-popliteal junction varicose veins Varicose veins of lower extremities	**1620 I83 L57** Excision of lesion(s) of skin and subcutaneous tissue of foot Varicose veins of lower extremities Skin changes due to chronic exposure to nonionising radiation	**1910 768 I83** Sedation Transcatheter embolisation of other blood vessels Varicose veins of lower extremities
2	**D68 O80 Z37** Other coagulation defects Single spontaneous delivery Outcome of delivery	**1344 O75 O80** Other suture of current obstetric laceration or rupture without perineal involvement Other complications of labor and delivery Single spontaneous delivery	**1344 O75 O82** Other suture of current obstetric laceration or rupture without perineal involvement Other complications of labour and delivery Single delivery by caesarean section
33	**1333 1340 O09** Neuraxial block during labour and delivery procedure Emergency lower segment caesarean section Duration of pregnancy	**1340 O14 Z37** Emergency lower segment caesarean section Gestational [pregnancy-induced] hypertension with significant proteinuria Outcome of delivery	**1340 3-6m O34** Emergency lower segment caesarean section Time-gap Maternal care for known or suspected abnormality of pelvic organs

Table IV
RETRIEVING 3 MOTIFS FOR EACH OF THE 6 FILTERS WHICH HAVE LARGEST WEIGHTS AND MOST FREQUENT WITH CODE CHAPTER O, I AND E.

VI. DISCUSSION

We have presented `Deepr`, a new deep learning architecture that provides an *end-to-end* predictive analytics in healthcare services. `Deepr` reads directly from raw medical records and predicts future outcomes. This departs from the traditional machine learning that relies on expensive manual feature extraction. `Deepr` learns to extract meaningful features by itself without expert supervision. This translates to uncovering the predictive local motifs in the space of diseases and interventions. These capacities are not seen in existing methods.

r) Significance: `Deepr` contributes to the growing literature of predictive medicine in multiple ways. First, it is able to uncover the underlying space of diseases and interventions, showing the relationships between them. The largest disease cluster in Fig. 3 suggests that diseases may interact in a complex way, and current representation of disease hierarchies such as those in ICD-10 may not reflect the true nature of medical disorders. Second, `Deepr` detects predictive motifs of comorbidity, care patterns and disease progression. The motifs suggest a new look into the complex interactions between diseases and between the diseases and cares. Third, similar patients can be retrieved not just using past history, but from likelihood of future risks as well. This would, for example, help to quickly identify an effective treatment regime based on similar patients who responded well to the treatment, or to alert the care team of a potential risk based on similar patients who had these before. Finally, `Deepr` predicts the future risk for a patient and explains why (through means of motifs responses), which is the core of modern prospective healthcare.

With these capabilities, `Deepr` can enable targeted monitoring, treatments and care packaging. This is highly important for chronic disease management that requires an on-going care and evaluation. For health services, a high predictive accuracy

of risk will lead to better resources prioritizing and allocation. For patients, accurate risk estimation is an important step toward personalized care. Patients and family will be promoted to become more aware of the conditions and risk, leading to proactive health management and help seeking. `Deepr` is generic and it can be implemented on existing EMR systems. This will enable innovative healthcare practices for better efficiency and outcomes to occur. For example, doctors, when seeing a patient, may consult the machine for a second opinion, with a *transparent, evidence-based reasoning*. Because they do not miss any piece of information in the database, they are less likely to overlook important signals.

s) Comparison to recent work on medical records: Deep learning in healthcare has recently attracted great interest. The most popular application is medical imaging using CNNs [8], motivated by the recent successes in cognitive vision [12], [22], [25]. However, there has been limited work on non-cognitive modalities. On time-series data (e.g., ICU measurements), the main difficulty is the handling missing data with recent work of [4], [23], [28], [38]. In [23], time-series are modeled using autoencoders (an unsupervised feedforward net) to discover meaningful phenotypes. In [4], [28], recurrent nets are used, and in [38], a convolutional net is employed. `Deepr` can be applied on these data, following a discretization of continuous signals into discrete words (e.g., through cut-points).

On routine medical records, `Deepr` is the only method that employs convolutional nets but there exist alternative architectures. Feedforward nets have been used [26], [10], [35]. Recurrent neural networks (RNN) on medical records include Doctor AI [6] and DeepCare [37]. Doctor AI is a RNN adapted for medical events, where both next events and time-gaps are predicted. DeepCare is a sophisticated model that represents time-gaps using a parametric model. Similar to our observation, the authors of DeepCare also noticed an interesting analogy between natural languages and EMR, where EMR is similar to a sentence, and diagnoses and interventions play the role of nouns and modifiers. While DeepCare is powerful on long records, it is less effective in short records, e.g., those with only one or two admissions. `Deepr`, on the other hand, does not suffer from this limitation. Stochastic deep neural nets such as deep Boltzmann machines are used in [32]. Deep non-neural nets have also been suggested in [13]. These methods are likely to be expensive to train and produce prediction.

Embedding of medical concepts has been proposed in contemporary work [5], [7], [37], [45]. In [7], medical concepts are embedded using *word2vec* [33], ignoring time gaps. The *Med2Vec* in [5] extends *word2vec* to embed visits. Both *word2vec* and *Med2Vec* model local collocations, but do not explicitly model motifs (with precise relative positions). In [45], a global model known as *eNRBM* embeds patients into vectors via regularized nonnegative restricted Boltzmann machines [36]. Local motifs are not modeled and and variable record length and time gaps are not properly handled. Discovering local motifs by means of convolutions has been suggested in [48] through matrix factorization. However, the work does not do prediction.

t) Limitations and future work: There are rooms for future work. First, long-term dependencies are simply captured through a max-pooling operation. This is rather simplistic due to a complex dynamic between care processes and disease processes [37]. A better model should pool information that is time-sensitive (e.g., recent events are more important to distant ones). At present, `Deepr` works exclusively on recorded events such as diagnoses and interventions. Integration with clinical narrative would be highly useful because rich information is buried in unstructured text. This can be done in the same framework of `Deepr` because of the sequential nature of text. Our evaluation has been limited to a common risk known as unplanned readmission. However, `Deepr` is not limited to any specific type of future risk. It can be well applied to predicting the onset or progression of a disease.

REFERENCES

[1] Martın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] Ognjen Arandjelović. Discovering hospital admission patterns using models learnt from electronic hospital records. *Bioinformatics*, page btv508, 2015.

[3] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. Theano: A CPU and GPU math compiler in Python. In *Proc. 9th Python in Science Conf*, pages 1–7, 2010.

[4] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *arXiv preprint arXiv:1606.01865*, 2016.

[5] Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, and Jimeng Sun. Multi-layer representation learning for medical concepts. *KDD*, 2016.

[6] Edward Choi, Mohammad Taha Bahadori, and Jimeng Sun. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. *arXiv preprint arXiv:1511.05942*, 2015.

[7] Youngduck Choi. Learning low-dimensional representations of medical concepts. *Proceedings of the AMIA Summit on Clinical Research Informatics (CRI)*, 2016.

[8] Dan C Cireşan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 411–418. Springer, 2013.

[9] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537, 2011.

[10] Joseph Futoma, Jonathan Morris, and Joseph Lucas. A comparison of models for predicting early hospital readmissions. *Journal of biomedical informatics*, 56:229–238, 2015.

[11] Danning He, Simon C Mathews, Anthony N Kalloo, and Susan Hutfless. Mining high-dimensional administrative claims data to predict early hospital readmissions. *Journal of the American Medical Informatics Association*, 21(2):272–279, 2014.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[13] Ricardo Henao, James T Lu, Joseph E Lucas, Jeffrey Ferranti, and Lawrence Carin. Electronic Health Record Analysis via Deep Poisson Factor Models. *JMLR*, 2016.

[14] Rui Henriques, Cláudia Antunes, and Sara C Madeira. Generative modeling of repositories of health records for predictive tasks. *Data Mining and Knowledge Discovery*, pages 1–34, 2014.

[15] G.E. Hinton and R.R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[16] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[17] Zhengxing Huang, Xudong Lu, and Huilong Duan. Latent treatment pattern discovery for clinical processes. *Journal of medical systems*, 37(2):1–10, 2013.

[18] Christopher H Jackson, Linda D Sharples, Simon G Thompson, Stephen W Duffy, and Elisabeth Couto. Multistate Markov models for disease progression with classification error. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(2):193–209, 2003.

[19] Anders Boeck Jensen, Pope L Moseley, Tudor I Oprea, Sabrina Gade Ellesøe, Robert Eriksson, Henriette Schmock, Peter Bjødstrup Jensen, Lars Juhl Jensen, and Søren Brunak. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature communications*, 5, 2014.

[20] Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012.

[21] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1106–1114, 2012.

[23] Thomas A Lasko, Joshua C Denny, and Mia A Levy. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS one*, 8(6):e66341, 2013.

[24] Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[25] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[26] Zhaohui Liang, Gang Zhang, Jimmy Xiangji Huang, and Qmming Vivian Hu. Deep learning for healthcare decision making with EMRs. In *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, pages 556–559. IEEE, 2014.

[27] Tsungnan Lin, Bill G Horne, Peter Tino, and C Lee Giles. Learning long-term dependencies in NARX recurrent neural networks. *IEEE Transactions on Neural Networks*, 7(6):1329–1338, 1996.

[28] Zachary C Lipton, David C Kale, and Randall Wetzel. Directly Modeling Missing Data in Sequences with RNNs: Improved Classification of Clinical Time Series. *arXiv preprint arXiv:1606.04130*, 2016.

[29] Chuanren Liu, Fei Wang, Jianying Hu, and Hui 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*, pages 705–714. ACM, 2015.

[30] Christopher D Manning. Computational linguistics and deep learning. *Computational Linguistics*, 2015.

[31] Jason Scott Mathias, Ankit Agrawal, Joe Feinglass, Andrew J Cooper, David William Baker, and Alok 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*, 20(e1):e118–e124, 2013.

[32] Saaed Mehrabi, Sunghwan Sohn, Dingheng Li, Joshua J Pankratz, Terry Therneau, Jennifer L St Sauver, Hongfang Liu, and Mathew Palakal. Temporal pattern and association discovery of diagnosis codes using deep learning. In *Healthcare Informatics (ICHI), 2015 International Conference on*, pages 408–416. IEEE, 2015.

[33] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. word2vec, 2014.

[34] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[35] Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6, 2016.

[36] T.D. 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.

[37] Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. DeepCare: A Deep Dynamic Memory Model for Predictive Medicine. *arXiv preprint arXiv:1602.00357*, 2016.

[38] Narges Razavian and David Sontag. Temporal convolutional neural networks for diagnosis from lab tests. *arXiv preprint arXiv:1511.07938*, 2015.

[39] Patrick Royston and Willi Sauerbrei. Interactions between treatment and continuous covariates: a step toward individualizing therapy. *Journal of Clinical Oncology*, 26(9):1397–1399, 2008.

[40] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[41] Mansour TA Sharabiani, Paul Aylin, and Alex Bottle. Systematic review of comorbidity indices for administrative data. *Medical care*, 50(12):1109–1118, 2012.

[42] Ralph Snyderman and R Sanders Williams. Prospective medicine: the next health care transformation. *Academic Medicine*, 78(11):1079–1084, 2003.

[43] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[44] Truyen Tran, Wei Luo, Dinh Phung, Sunil Gupta, Santu Rana, Richard L Kennedy, Ann Larkins, and Svetha Venkatesh. A framework for feature extraction from hospital medical data with applications in risk prediction. *BMC bioinformatics*, 15(1):6596, 2014.

[45] Truyen Tran, Tu Dinh Nguyen, Dinh Phung, and Svetha Venkatesh. Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM). *Journal of biomedical informatics*, 54:96–105, 2015.

[46] Truyen Tran, Dinh Phung, Wei Luo, and Svetha Venkatesh. Stabilized sparse ordinal regression for medical risk stratification. *Knowledge and Information Systems*, 2014. DOI: 10.1007/s10115-014-0740-4.

[47] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[48] F. Wang, N. Lee, J. Hu, J. Sun, and S. Ebadollahi. Towards heterogeneous temporal clinical event pattern discovery: a convolutional approach. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 453–461. ACM, 2012.

[49] Xiang Wang, David Sontag, and Fei Wang. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 85–94. ACM, 2014.

[50] DRGHR Williams and GE Hinton. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

[51] Xiang Zhang and Yann LeCun. Text understanding from scratch. *arXiv preprint arXiv:1502.01710*, 2015.