Journal of Biomedical Informatics 97 (2019) 103256

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Readmission prediction using deep learning on electronic health records

Awais Ashfaq[a,b,*], Anita Sant'Anna[a], Markus Lingman[b,c], Sławomir Nowaczyk[a]

[a] Center for Applied Intelligent Systems Research, Halmstad University, Sweden
[b] Halland Hospital, Region Halland, Sweden
[c] Institute of Medicine, Dept. of Molecular and Clinical Medicine/Cardiology, Sahlgrenska Academy, University of Gothenburg, Sweden

ARTICLE INFO

Keywords:
Electronic health records
Readmission prediction
Long short-term memory networks
Contextual embeddings

ABSTRACT

Unscheduled 30-day readmissions are a hallmark of Congestive Heart Failure (CHF) patients that pose significant health risks and escalate care cost. In order to reduce readmissions and curb the cost of care, it is important to initiate targeted intervention programs for patients at risk of readmission. This requires identifying high-risk patients at the time of discharge from hospital. Here, using real data from over 7500 CHF patients hospitalized between 2012 and 2016 in Sweden, we built and tested a deep learning framework to predict 30-day unscheduled readmission. We present a cost-sensitive formulation of Long Short-Term Memory (LSTM) neural network using expert features and contextual embedding of clinical concepts. This study targets key elements of an Electronic Health Record (EHR) driven prediction model in a single framework: using both expert and machine derived features, incorporating sequential patterns and addressing the class imbalance problem. We evaluate the contribution of each element towards prediction performance (ROC-AUC, F1-measure) and cost-savings. We show that the model with all key elements achieves higher discrimination ability (AUC: 0.77; F1: 0.51; Cost: 22% of maximum possible savings) outperforming the reduced models in at least two evaluation metrics. Additionally, we present a simple financial analysis to estimate annual savings if targeted interventions are offered to high risk patients.

1. Introduction

Unscheduled readmissions are a hallmark of Congestive Heart Failure (CHF), with 1 in 4 patients being readmitted within 30 days of discharge [1]. Readmissions are problematic because they pose additional economic burden on the healthcare system and put patients at risk of hospital-acquired infections and clinical errors [2]. They are also considered as a proxy by authorities (like the Centre for Medicare and Medicaid Services in the US) to measure care quality since readmissions are often related to premature discharge or improper treatment [3]. Precise prediction of readmission risk can support care-providers to decide if a patient is ready for discharge or should be considered for an intervention program, eventually reducing the number of unscheduled readmissions and curbing healthcare cost [4]. Put differently, an important decision related to the readmission problem is made by care-providers at the time of patient discharge, and rests on a simple prediction challenge: *what is the readmission probability of the patient at the time of discharge?*

A recent review investigated 60 studies with 73 models to predict unscheduled 30-day readmissions [5]. It reported moderate discrimination ability. Common limitations include building a predictive model that:

- Uses either only human-derived features [6,7] or machine-derived features [8,9]. The former discards a huge proportion of information in each patient's record, while the latter ignores knowledge and guidelines coming from human intelligence.
- Ignores the sequential or temporal trajectory of events embedded in Electronic Health Records (EHRs) [6,10–13]. EHRs include a sequence of measurements (clinical visits) over time which contains important information about the progression of disease and patient state.
- Fails to consider the skewness in terms of class imbalance and different costs of misclassification errors [14,6,12,13,15,8]. Class imbalance problems are common with EHR data [16]. A favourable prediction model is often the one with a high precision on the minority class with a reasonable precision on the majority class.

Multiple studies have attempted to address the aforementioned challenges inherent to EHRs [17,18]. Researchers have used word-embedding techniques [8] and auto-encoders [19] to automatically generate clinical representations by considering all clinical codes

* Corresponding author at: Center for Applied Intelligent Systems Research, Halmstad University, Sweden.
E-mail address: awais.ashfaq@hh.se (A. Ashfaq).

https://doi.org/10.1016/j.jbi.2019.103256
Received 17 October 2018; Received in revised form 24 June 2019; Accepted 22 July 2019
Available online 24 July 2019

(diagnoses, procedures, medications and labs) related to individual patients. Another study compounded these embeddings with human-selected features to enhance clinical representations [20]. Similarly, some studies have documented the use of Recurrent Neural Networks (RNN) and its variants to capture the sequential information embedded in EHRs [21–24]. Addressing the class imbalance problem, a recent study advocated for a modified loss function that biases the decision boundary of the model in favour of the minority (or readmission) class [11]. Though these studies have independently shown improved performance on various prediction tasks; to the best of our knowledge, there is no single model that addresses all the three limitations mentioned earlier. The contribution of this work is twofold:

- We present a predictive model that is cost-sensitive and leverages both human and machine-derived features in a sequential manner. We also present a direct approach, originally proposed in [25], to generate patient representations that are fed as input to the prediction model.
- We present a financial analysis to estimate possible cost-savings if the prediction model is implemented in real clinical workflow. This will support policy-makers in healthcare and facilitate the implementation of cost-effective interventions.

2. Background

This study lies at the intersection of three mature research fields with rich literature: representation learning, long short-term memory (LSTM) networks and cost-sensitive classification. Herein we present a brief overview of these techniques applied on EHRs. We begin with a short introduction to the EHR structure.

2.1. EHR representation

A patient in an EHR is often represented as a sequence of care visits. These include visits to primary care, specialist outpatient care, emergency care and hospital admissions. The information in each visit can be broadly categorized into demographics (of patient and care provider) and clinical state of the patient. Demographic features include patient age, gender, place, type of visit and more. The clinical state in EHR is represented as a list of clinical codes and values pertaining to diagnoses, procedures, lab results, vitals or medication recorded in that visit. The state can be numerically represented as a scalar severity score, or a feature vector.

Clinical severity scores are ubiquitous in medical literature. They are built to include a handful of ad hoc features to quantify the clinical state or severity of the patient [26,27]. Clinical feature vectors are generated as high-dimensional sparse binary vectors [28], or low-dimensional dense vectors using Natural Language Processing (NLP) inspired word embedding techniques [29] or auto-encoders [19]. An advantage of word-embedding's over auto-encoders and binary vectors is that the former preserves contextual similarities among representations. Word embeddings have been extensively used in recent years to represent clinical concepts in EHR [30–32,9,23]. Intuitively the learning is based on the co-occurrence of clinical codes in a single profile. Put differently, codes with similar neighbours are considered to have similar meaning. This is referred to as *contextual similarity*.

Having the clinical feature vector or severity score, a visit vector is often generated by concatenating the former with visit demographics [20]. Finally a patient is represented by concatenating corresponding visit vectors.

2.2. Sequential modelling

To input a patient representation while capturing the sequence of visits requires model architecture equipped with memory and capable of handling variable size inputs. RNNs qualify as a logical choice. Originally designed for time series data and NLP applications, RNNs have garnered significant attention for modelling the sequential nature of EHRs [17,18]. RNNs belong to the family of artificial neural networks (ANN) with recurrent connections in hidden layer [33,18]. Operationally, the hidden state h_t is sequentially updated depending on both the activation of the current input x_t at time t, and the previous hidden state of the layer h_{t-1}. Popular RNN variants include the long short-term memory (LSTM) and gated recurrent unit (GRU) models [34]. Contrary to traditional RNNs, LSTMs and GRUs contain an internal recurrence loop along with three and two gates respectively to control information flow. Gated RNNs have shown to capture long-term dependencies in data and overcome the vanishing gradient problem [35]. A young popular extension to train RNNs on long historical patterns is attention-based-learning [36]. It is loosely based on our understanding of visual attention: humans (at a time) focus more on a particular region of an image than others for further processing. This is because not all parts of the information are often equally relevant for the task in hand. Attention-based learning has been widely used for machine translation, image captioning and text summarization [37–39]. In the context of EHRs, [40–42] have used it together with gated RNNs to infer what historical information is pertinent for predicting future events.

2.3. Cost-sensitive classification

A dataset is referred to as *skewed* or *imbalanced* if some classes are highly under represented compared to others. General prediction algorithms assume that training sets have evenly distributed classes which - in case of skewed datasets - bias the algorithm towards the majority class. As a result the distribution of the minority class is not adequately learned [43]. Class imbalance problems (CIP) are troublesome in fields where correctly predicting the minority class is often significant than the other. For instance, in the clinical domain, a false-negative HIV result prior to renal dialysis can be catastrophic [44]. Despite the pertinence of CIP in real-world data, there has been little research in recent years [45]. Two often applied techniques in the realm of CIP include over-sampling of under-represented class in the training set [46] and embedding a cost matrix in the loss function that encodes the penalty of misclassifying samples from a particular (minority) class [11].

For instance, consider a neural network with softmax output layer and cross-entropy loss function $L = -\sum_i t_i \log y_i$ where t_i is the target class indicator. y_i is the output of the softmax function. The loss gradient with respect to the input of the sotmax layer z_i is given as $\frac{\partial L}{\partial z_i} = y_i - t_i$. If we add a cost weight $C(t_i)$ for each target class t_i, then the loss function is $-\sum_i C(t_i)^* t_i \log y_i$ and corresponding gradient is $C(t_i)^*(y_i - t_i)$. Put differently, if class A has larger cost C, and $y_i \neq t_i$ then the gradients computed from samples of class A will be larger which in turn will affect weight updates in favour of class A.

3. Methods and Materials

The overall framework is illustrated in Fig. 1.

3.1. Ethics

The study was conducted using real-world EHR data from southern Sweden with approval from the Ethics Committee in Lund, Sweden (Dnr. 2016/517). The data was anonymized by removing all identifiable features: names, dates of birth, addresses and telephone numbers prior to the study.

3.2. Data summary

We used data from a CHF population. A CHF patient was defined

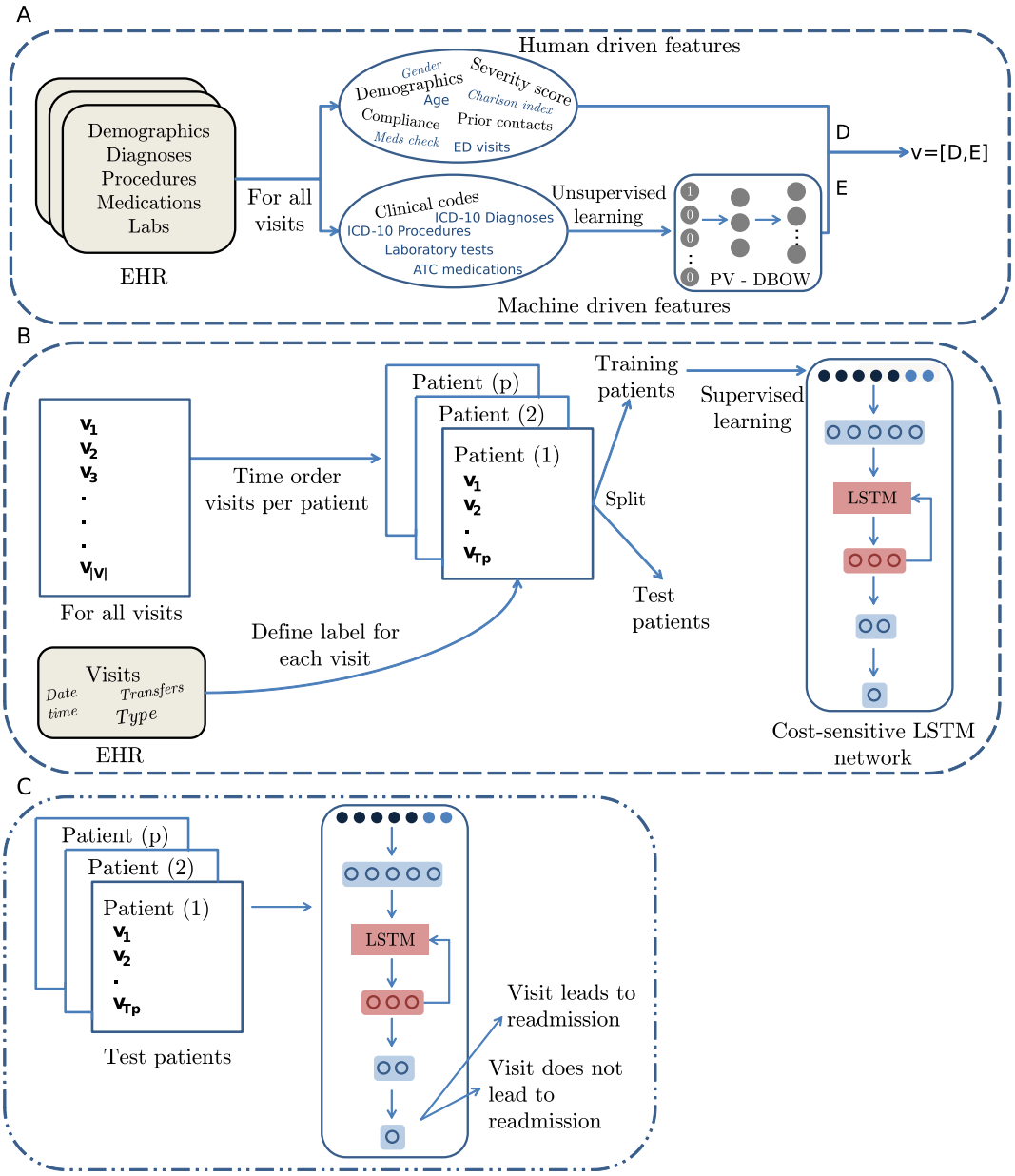

Fig. 1. Complete framework. (A) Generate visit representations from human and machine-derived features. The output is a feature vector for each visit. (B) The visit representations are fed sequentially (per patient) in a cost-sensitive LSTM network for training. (C) The test patients are fed to the trained network to predict 30-day readmission risk at each visit.

according to the guidelines[1] of Swedish Board of Health and Social Welfare. We included patients with at least one CHF related diagnosis and one hospital admission between 2012 and 2016. Only structured data pertaining to each patient was used in this work. Clinical notes and waveforms were not considered. Table 1 shows an overview of the population.

Table 1
Data summary.

Number of patients	7655
Gender distribution (female:male)	43:57
Mean, median age at the time of visits	78.8, 81
Total in-hospital visits or admissions	32,287
Total readmissions	9004
Mean duration of stay (days)	6.8
Mean clinical codes per visit	38
Min, max codes in visit	1, 121

3.3. EHR representation

Let P and V denote a set of all patients and visits where $|P|$ and $|V|$ are the total number of patients and visits in the dataset respectively. Then, for each patient $p \in P$, there exists T_p time-ordered visits $v_p \in V$: $v_{p1}, v_{p2}, ..., v_{pt}, ..., v_{pT_p}$. A visit v is represented via human-derived features (D) and machine-derived features (E). Herein we describe the two feature sources.

3.3.1. Human-derived features: D

Based on relevant medical literature, significant variables for readmission prediction include comorbidities, length of stay,

[1] I50 and all sub-codes; I11.0; I42 and all sub-codes except I42.1 and I42.2; I43 and all sub-codes.

Table 2
Human-derived features for each visit. The feature 'compliance' reports prescription for both ACE inhibitors and Beta Blockers. Types of visit include scheduled and unscheduled admissions. The prior visits were counted on a 6-month window starting from the date of admission.

Age at the time of visit	Discrete
Gender	Binary
Medication compliance	Binary
Total procedures performed	Discrete
Duration of stay	Discrete
Duration of all stays	Discrete
Type of visit	Binary
Charlson comorbidity score	Discrete
Number of prior emergency care visits	Discrete
Number of prior admissions	Discrete
Number of prior outpatient visits	Discrete

demographics, type of admission and medications [5,47,48]. Table 2 shows the list of features used to describe a visit in this study.

3.3.2. Machine-derived features: E

We considered all clinical codes related to diagnoses, procedures medications and lab tests. Diagnoses (including procedures) and medications in the EHR were represented according to standard schemas: ICD-10-SE[2] and ATC respectively.[3] It is worth adding that the schemas have codes in the order of thousands; many of which are very granular. Both ICD-10 and ATC codes follow a hierarchical alpha-numeric schema where the granularity of the disease, procedure or medicine decreases as we go from right to left. To reduce information overload and have a generalized specificity level, we grouped the codes into high-order categories by selecting the first three characters of each code. Grouping clinical codes has also been previously practiced in [22,40,49]. Laboratory records in the EHR did not follow a standard schema, rather recorded by the name of the test. Thus lab tests were not grouped; however, only tests that had an abnormal value were included when building the clinical state of the patient.

Let $S = \{s_1, s_2, ...s_{|S|}\}$ denote a set of all clinical codes in the EHR where $|S|$ is total number of unique codes. Each visit v consists of a subset of S, called C_v - a set of heterogeneous clinical codes corresponding to the visit. Here, we leveraged the Paragraph Vector for Distributed Bag of Words (PV-DBOW) [25] to generate numerical representations of C_v. Traditional document embedding tools are built for language texts and consider the order of the words using a fixed size sliding window. Clinical codes in visits, on the other hand, are unordered and each visit may have different numbers of codes. Thus, we extended PV-DBOW to support dynamic window size with respect to the size of C_v. The training objective of PV-DBOW is to maximise the probability of all clinical codes in a visit given a visit index v. Mathematically,

$$\arg\max_{\boldsymbol{E}} \prod_{v \in V} p(C_v|v) \tag{1}$$

$\boldsymbol{E}$ is the learned embedding matrix that stores numerical representations of each visit and was trained using mini-batch gradient descent to minimize the cross-entropy loss with Gensim 3.4.0 in Python [50]. The training steps are detailed in supplement.

3.3.3. Visit representation: v

We appended the embedding matrix $\boldsymbol{E}$ with 11 human-derived features $\boldsymbol{D}$ to get the visit representations $\boldsymbol{v} = [\boldsymbol{D}, \boldsymbol{E}]$.

3.3.4. Patient representation: p

Finally each patient $p \in P$, was represented by appending sequences of time-ordered visits $\boldsymbol{p} = [\boldsymbol{v_{p1}}, \boldsymbol{v_{p2}}, ..., \boldsymbol{v_{pT_p}}]$.

3.4. Sequential modelling

The LSTM network is depicted in Fig. 2. Given a patient p with T_p inpatient visits, the model accepts one visit at each time step. From $t = 2$ till $t = T_p$, new LSTM state is dependent on the state at the previous time step and the new input visit. Of note, at every time step, the LSTM block propagates its state to the next dense layers. This LSTM configuration is often referred to as *sequence to sequence* prediction. This means that at each time step, there is an input and output for the network. We detail the training steps involved in the supplement.

3.5. Cost-sensitive classification

Our EHR data has class imbalance ratio of 0.28. Thus, we add a cost term to our loss function $C(y'_{pt})$ that corresponds to the penalty of wrongly predicting y'_{pt}. We set the cost for misclassifying a readmission visit to be three times higher than the cost for misclassifying a non-readmission visit. To learn and optimize the parameters of the model, we set the binary cross entropy as the loss function and minimize with respect to weights and bias terms W.

$$\min_{\boldsymbol{W}} \sum_{p=1}^{|P|} \sum_{t=1}^{T_p} [-y'_{pt} log(y_{pt}) - (1 - y'_{pt}) log(1 - y_{pt})] \cdot C(y'_{pt}) \tag{2}$$

where y'_{pt} is the readmission indicator for the t^{th} visit of p^{th} patient where 1 indicates readmission and 0 control. The loss minimization and parameter (W) optimization was performed through back-propagation using mini-batch gradient descent implemented via Keras 2.2.2 [51]. We detail the hyper-parameters used and training steps in the supplement.

3.6. Defining labels

The target label 'readmission' is a binary variable, designated to each visit if a subsequent visit by the same patient (from any cause) occurred within 30 days of discharge from the former visit. Within hospital transfers or scheduled visits were not considered as readmission.

3.7. Cost savings analysis

We estimate the potential annual cost savings C_{saved} if an intervention I is selectively offered to patients at high risk of readmission according to Eq. (3).

$$C_{saved} = (C_r \cdot T_r \cdot I_{sr}) - (C_i \cdot P_r) \tag{3}$$

where C_r and C_i are the readmission and intervention cost per patient; and I_{sr} is the intervention success rate. T_r and P_r are the number of *truly predicted* and *all predicted* readmissions. T_r and P_r were calculated over a range of thresholds $\in (0, 1)$. The intervention cost is the cost spent on each patient that is predicted with a high readmission risk. Put differently, it includes true positives and false positives. Of note, the word *cost* is expressed in monetary terms here.

3.8. Evaluation

The model performance was assessed on predicting 30-day readmissions among hospitalized CHF patients. We conducted an iterative design of 12 experiments with possible combinations of model characteristics that are summarized in Table 3. The characteristics are not mutually exclusive. This means that a model can have more than one characteristic in a single experiment. We summarize the application of different characteristics in recent studies in the supplement. We used the area under the receiver operating characteristic curve (ROC-AUC),

[2] International Classification of Diseases, revision 10, Sweden.
[3] Anatomical Therapeutic Chemical classification system.

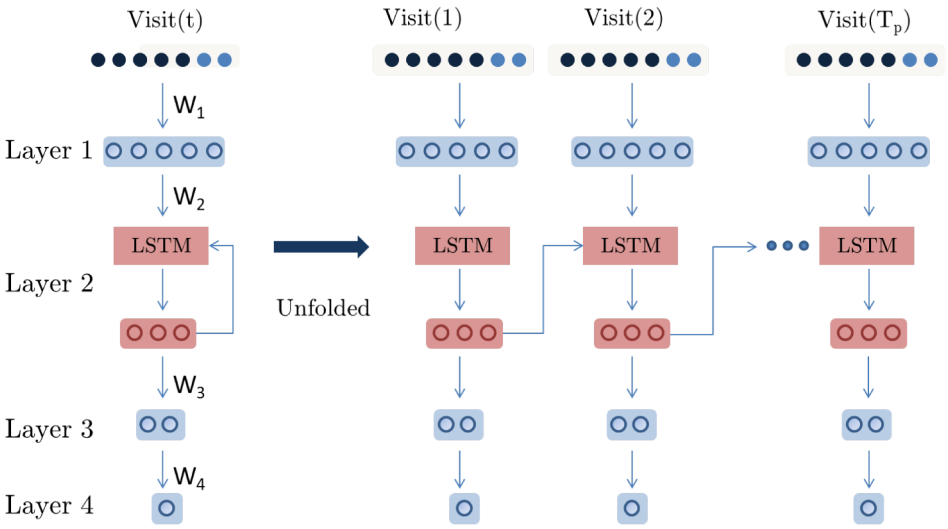

Fig. 2. Sequential modelling via LSTM. *W* include the weight matrices and bias terms between layers.

Table 3
Model characteristics explained.

Characteristic	Meaning
HDF	Human derived features are fed as input to the model
MDF	Machine derived contextual embeddings are fed as input to the model
LSTM	The model captures the sequential visit patterns in the EHR
CA	The model adjusts for misclassification costs

together with F1-scores and cost-savings for evaluation. The discrimination threshold θ used to classify readmission/no-readmission visits was selected based on maximum cost-savings (Eq. (3)). Here we consider a telemonitoring intervention program where C_i is 20% of C_r and $I_{sr} = 0.5$[52]. From financial records, C_r was found to be around SEK 48,000 in 2016 in Halland Sweden. Correspondingly, F1-scores are reported along with cost savings as a fraction of maximum savings. Maximum savings $= I_{sr} *(C_r - C_i)* A_r$, where A_r are all actual readmissions. Later, iterating through a range of θ levels, we plot the maximum annual savings for different values of C_i and I_{sr}.

4. Results

All experiments were performed on a machine with Intel Xeon(R) E5-2660 v2-4790 K CPU, 16 GB RAM. The source codes are publicly available.[4]

We split the data into training and test sets in two different ways. In case 1, we split based on patients. 70% patients were used for training and 30% for testing and we report the performance on the test set. In case 2, we split based on time. We used all patient data from 2012 to 2015 for training and tested on the complete set 2012–2016. However, we evaluate model performance on visits that occurred in 2016 only. We used the complete dataset for testing because most patients in the data have visits across different years, and this prior patient information is significant to predict readmissions in the following year. In both cases we repeated the experiment 10 times and report the mean AUC, F1-score and cost savings.

In the first case, we make sure that no patient overlaps between training and test set. Thus, the model is tested on *new* patients that weren't part of training. This is the usual way of evaluating readmission predictions in previous studies [8,15]. However, practically if a model is used in a clinical workflow, it would require predicting readmission risk for new patients (with no prior admission) and patients with prior admissions. The model performance on future visits by the latter patients can be improved if we include their prior data in the training set. Thus we expect better performance if data is split according to case 2. However, the results from case 2 are slightly less generalizable (or more data-specific) than case 1.

Table 4 shows case 1 performance for predicting 30-day readmissions with respect to different model characteristics: adding machine-derived and human-derived features; capturing the sequential aspect via LSTM and adjusting for prediction costs. Given the model with all four characteristics, Fig. 3 compares it with each reduced model (one characteristic ignored). Table 5 evaluates if the decrease in either metric is statistically significant. We show that ignoring LSTM or MDF results in significant ($p < 0.05$) decrease across all metrics. Ignoring HDF, results in a significant decrease in ROC-AUC and cost savings. Adjusting for prediction cost (CA) results in no significant change in the ROC-AUC, however, it improves F1-score and cost-savings.

We also analysed some attributes of correct and incorrect predictions made by the complete model. At a threshold θ that corresponded with maximum cost-savings for the hospital, the predicted risk scores were classified into true positives (TP), false negatives (FN), false positives (FP) and true negatives (TN). Table 6 suggests that the model tends to give high risk scores to sicker (and slightly aged) patients with frequent and long visits. In general, we see much similar pattern between TP and FP visits, and also between TN and FN visits. We suggest two probable explanations for the similarities between TP and FP visits. First, FP visits might also be at risk of readmission but not necessarily within 30 days. They might be at a readmission risk within 35 days or similar. Second, FP visits might be considered for a readmission-preventing intervention such as a care-counselling or home nurse visit that could have reduced the risk of readmission. However, we do not include the information of existing intervention programs in our prediction model. FN visits are at a risk of readmission, though their basic characteristics seem similar to TN visits. A probable explanation might be that the cause of readmission was not necessarily CHF or a related diagnose but something different, such as bone fracture or similar. Thus, we believe that an in-depth investigation using even richer information such as clinical notes will help understand and overcome the pitfalls of the prediction model and yield an even higher AUC. In addition to predictive performance, we also look into individual characteristics of high and low readmission risk visits and the change in risk score when a particular feature is altered. Fig. 4 suggests that the model gives a higher readmission risk score for males and the risk increases with

[4] https://github.com/caisr-hh/lstm_predict.

Table 4
Model performance on predicting 30-day readmission on test patients. The model was trained on 70% of CHF patients and tested on the rest.

Model characteristic				Metric (std. deviation)		
HDF	*MDF*	*LSTM*	*CA*	*ROC-AUC*	*F1-score*	*Cost-savings*
0	1	0	0	0.54 (0.009)	0.09 (0.070)	0.00 (0.003)
0	1	0	1	0.55 (0.010)	0.00 (0.061)	0.00 (0.001)
0	1	1	0	0.75 (0.014)	0.48 (0.006)	0.18 (0.012)
0	1	1	1	0.75 (0.003)	0.50 (0.008)	0.20 (0.013)
1	0	0	0	0.60 (0.005)	0.26 (0.063)	0.03 (0.015)
1	0	0	1	0.60 (0.001)	0.23 (0.107)	0.03 (0.016)
1	0	1	0	0.72 (0.006)	0.44 (0.008)	0.12 (0.013)
1	0	1	1	0.73 (0.009)	0.46 (0.010)	0.15 (0.018)
1	1	0	0	0.61 (0.005)	0.31 (0.028)	0.04 (0.013)
1	1	0	1	0.61 (0.010)	0.31 (0.021)	0.04 (0.014)
1	1	1	0	0.76 (0.005)	0.49 (0.007)	0.19 (0.011)
1	1	1	1	0.77 (0.005)	0.51 (0.008)	0.22 (0.010)

respect to age. We also notice that 5 days increase in the duration of hospital stay reduces the readmission risk for sicker patients while increases the risk for less sick patients. This is consistent with [53] in which unnecessary prolonged hospitalization was associated with increased health risk and hospital-related complications. It also supports [54] in which an additional hospitalization day for acute CHF patients was found to reduce readmission risk by 7%. Of note, we only alter a single feature in this analysis while keeping the remaining visit vector intact. In practical settings, additional hospitalization days might also result in additional lab exams, procedures or medications which will influence the readmission risk.

Table 7 shows case 2 performance to predict 30-day readmissions. We use this model to estimate annual savings in Halland, Sweden. Fig. 5 shows potential annual savings (in million SEK) for different intervention costs and success rates. For each C_i and I_{sr} pair, we calculated the cost saved over a range of θ levels and report the maximum savings. Correspondingly, for that value of θ, we report the fraction of high risk patient visits considered for a readmission-preventing intervention.

5. Discussion

This study presents a deep learning model to predict 30-day readmission in patients with CHF at the time of discharge. The model achieves high discrimination ability with AUC 0.77 and F1-score 0.51. We show that leveraging both human and machine-derived features from EHR, together with an LSTM network outperforms models that ignore any of these characteristics. Following practical validation to assure efficacy, the proposed model has several benefits if used as a decision support system in hospitals. From a clinical perspective, patients at risk of 30-day readmission may benefit from personalized discharge planning, care counselling or be considered for home nurse visits. From an economic perspective, a high precision model (Eq. (3): more true positives and fewer false positives) will have significant cost savings if interventions prove effective. This will also facilitate precision resource utilization in hospitals.

Table 5
T-test output: P-values comparing the evaluation metrics of the complete model (CA + LSTM + MDF + HDF) with others.

Model	ROC-AUC	F1-score	Cost-savings
CA + LSTM + MDF	< 0.05	0.09	< 0.05
CA + LSTM + HDF	< 0.05	< 0.05	< 0.05
CA + HDF + MDF	< 0.05	< 0.05	< 0.05
LSTM + HDF + MDF	0.10	< 0.05	< 0.05

In addition to the domain contribution, this study also highlights key findings from a data science perspective. We show that for all combinations of LSTM and cost adjustments (Table 4), combining hand-derived features with machine-derived features results in better performance than using either of them alone. Though deep learning approaches have shown to eliminate the need for feature selection in continuous data types like images [55], these approaches are somehow dependent on feature selection from heterogeneous, discrete or categorical data types like EHRs, supporting [20,56]. A recent study [9] hypothesized that appending human-derived demographical features to embeddings might not significantly improve the model performance because some basic features like *gender* are indirectly captured in the clinical codes. For instance, the code Z-37 (normal delivery) would imply a female. Here, we added human-derived features and found a significant 3% increase in the model AUC compared to a model that relied only on embeddings. Despite the improvement, we don't completely negate the hypothesis set in the previous study [9]. Consider the two models, with and without LSTM block regardless of cost adjustments, trained using machine-derived features. Adding human-derived features to the former resulted in a 3% increase in AUC whereas, for the latter, we saw a 11% increase in AUC. This is expected because some human-derived features (like prior admissions) are automatically captured by the LSTM model. Thus adding the feature explicitly did not improve the model performance comparatively.

We also show that an important piece of information in EHR is embedded in the sequential trajectory of patient which is consistent with previous studies [24,21]. Capturing the visits sequentially (as they

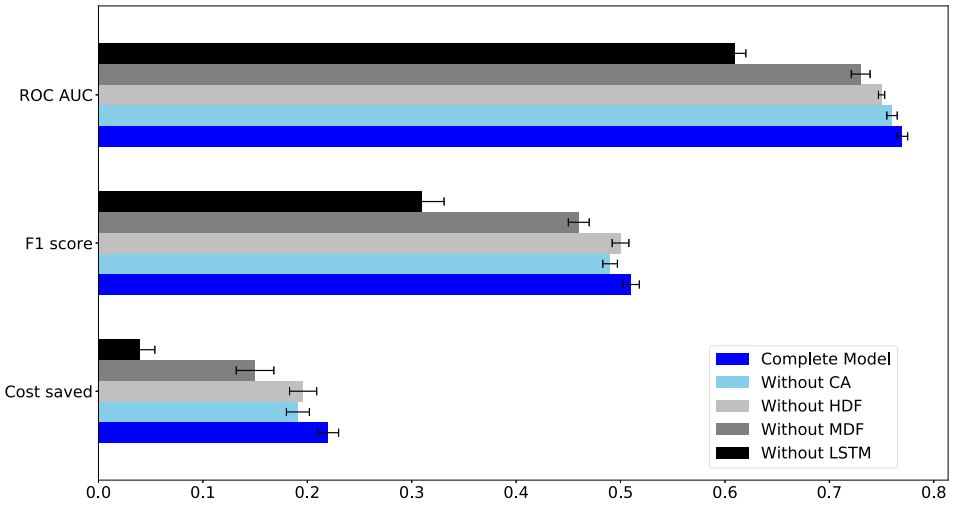

Fig. 3. Performance comparison of the complete model with reduced ones.

Table 6
Characteristics of TP, FN, FP, and TN test patient visits. Reporting Mean (std. deviation).

Metric	*TP*	*FN*	*FP*	*TN*
Age	78.82 (10.67)	80.20 (11.16)	80.80 (9.36)	79.93 (11.28)
Duration of stay (days)	9.36 (23.09)	6.92 (27.90)	7.52 (10.81)	7.07 (31.30)
No. of prior admissions	9.13 (9.54)	2.87 (5.76)	7.01 (6.39)	3.61 (4.76)
Charlson score	3.19 (1.97)	1.66 (1.36)	2.88 (1.86)	2.16 (1.73)
No. of diagnoses	7.74 (2.98)	5.62 (2.56)	7.32 (2.75)	6.25 (2.99)
No. of medications	5.07 (5.21)	4.38 (4.01)	4.49 (4.88)	3.77 (4.31)
No. of procedures	1.01 (1.46)	1.02 (1.61)	0.87 (1.36)	1.06 (1.55)
No. of labs	26.71 (14.96)	17.75 (12.74)	22.93 (13.66)	20.58 (15.01)

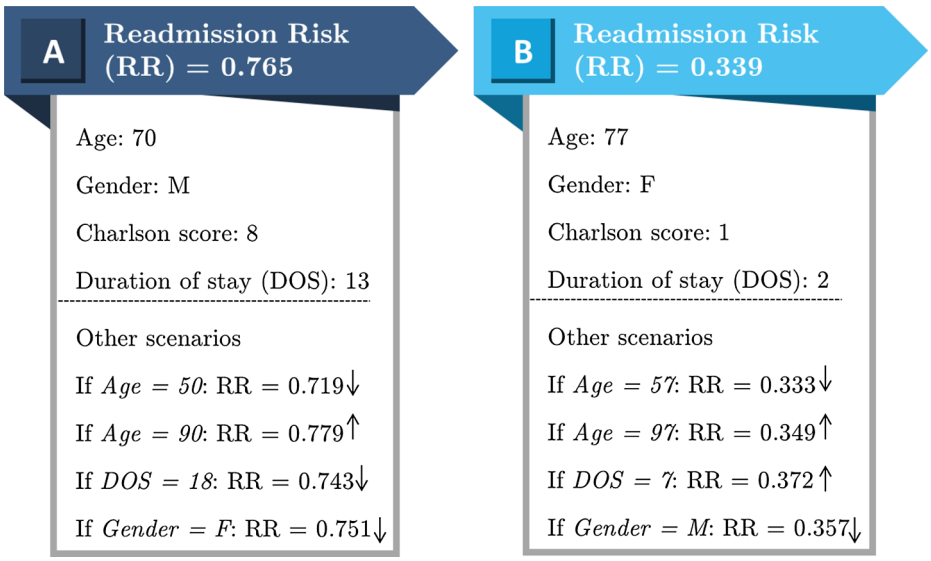

Fig. 4. Sample visits.

Table 7
Evaluating the complete model under different experiment settings.

Data split	Metric (std. deviation)		
	ROC-AUC	*F1-score*	*Cost-savings*
1: Trained on 70% patients, tested on 30%	0.77 (0.005)	0.51 (0.008)	0.22 (0.010)
2: Trained 2012–2015 data, tested on 2016 data	0.83 (0.006)	0.64 (0.005)	0.43 (0.005)

appear in EHR) adds a significant 26% rise in the AUC compared to a memory-less neural network like multilayer perceptron.

We found that adjusting for class imbalance does not contribute in terms of AUC. This is inconsistent with [11] that proposed a cost-sensitive deep learning model for readmission prediction. We suspect that the improved model performance in that study was because of using both static and continuous EHR features in an ANN and not because of cost adjustments. However, it requires further investigation. Despite the overall AUCs being similar for complete and the CA reduced model, Fig. 6 illustrates that the complete model performs better in the high sensitivity range (at the cost of high FPR) than its reduced version and vice versa in the low sensitivity range (low FPR). In this case, the prediction models should be rated depending on the context of the clinical situation and thus a portion of the overall ROC area may need to be considered. In this setup, we target *low risk/low cost* interventions where higher sensitivities are preferred and thus the complete model would be useful.

Recently, a study leveraged the FHIR[5] standard to represent patient events and built several models including a generic 30-day readmission prediction model reporting an AUC of 0.76 [15]. However, the proposed framework is complex and computationally intensive since it captures complete patient records including clinical notes. Following the new EU's General Data Protection Regulations (GDPR) [57], accessing complete healthcare datasets, especially clinical notes, is complicated because their anonymization is challenging compared to other structured data in EHRs [58]. This limits the use of clinical notes for research purposes. In this study, we show that even without including clinical notes, our model achieves similar performance on predicting 30-day readmissions. However, the fact that our model was trained and tested on a specific (CHF) population is reiterated.

Fig. 5 facilitates in defining cost-effective interventions based on estimated annual savings and underlying model performance. This result is interesting from an economic standpoint because it also defines the utility of the prediction model against different interventions. For instance, given an intervention I with $C_i = 1000$ and $I_{sr} = 0.9$, the economic utility of the proposed model is limited. This is because the maximum cost saved is when every visit is given that intervention. Though it is practically challenging to design such a cost-effective intervention, it is important to know that the economic benefit of prediction models depends on underlying interventions. Moreover, having an intervention of interest, the annual cost savings is another evaluation metric for readmission prediction models.

In this study, we highlight different characteristics of the prediction

[5] Fast Healthcare Interoperability Resources.

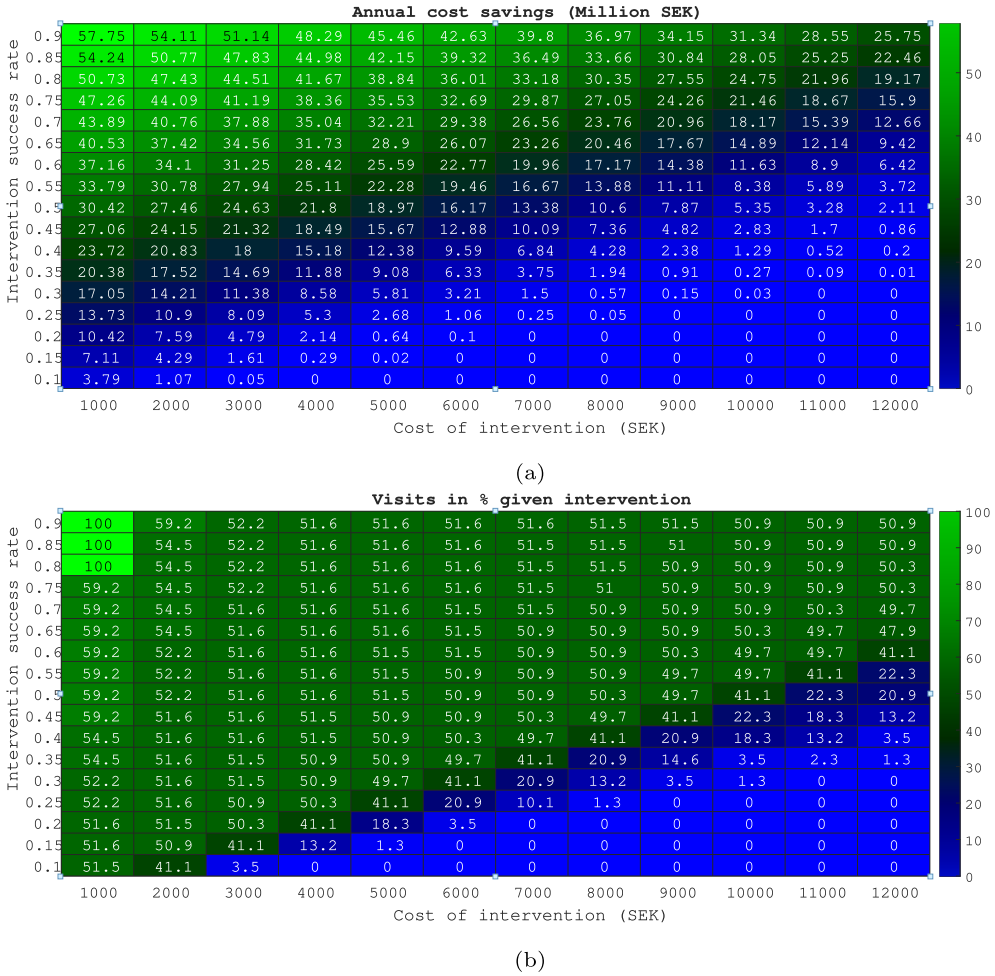

Fig. 5. Economic utility of the model. (a) The estimated annual savings (in Million SEK) for different intervention costs and success rates. (b) Corresponding percentage of total visits given readmission-preventing intervention based on the model output.

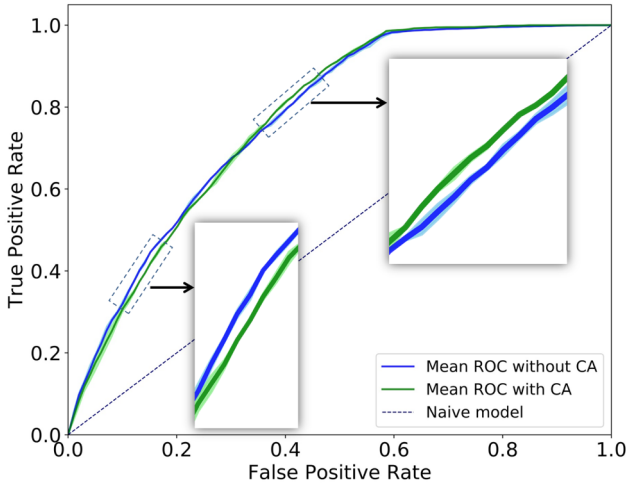

Fig. 6. Visual inspection of the ROC.

framework (Table 3) and attempt a single method for including each of them. For instance, we used LSTMs to model the sequential trajectory of patient visits. However, the sequential information can also be modelled in other ways such as via GRUs [8] or Convolutional Neural Nets [49], etc. Similarly, apart from paragraph embeddings, MDF can also be generated using aggregated word embeddings [24] or auto-encoders [19], etc. Regardless of how a particular characteristic is embedded in the prediction framework, we propose that they all should be considered when building readmission prediction models.

5.1. Limitations and future work

First, the model was trained and tested on data from a single region, Halland, Sweden. Halland has a homogeneous population with nearly 90%[6] of inhabitants being local residents. Thus similar performance cannot be guaranteed if the trained model is applied on EHRs from other parts of the globe. However, the training time of the model is fast (≈ 45 min) and can be fine-tuned on other EHRs provided they include similar features.

In this study, we ignored true laboratory values in EHR and only considered labs as a binary indicator. Several studies tend to ignore true lab values for EHR driven prediction tasks because their automatic integration to a deep learning model is challenging [9,19,24]. In future work, we will attempt to address this challenge.

A key limitation of NLP inspired embedding techniques is that they do not account for rare words in the vocabulary. Rare clinical codes like cancer or HIV can have significant importance if present in the patient profile. In this study, we indirectly capture these rare events using the Charlson score. However, future work will explore the use compositional character models to represent the complete list clinical codes [59].

Insights into future (like risk of readmission) allow informed

[6] http://www.citypopulation.de/php/sweden-admin.php?adm1id=13.

decision-making. However, to address the root cause of readmission and select effective interventions, it is important to understand what group of features contributed to the prediction. Future work will focus on leveraging attention based mechanisms that can infer features pertinent to the prediction score.

6. Conclusion

Unscheduled readmissions are a major source of burden for healthcare and timely identification of patients at risk of readmission can initiate relevant interventions to reduce unnecessary cost and improve care quality. Following an iterative design of experiments performed on real EHR data, this study targets key elements of an EHR driven prediction model and evaluates their contribution in terms of prediction AUC, F1-score and savings. The value of this study is a deep learning framework in which both human and machine derived features are fed sequentially in a cost-sensitive LSTM model to identify patients at risk of readmission at the point of discharge. The study also highlights potential annual cost savings if the model were implemented in a real hospital environment. As a next stage of the project, we aim to overcome the model limitations and conduct a feasibility study to implement the model in the clinical workflow.

Declaration of Competing Interest

The authors declare no conflict of interest in regards to the content published in this article.

Acknowledgements

We thank Antanas Verikas (PhD) and Stefan Lönn (MD, PhD) for providing methodological and clinical support during research.

Funding: The authors thank the European Regional Development Fund (ERDF), Health Technology Center and CAISR at Halmstad University and Hallands Hospital for financing the research work under the project *HiCube - behovsmotiverad hälsoinnovation*. We also acknowledge Group Inc. Consultants to Government and Industries (CGI) Sweden for setting up the data platform and granting legal access. Opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the aforementioned organizations.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jbi.2019.103256.

References

[1] K.E. Bergethon, C. Ju, A.D. DeVore, N.C. Hardy, G.C. Fonarow, C.W. Yancy, P.A. Heidenreich, D.L. Bhatt, E.D. Peterson, A.F. Hernandez, Trends in 30-day readmission rates for patients hospitalized with heart failure: findings from the get with the guidelines-heart failure registry, Circul.: Heart Fail. 9 (6) (2016) e002594.

[2] H.C. Felix, B. Seaberg, Z. Bursac, J. Thostenson, M.K. Stewart, Why do patients keep coming back? Results of a readmitted patient survey, Soc. Work Health Care 54 (1) (2015) 1–15.

[3] C.K. McIlvennan, Z.J. Eapen, L.A. Allen, Hospital readmissions reduction program, Circulation 131 (20) (2015) 1796–1803.

[4] S. Kripalani, C.N. Theobald, B. Anctil, E.E. Vasilevskis, Reducing hospital readmission rates: current strategies and future directions, Annu. Rev. Med. 65 (2014) 471–485.

[5] H. Zhou, P.R. Della, P. Roberts, L. Goh, S.S. Dhaliwal, Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review, BMJ Open 6 (6) (2016) e011060.

[6] S. Basu Roy, A. Teredesai, K. Zolfaghar, R. Liu, D. Hazel, S. Newman, A. Marinez, Dynamic hierarchical classification for patient risk-of-readmission, Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2015, pp. 1691–1700.

[7] B.A. Goldstein, A.M. Navar, M.J. Pencina, J. Ioannidis, Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review, J. Am. Med. Inform. Assoc. 24 (1) (2017) 198–208.

[8] C. Xiao, T. Ma, A.B. Dieng, D.M. Blei, F. Wang, Readmission prediction via deep contextual embedding of clinical concepts, PloS One 13 (4) (2018) e0195024.

[9] W. Farhan, Z. Wang, Y. Huang, S. Wang, F. Wang, X. Jiang, A predictive model for medical events based on contextual embedding of temporal sequences, JMIR Med. Inform. 4 (4).

[10] Y. Maali, O. Perez-Concha, E. Coiera, D. Roffe, R.O. Day, B. Gallego, Predicting 7-day, 30-day and 60-day all-cause unplanned readmission: a case study of a Sydney hospital, BMC Med. Inform. Decis. Making 18 (1) (2018) 1.

[11] H. Wang, Z. Cui, Y. Chen, M. Avidan, A.B. Abdallah, A. Kronzer, Predicting hospital readmission via cost-sensitive deep learning, IEEE/ACM Trans. Comput. Biol. Bioinform.

[12] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission, Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2015, pp. 1721–1730.

[13] M. Jamei, A. Nisnevich, E. Wetchler, S. Sudat, E. Liu, Predicting all-cause risk of 30-day hospital readmission using artificial neural networks, PloS One 12 (7) (2017) e0181173.

[14] S.B. Golas, T. Shibahara, S. Agboola, H. Otaki, J. Sato, T. Nakae, T. Hisamitsu, G. Kojima, J. Felsted, S. Kakarmath, et al., A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data, BMC Med. Inform. Decis. Making 18 (1) (2018) 44.

[15] A. Rajkomar, E. Oren, K. Chen, A.M. Dai, N. Hajaj, M. Hardt, P.J. Liu, X. Liu, J. Marcus, M. Sun, et al., Scalable and accurate deep learning with electronic health records, npj Digital Med. 1 (1) (2018) 18.

[16] J. Zhao, Learning predictive models from electronic health records, PhD thesis Department of Computer and Systems Sciences, Stockholm University, 2017.

[17] B. Shickel, P. Tighe, A. Bihorac, P. Rashidi, Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis, arXiv preprint arXiv:1706.03446.

[18] C. Xiao, E. Choi, J. Sun, Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review, J. Am. Med. Inform. Assoc.

[19] R. Miotto, L. Li, B.A. Kidd, J.T. Dudley, Deep patient: an unsupervised representation to predict the future of patients from the electronic health records, Scient. Rep. 6 (2016) 26094.

[20] E. Choi, M.T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, J. Sun, Multi-layer representation learning for medical concepts, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 1495–1504.

[21] Z.C. Lipton, D.C. Kale, C. Elkan, R. Wetzel, Learning to diagnose with lstm recurrent neural networks, arXiv preprint arXiv:1511.03677.

[22] E. Choi, M.T. Bahadori, A. Schuetz, W.F. Stewart, J. Sun, Doctor ai: predicting clinical events via recurrent neural networks, Machine Learning for Healthcare Conference, 2016, pp. 301–318.

[23] T. Pham, T. Tran, D. Phung, S. Venkatesh, Deepcare: a deep dynamic memory model for predictive medicine, Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2016, pp. 30–41.

[24] E. Choi, A. Schuetz, W.F. Stewart, J. Sun, Using recurrent neural network models for early detection of heart failure onset, J. Am. Med. Inform. Assoc. 24 (2) (2016) 361–370.

[25] Q. Le, T. Mikolov, Distributed representations of sentences and documents, International Conference on Machine Learning, 2014, pp. 1188–1196.

[26] M.E. Charlson, P. Pompei, K.L. Ales, C.R. MacKenzie, A new method of classifying prognostic comorbidity in longitudinal studies: development and validation, J. Chron. Diseases 40 (5) (1987) 373–383.

[27] E.F. Philbin, T.G. DiSalvo, Prediction of hospital readmission for heart failure: development of a simple risk score based on administrative data, J. Am. Coll. Cardiol. 33 (6) (1999) 1560–1566.

[28] R. Chen, H. Su, M. Khalilia, S. Lin, Y. Peng, T. Davis, D.A. Hirsh, E. Searles, J. Tejedor-Sojo, M. Thompson, et al., Cloud-based predictive modeling system and its application to asthma readmission prediction, AMIA Annual Symposium Proceedings, Vol. 2015 American Medical Informatics Association, 2015, p. 406.

[29] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Adv. Neural Inform. Process. Syst. (2013) 3111–3119.

[30] B.S. Glicksberg, R. Miotto, K.W. Johnson, K. Shameer, L. Li, R. Chen, J.T. Dudley, Automated disease cohort selection using word embeddings from electronic health records, Pac Symp Biocomput, Vol. 23 World Scientific, 2018, pp. 145–156.

[31] A.L. Beam, B. Kompa, I. Fried, N.P. Palmer, X. Shi, T. Cai, I.S. Kohane, Clinical concept embeddings learned from massive sources of medical data, arXiv preprint arXiv:1804.01486.

[32] E. Choi, A. Schuetz, W.F. Stewart, J. Sun, Medical concept representation learning from electronic health records and its application on heart failure prediction, arXiv preprint arXiv:1602.03686.

[33] R.J. Schalkoff, Artificial Neural Networks Vol. 1 McGraw-Hill, New York, 1997.

[34] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[35] S. Ayyar, O. Don, W. Iv, Tagging patient notes with icd-9 codes, Proceedings of the 29th Conference on Neural Information Processing Systems, 2016.

[36] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473.

[37] J.K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, Y. Bengio, Attention-based models for speech recognition, Adv. Neural Inform. Process. Syst. (2015) 577–585.

[38] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo, Image captioning with semantic attention, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

2016, pp. 4651–4659.
[39] M.-T. Luong, H. Pham, C.D. Manning, Effective approaches to attention-based neural machine translation, arXiv preprint arXiv:1508.04025.
[40] E. Choi, M.T. Bahadori, J. Sun, J. Kulas, A. Schuetz, W. Stewart, Retain: An interpretable predictive model for healthcare using reverse time attention mechanism, Advances in Neural Information Processing Systems, 2016, pp. 3504–3512.
[41] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, J. Gao, Dipole: diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks, Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2017, pp. 1903–1911.
[42] T. Baumel, J. Nassour-Kassis, M. Elhadad, N. Elhadad, Multi label classification of patient notes a case study on icd code assignment, arXiv preprint arXiv:1709.09587.
[43] N.V. Chawla, N. Japkowicz, A. Kotcz, Special issue on learning from imbalanced data sets, ACM Sigkdd Explor. Newslett. 6 (1) (2004) 1–6.
[44] Special report: The making of an hiv catastrophe, DAWN news.
[45] S.H. Khan, M. Hayat, M. Bennamoun, F.A. Sohel, R. Togneri, Cost-sensitive learning of deep feature representations from imbalanced data, IEEE Trans. Neural Networks Learn. Syst.
[46] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Artific. Intell. Res. 16 (2002) 321–357.
[47] C. van Walraven, I.A. Dhalla, C. Bell, E. Etchells, I.G. Stiell, K. Zarnke, P.C. Austin, A.J. Forster, Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community, Can. Med. Assoc. J. 182 (6) (2010) 551–557.
[48] J.B. Hamner, K.J. Ellison, Predictors of hospital readmission after discharge in patients with congestive heart failure, Heart Lung: J. Acute Crit. Care 34 (4) (2005) 231–239.
[49] P. Nguyen, T. Tran, N. Wickramasinghe, S. Venkatesh, Deepr: a convolutional net for medical records, IEEE J. Biomed. Health Inform. 21 (1) (2017) 22–30.
[50] R. Řehůřek, P. Sojka, Software Framework for Topic Modelling with Large Corpora, Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, ELRA, Valletta, Malta, 2010, pp. 45–50.
[51] F. Chollet et al., Keras, <https://keras.io>, 2015.
[52] A. Broderick, Partners healthcare: Connecting heart failure patients to providers through remote monitoring, The Commonwealth Fund, New York.
[53] D. Ruangkriengsin, P. Phisalprapa, Causes of prolonged hospitalization among general internal medicine patients of a tertiary care center, J. Med. Assoc. Thai 97 (3) (2014) 206–215.
[54] A.P. Bartel, C.W. Chan, H. Kim, Should hospitals keep their patients longer? The role of inpatient and outpatient care in reducing readmissions, Citeseer, 2014.
[55] A. Ashfaq, Segmentation of Cone Beam ct in Stereotactic Radiosurgery, Master's thesis, TRITA-STH, KTH, Sweden, 2016.
[56] X. Min, B. Yu, F. Wang, Predictive modeling of the hospital readmission risk from patients' claims data using machine learning: a case study on copd, Scient. Rep. 9 (1) (2019) 2362.
[57] M. Mostert, A.L. Bredenoord, M.C. Biesaart, J.J. van Delden, Big data in medical research and eu data protection law: challenges to the consent or anonymise approach, Eur. J. Hum. Genet. 24 (7) (2016) 956.
[58] X.-B. Li, J. Qin, Anonymizing and sharing medical text records, Inform. Syst. Res. 28 (2) (2017) 332–352.
[59] W. Ling, T. Luís, L. Marujo, R.F. Astudillo, S. Amir, C. Dyer, A.W. Black, I. Trancoso, Finding function in form: Compositional character models for open vocabulary word representation, arXiv preprint arXiv:1508.02096.