Convolutional Neural Network-based Model for Patient Representation Learning to Uncover Temporal Phenotypes for Heart Failure

Chongchao Zhao[1] Yichen Shen[2] Li-Pan Yao[1]
[1] MS Analytics, Georgia Institute of Technology
[2] MS Computer Science, Georgia Institute of Technology

Abstract

Electronic Health Records (EHR) are immense datasets filled with a wealth of all kinds of medical data for each patient for each medical visit. However, traditional methods of data analysis on EHR datasets prove impenetrable due to its size, dimensionality, and irregularity. Heart failure (HF) has frustrated caregivers due to difficulty in prediction as well as its nature as an overarching condition rather than a distinct phenotype. In this work, we propose to utilize convolutional neural networks (CNN) on EHR data, which would allow us to precisely predict risks of heart failure and efficiently extract informative representations of HF patients. The representation can help identify a new set of phenotypes with similar EHR profiles to facilitate treatment of patients with HF.

Presentation Video: https://youtu.be/KjQEyZH6I-M

1 Introduction

Currently, there are about 5.7 million adults in the United States suffering from heart failure (HF). HF has contributed around 1/9 of the deaths per year, and about half of the patients die within 5 years of diagnosis [Mozaffarian 2016]. Furthermore, HF costs the nation an estimated $30.7 billion each year [Heidenreich 2011]. Due to the high morbidity and mortality, an increasing prevalence, and the escalating healthcare costs associated with HF, there is ongoing effort in clinical research to identify high-risk patients in order to provide more precise prediction and apply customized treatment [Alba 2013]. Such predictions include differentiating HF patients with the reduced ejection fraction (rEF) phenotype which has ample effective treatments from the patients with preserved ejection fraction (pEF) which currently has no effective treatment options [Lindman 2017, Austin 2013].

Among the various efforts to this end, the data mining approach became increasingly popular. Modern data mining techniques such as advanced machine learning and the more recent deep-learning methods have particular merits. Through utilizing health information technology (HIT) infrastructure in the form of electronic health records (EHRs), many aspects of care including diagnosis, medication, laboratory results and imaging data could be captured [Jensen 2012]. Extracting meaningful representation of various groups of similar patients from electronic records, sometimes called computational phenotyping, would have a wide range of beneficial impact on facilitating clinical research in HF as well as other areas in health care.

However, mining useful information in EHR remains extremely challenging. The electronic record data are usually complex, high-dimensional, noisy, incomplete, and longitudinally irregular. Currently, Existing EHR driven studies in HF or other areas often rely on carefully engineered features chosen by domain experts [Smith 2011, Taslimitehrani 2016]. However, there are a couple of limitations to such analyses. First, these types of feature engineering require extensive domain knowledge and are sometimes costly and difficult to generalize. Second, the context and temporal relationship between different features are ignored in the setting of traditional HF research, which can sometimes contain crucial information to identify risk groups.

To cope with these limitations, we utilize a deep learning method to efficiently extract robust patient representation for HF risk prediction from EHR without manual feature engineering. The concept of our model is inspired by the Word2Vec model [Mikolov 2013] and Convolutional Neural Networks (CNN) in the context of natural language processing (NLP) [Kim 2014]. Specifically, via the Word2Vec model, single patient events are transformed to distributed embedding forms that contain hidden semantic relationships between different medical events. We can then utilize a CNN model taking in a sequence of patient events and predict the risk of HF. Owing to its special convolutional operation, CNN model can automatically learn latent structure in sequential relationships between events. Combining those two models we are able to extract the context and temporal relations in patient-level medical events from EHR and provide robust patient representation pertaining to HF risk prediction.

Since deep learning applications especially CNN in clinical studies is still in its nascency, there are only a few papers that incorporated CNN in EHR driven studies. Previous tentative research by Che et. al [2017] has highlighted the efficacy of the application of CNN and Word2Vec to EHR-driven prediction tasks, but their approach is coarse and still needs refinement. We improve their predictive model and further interpret patient-level representation learned by CNN in HF prediction. Our contributions are listed as follows:

1. The addition of time tokens indicating the elapsed time between two consecutive events as a feature for a patient.

2. The inclusion of demographic features to the hidden layer in CNN.

3. The interpretation and extraction of phenotypes from hidden representation learned by the CNN model.

2 Methodology

2.1 Feature Embedding Learning via Word2Vec

Since its publication in 2013, word2vec, developed by Google, has gained much popularity in multiple areas such as NLP and deep learning. It takes a text corpus as input and outputs word vectors. Learned representations are based on distance which calculates the similarity between words. It does not require labels, and with enough training data, the resulting word vectors have embeddings with intriguing characteristics in such a manner that words with similar meanings appear in the same clusters. EHRs are essentially time-series data comprising sequences of encounter episodes. For a given patient, we observe his/her events in a temporal order. For MIMIC-III dataset, we focused on diagnoses, procedures and prescriptions. In the context of word2vec, we extracted the information of ICD9 codes, procedure items, and drug names from a patients medical record and

arranged these words into a sequence which can be likened to a "sentence". Then word2vec is used to train this collection of sentences. As a result, each word in a sentence representing a patient will be converted to an array. These arrays could then be stacked to become a matrix which can be fed into a CNN. However, this process assumes that the intervals of hospital visits are uniformly distributed. To capture the heterogeneity of the intervals, we include an interval tag or "time token" to the records. For example, if time difference between two consecutive codes is approximately 3 days, we add an event token "3-7days" between them. We have discretized the observation window into several intervals associated with different tokens.

2.2 CNN Risk Prediction Model

Convolutional Neural Networks (CNN) is efficient in extracting underlying local structures. It has been widely used in computer vision and speech recognition area and achieved significant success. Its more recent application in NLP and video processing demonstrates its capability of extracting temporal and sequential information [Kim 2014, Karpathy 2014]. Hence, in temporally ordered EHR data, CNN is very suitable for mining hidden structure and local dependencies for each patients records that are useful for predicting diseases.

Specifically, we construct input for given patient p as a embedding matrix $X_p \in R^{n_p * d}$, where n_p is the number of records the patient have in the observation window and d is the dimension of embedding for each patients medication records, diagnosis records, and input tags. Note that the records are ordered temporally and the temporal dimension of of the matrix n_p is the maximum of number of events among all patients record (those patients records with less events are padded with dummy tokens). We then apply 1D convolutional operation over temporal dimensions of the matrix. Suppose the size of filter we use is F, the resultant vector from the convolutional layer is $n_p - F + 1$. Using combinations of filters of different lengths are proven to be beneficial in prediction results [Che 2017]. Thus we use K filters in different sizes, range from 2 to 5, to capture temporal length variations. After the convolutional layer, we use a max pooling layer that maps user output vector from each filter to a real number. We then create a demographic feature vector for each patient that encodes patients gender and age at the end of his/her observation window. By concatenating the pooling result across all filters, along with patients demographic vector, we will eventually get an dense patient representation in a vector of size $K + 2$. We then feed this representation into a fully connected network and the prediction result will be obtained from the outermost softmax layer. The full architecture of our prediction model is illustrated in Figure 1.

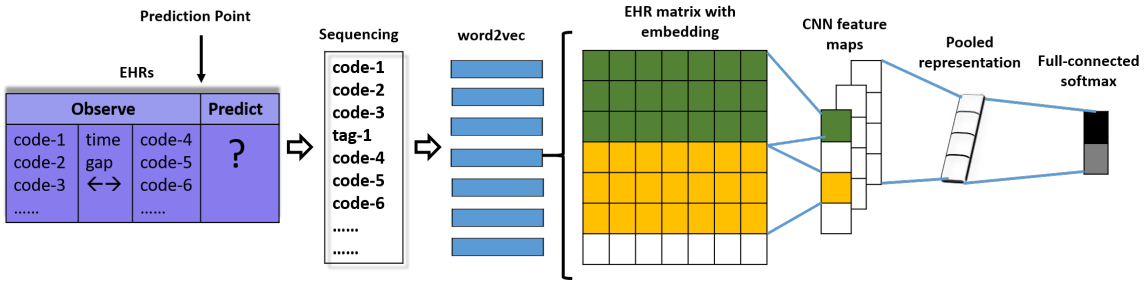

Figure 1: Structure of Convolutional Neural Network Prediction Model

3 Implementation

3.1 Data

We utilized MIMIC-III v1.4 data set, which was released in September 2016. There are 46,520 patients and 651,047 diagnosis events, 240,095 procedures and 4,156,450 prescriptions. Each diagnosis event is associated with a hospital admission time, which is used to assign a sequential order to the ICD9 codes. Each procedure and prescription event are in forms of text names.

3.2 Implementation Details and Evaluation

We applied CNN to the word2vec embeddings to predict the diagnosis of HF. We took MIMIC-III data and trained the word2vec model using all patients prescription, procedure, and diagnosis records, padded with appropriate time tags. We then set the dimension of the embedding as 100. In the risk prediction stage, we set the observation window and prediction window as 2000 days and 90 days, respectively. In order to train a more robust model, we only kept patients that had between 30 and 500 event records in their observation window. We subsequently identified all the case group with patients diagnosed with any type of HF (ICD9 codes with 428.XX). We found that only a relatively small portion of patients could be categorized as case group. To keep the case and control classes balanced, for each patient in case group, we randomly chose two control patients from those without HF diagnosis code. The resulting size of case and control group was 651 and 1302, respectively. In the MIMIC-III data, diagnosis records are not directly timestamped, and thus, we used the admission time of the hospital stay during which diagnosis was given as the timestamp of each diagnosis. We set 16 filters with sizes from 2 to 5 and 4 channel per kernel size in the convolutional layer followed by one pooling layer and two fully connected layers. We used the rectified linear unit (ReLU) activation function in convolution layer and fully connected layers. We used Nesterov Adam (Nadam) as the optimizer. The CNN model was implemented in Keras with Tensorflow as the backend.

We randomly separated the data set by 7:1:2 into the training set, validation set, and test set and the performance of the prediction were evaluated by the test set using accuracy (ACC) and area under the receiver operating characteristic curve (AUC) as metrics.

In order to better understand what CNN has learned in terms of phenotypes of HF, we performed the following: We determined the most salient features that were most informative for the model in the prediction task. Specifically, we approximated the saliency by the output after the convolutional layer. We extracted the output of the pooling layer concatenated with demographic features as the representation of each patient. We used the t-SNE algorithm to visualize different groups in case and control patients. We performed a k-means clustering algorithm on the case group and found common patterns among subgroups. We then identified which features were among the top occurrences of consecutive events (2-grams) in each subgroup.

4 Results

4.1 Risk Prediction

Table 1 shows the summary of results of the prediction with different models. Two columns shows results from models with and without "time tokens" as features. In each group, four types CNN are

tested: a base model with a randomly initialized word embedding layer (**CNN-base**), a model with randomly initialized embedding and demographic information (**CNN-demo**), a model initialized with context (W2V) embeddings trained by Word2Vec model (**CNN-w2v**), and a model with both W2V embedding and demographic information (**CNN-full**). For comparison, we tested logistic regression (**LR**), support vector machine (**SVM**), decision tree (**DT**), and random forest (**RF**) using traditional event count as baseline models.

Overall our CNN models achieved outstanding results that are around 5% better in our selected metrics than the baseline models, showing robustness of CNN in terms of predicting heart failure diagnosis within 90 days. Specifically, the models with "time tokens" feature are markedly better than those without, suggesting the effectiveness of featuring time span and frequency of a patients event records. Models without W2V provided slightly better performance over the W2V model, suggesting that the context information contained in W2V features does not contribute to prediction in the meaningful way that it had been expected to. Models with demographic information showed a slight advantage over other CNN models.

Table 1: HF Prediction results using different models

No Time Tokens	ACC	AUC	Time Tokens	ACC	AUC
LR	72.93%	69.69%	LR	92.29%	91.69%
SVM	69.76%	66.61%	SVM	90.49%	89.96%
DT	72.20%	65.37%	DT	91.77%	91.50%
RF	73.41%	64.44%	RF	82.01%	74.43%
CNN-base	74.14%	77.08%	CNN-base	95.37%	97.06%
CNN-demo	70.70%	76.10%	CNN-demo	96.14%	98.60%
CNN-W2V	72.68%	79.52%	CNN-W2V	93.03%	97.50%
CNN-full	72.68%	77.48%	CNN-full	93.31%	96.97%

4.2 Interpretation of Patient Representation and Phenotyping

Figure 2 shows the control and case groups in form of patient representation extracted from CNN, with dimensions reduced by t-SNE. It clearly shows how effectively learned representations separate the two classes. Table 2 shows a few of the most salient events that contributes to HF prediction. The importance of the events is determined by their degree of output after the convolutional layer, which is approximately the size of their gradient during training.

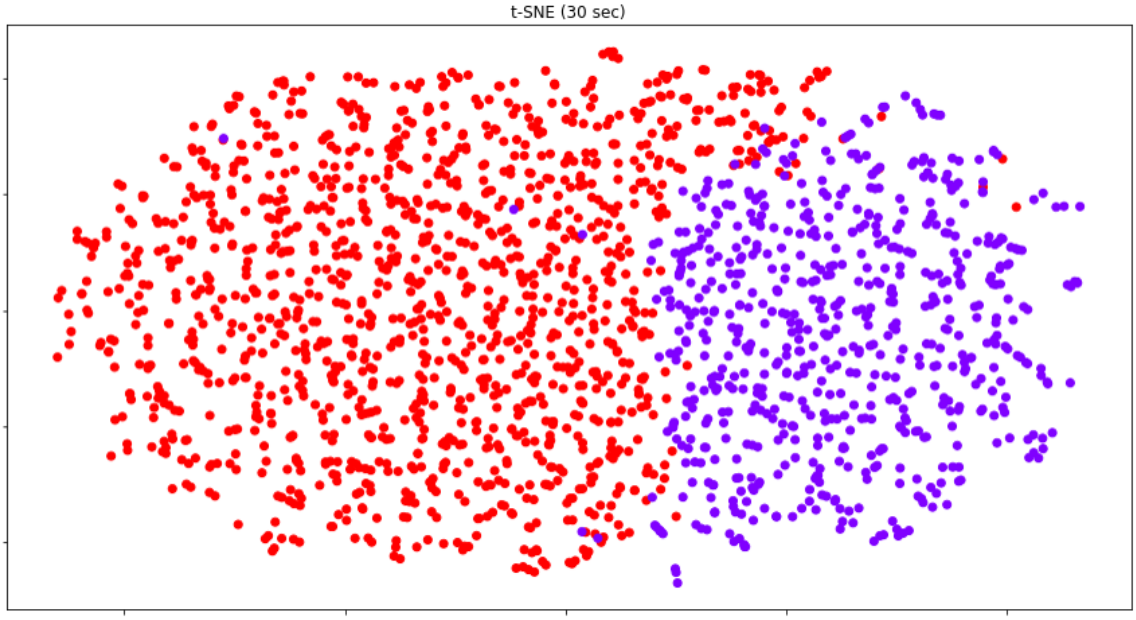

Figure 2: t-SNE visualization of control (red) and case (purple) groups based-on patient representations extracted from CNN

Table 2: Examples of most salient events in different types extracted from CNN

Diagnosis (ICD9)	Prescription/Time Tokens	Procedure (ICD9)
276	Time token 0-2day	389
599	insulin	885
401	potassium chloride	361
250	ns	372
285	d5w	399
427	0.9% sodium chloride	396
585	furosemide	967
V45	sodium chloride 0.9% flush	966
584	metoprolol	960
518	iso-osmotic dextrose	360

It is known that there exists many types of HF that require distinct types of treatment. In order to further investigate subgroups among HF patients, we clustered the representation from case patients by k-means clustering algorithm. Figure 3 shows the subgroups found in the case group with the patient representation and groups with cluster labels identified by k-means algorithm. It is clear that there exists many identifiable subgroups of representation among patients diagnosed with HF.

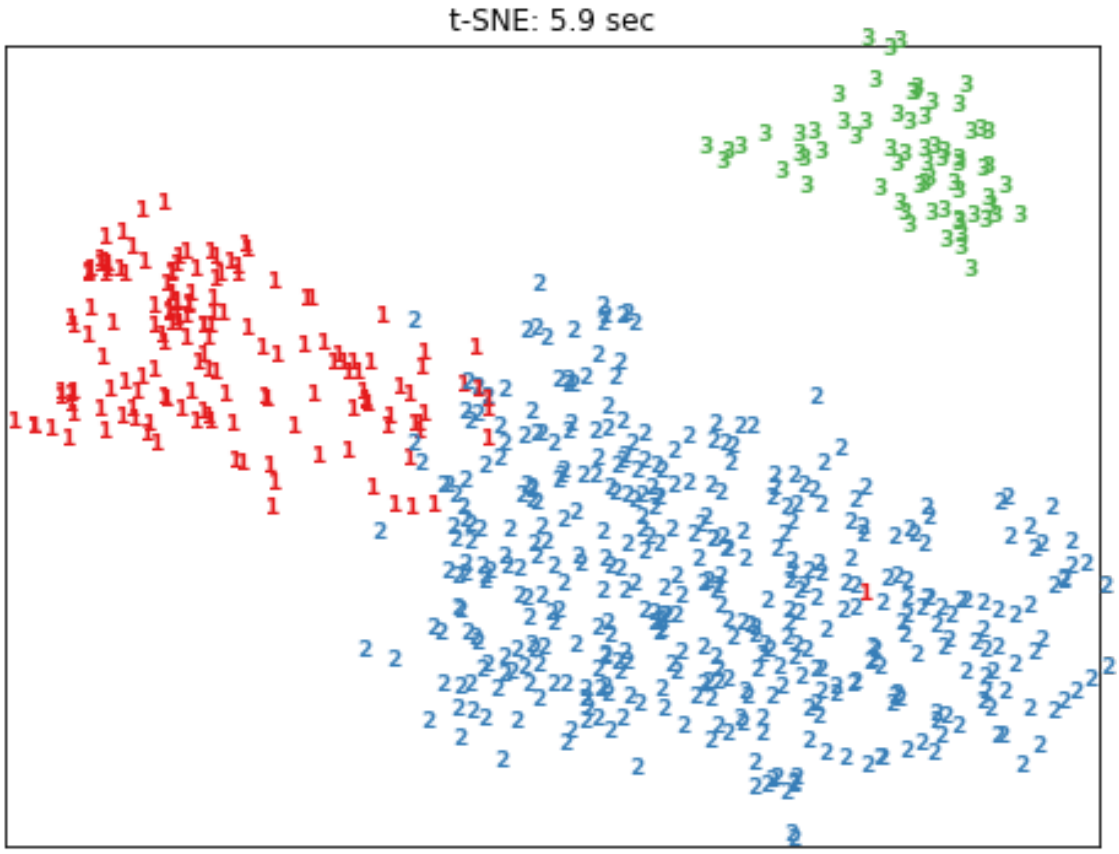

Figure 3: t-SNE visualization of three different case subgroups clustered by K-means

To better get sense of what distinguishes those subgroups, we list common occurrences of event sequences in each subgroup we identified. Tables 3, 3, and 5 show some most common 2-gram diagnosis, procedure and prescription events in each group. Interestingly, we observed that the three clusters have some different phenotyping features.

1. As shown in Figure 3, Group 3 are far from the other two. As expected, Group 3 has very different features than the other two groups. Some attributes include frequent administration of metoprolol and diltiazem, both used for treating hypertension, and calcium gluconate which treats calcium deficiencies. Also, these patients suffer from anemia (285) and kidney disease/failure (403, 584). There was also unique procedures (399,991) which indicate injections into blood vessels. All these evidences show that this group has severe health conditions.

2. Group 1 and Group 2 are in some ways similar. The requirement for administering insulin is very common, suggesting these patients might suffer from diabetes-induced HF. Accordingly, they both frequently display diabetes mellitus (250). But they are distinguishable from each other in terms of the stage of HF development.

3. Group 1 has the V10 and V45 diagnostic codes (which Group 3 also has), and the (396,967) procedure codes, which stand for heart surgery and related cardiac device. This suggests

patients in Group 1 are in worse health conditions than Group 2. Group 2 patients do not have major surgeries, but they tend to have more ultrasound scans (887). This could mean their heart conditions were still relatively stable.

In general, from this clustering method, we probably can conclude that three groups are at different stages of heart failure development: Group 2 ¡ Group 1 ¡ Group 3. Group 2 patients are at early stage, Group 1 already have major heart surgery, and Group 3 have many complications in addition to heart disease. To summarize, this information can be very helpful in facilitating providing the best treatments for each HF patient.

Table 3: Medications, Top 10 2-gram Medications

Group 1
(potassium chloride, potassium chloride)
(insulin, insulin)
(d5w, d5w)
(metoprolol, metoprolol)
(ns, ns)
(furosemide, furosemide)
(acetaminophen, sodium chloride 0.9% flush)
(0.9% sodium chloride, 0.9% sodium chloride)
(magnesium sulfate, potassium chloride)
(metoprolol, potassium chloride)
Group 2
(insulin, insulin)
(d5w, d5w)
(ns, ns)
(potassium chloride, potassium chloride)
(metoprolol, metoprolol)
(magnesium sulfate, potassium chloride)
(acetaminophen, sodium chloride 0.9% flush)
(furosemide, furosemide)
(d5w, acetaminophen)
(0.9% sodium chloride, 0.9% sodium chloride)
Group 3
(metoprolol, metoprolol)
(potassium chloride, potassium chloride)
(d5w, d5w)
(ns, ns)
(0.9% sodium chloride, 0.9% sodium chloride)
(insulin, insulin)
(diltiazem, diltiazem)
(calcium gluconate, calcium gluconate)
(d5w, acetaminophen)
(acetaminophen, sodium chloride 0.9% flush)

Table 4: Diagnostics, Top 10 2-gram ICD9 Codes

Group 1	Group 2	Group 3
(414, 427)	(401, 250)	(276, 427)
(427, 401)	(250, 250)	(403, 584)
(401, 250)	(414, 401)	(401, 250)
(V10, 401)	(401, 530)	(285, 276)
(414, 401)	(530, 250)	(427, V10)
(427, 997)	(414, 427)	(427, 401)
(272, V45)	(250, 403)	(414, 427)
(427, V10)	(276, 276)	(250, 584)
(401, 530)	(276, 518)	(276, 518)
(276, 518)	(427, 997)	(V45, 585)

Table 5: Procedures, Top 10 2-gram ICD9 Codes

Group 1	Group 2	Group 3
(885, 372)	(885, 372)	(990, 990)
(885, 885)	(885, 885)	(389, 389)
(361, 361)	(389, 389)	(885, 372)
(960, 389)	(361, 361)	(004, 004)
(360, 360)	(389, 990)	(389, 991)
(361, 396)	(960, 389)	(389, 399)
(992, 885)	(389, 887)	(885, 885)
(389, 967)	(992, 885)	(991, 991)
(389, 990)	(372, 004)	(360, 360)
(004, 004)	(360, 360)	(399, 399)

5 Discussion

We built a CNN model to predict heart failure. We found that CNN is very effective in this prediction task. Specifically, adding time gap features in patients record significantly improved prediction results. Furthermore, we were able to visualize and interpret phenotypes based on learned hidden representations of the patients. We found this representation helpful in identifying subgroups of HF patients. However, since the MIMIC-III data only contains ICU patients, the scope of our predictive model is limited. Future work can involve applying similar methodologies to larger datasets with more extensive patient encounter histories in order to make predictions applicable to the general population and generate more refined patient representations useful in facilitating treatments of patients with HF.

6 References

1. Alba AC, Agoritsas T, Jankowski M, Courvoisier D, Walter SD, Guyatt GH, Ross HJ. Risk Prediction Models for Mortality in Ambulatory Patients With Heart Failure. A Systematic Review. Circ Heart Failure. 2013. 6(5):8819.

2. Austin PC, Tu JV, Ho JE, Levy D, Lee DS. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. Journal of clinical epidemiology. 2013. 66(4):398-407.

3. Che Z, Cheng Y, Sun Z, Liu Y. Exploiting Convolutional Neural Network for Risk Prediction with Medical Feature Embedding. arXiv preprint arXiv:1701.07474. 2017.

4. Heidenreich PA, Trogdon JG, Khavjou OA, Butler J, Dracup K, Ezekowitz MD. Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association. Circulation. 2011. 123(8):93344.

5. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics, 2012. 13(6):395-405.

6. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L. Large-scale video classification with convolutional neural networks. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2014. 1725-32.

7. Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882 (2014).

8. Lindman BR. The Diabetic Heart Failure with Preserved Ejection Fraction Phenotype: Is It Real and Is It Worth Targeting Therapeutically? Circulation. 2017. 135:736-40.

9. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean, J. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems. 2013. 3111-9.

10. Mozaffarian D, Benjamin EJ, Go AS. on behalf of the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics2016 update: a report from the American Heart Association. Circulation. 2016. 133:e38-e360.

11. Nguyen P, Tran T, Wickramasinghe N, Venkatesh S. Deepr: A Convolutional Net for Medical Records. IEEE journal of biomedical and health informatics. 2017. 21(1):22-30.

12. Smith DH, Johnson ES, Thorp, ML, Yang X, Petrik A, Platt RW, Crispell K. Predicting poor outcomes in heart failure. The Permanente Journal, 2011. 15(4), 4.

13. Taslimitehrani V, Dong G, Pereira NL, Panahiazar M, Pathak J. Developing EHR-driven heart failure risk prediction models using CPXR (Log) with the probabilistic loss function. Journal of biomedical informatics. 2016. 60:260-269.

7 Appendix

Table 6: Diagnosis and Procedure Codes and Descriptions

diagnostic	description
272	hyperglyceridemia
276	hyposmolality,hypernatremia,acid-base balance disorder
285	anemia
250	diabetes mellitus
401	hypertension
403	hypertensive chronic kidney disease
414	coronary atherosclerosis
427	parox atrial tachycardia
518	pulmonary interstitial emphysema
530	achalasia & cardiospasm
584	acute kidney failure
585	chronic kidney disease
997	nervous system complication
v10	neoplasm of unspecified site in gastrointestinal tract
v45	cardiac device in situ
procedure	**description**
4	adjunct vascular system procedures
360	removal of coronary artery obstruction and insertion of stent(s)
361	bypass anastomosis for heart revascularization
372	right/left heart cardiac catheterization
389	puncture of vessel
396	extracorporeal circulation and procedures auxiliary to heart surgery
399	other operations on vessels
885	angiocardiography using contrast material
887	diagnostic ultrasound
960	nonoperative intubation of gastrointestinal and respiratory tracts
967	other continuous invasive mechanical ventilation
990	transfusion of blood and blood components
991	injection or infusion of therapeutic or prophylactic substance
992	injection or infusion of other therapeutic or prophylactic substance