Predicting Diabetes Disease Evolution Using Financial Records and Recurrent Neural Networks

Rafael T. Sousa
Universidade Federal de Goiás
rafaelts777@gmail.com

Lucas A. Pereira
Universidade Federal de Goiás
apereiral@outlook.com

Anderson S. Soares
Universidade Federal de Goiás
engsoares@gmail.com

Abstract

Managing patients with chronic diseases is a major and growing healthcare challenge in several countries. A chronic condition, such as diabetes, is an illness that lasts a long time and does not go away, and often leads to the patient's health gradually getting worse. While recent works involve raw electronic health record (EHR) from hospitals, this work uses only financial records from health plan providers to predict diabetes disease evolution with a self-attentive recurrent neural network. The use of financial data is due to the possibility of being an interface to international standards, as the records standard encodes medical procedures. The main goal was to assess high risk diabetics, so we predict records related to diabetes acute complications such as amputations and debridements, revascularization and hemodialysis. Our work succeeds to anticipate complications between 60 to 240 days with an area under ROC curve ranging from 0.81 to 0.94. In this paper we describe the first half of a work-in-progress developed within a health plan provider with ROC curve ranging from 0.81 to 0.83. This assessment will give healthcare providers the chance to intervene earlier and head off hospitalizations. We are aiming to deliver personalized predictions and personalized recommendations to individual patients, with the goal of improving outcomes and reducing costs

1 Introduction

The last World Health Organization (WHO) diabetes report [1] point that the number of people with diabetes raised from 108 million in 1980 to 422 million in 2014, resulting on a global prevalence of 8.5% among adults over 18 years of age. The estimated number of death directly caused by diabetes in 2015 was 1.6 million[2]. According to the International Diabetes Federation, Brazil ranks fourth among countries with the highest number of diabetics, about 12.4 million in 2017 [3, 4, 5].

Prevention and management of diabetic population is one of the major challenges for health companies. These tasks can not only improve outcomes for patients but also substantially balance healthcare spending. As the disease progresses it can damage the heart, blood vessels, eyes, kidneys and nerves. It also increases the risk of heart diseases and strokes. The blood vessels damages combined with deterioration of the nervous system can result in foot ulcers, infection and the eventual need for limp amputation. In the same way, these damages can also make diabetes the main cause of kidney failure.

According to World Health Organization report, it is possible to prevent diabetes progress, however, efficient tools are required to identify and assess high-risk groups[1]. In this way, machine learning can come up with personalized predictions and recommendations significantly strengthening prevention efforts. We want to describe an approach to assess the complication risk as a diabetes progress index.

Evaluate and assess patient outcomes is a complex problem and some authors have been using machine learning in an attempt to solve it [6, 7, 8, 9]. Their approach models the disease progression based on Eletronic Health Records (EHR) of past patients. We can highlight some remarkable works like Choi's Doctor AI [8] which performs a multilabel prediction with Recurrent Neural Networks (RNN) applied to EHR. The results on a large real-world dataset achieved 79,58% recall@30 and report that the modelling not only mimics the predictive power of human doctor, but also provides clinically meaningful diagnostics. Another work [9] presented a diabetes complication predicting technique using classic machine learning algorithms, which achieved an average 0.75 of area under ROC curve in the prediction of retinopathy, nephropathy and neuropathy.

Both works cited and others [6, 7] relies on a complete electronic report about patients, like demographic data, body mass index, habits and results of exams and laboratory tests. Based on the scenario of most developing countries, in several healthcare companies electronic medical records and exam results are usually either unavailable in digital format or too heterogeneous to be integrated. In this work, we propose the use of financial records as an alternative since it is, among the available data, more reliable and easy to gather. However, financial records are in a different data domain, with more sparsity and irrelevant records. To deal with it, we are also proposing a recurrent model based on recent advances on natural language process.

Inspired by the Brazilian data availability and the difference in data domain, we propose a recurrent neural network architecture to predict diabetes complications through financial records of health plan providers. We assess the risk of diabetic complications as an index to diabetes progress. In the next sections we will expound details of the data used, method and result obtained, limitations, conclusions and future works.

2   Financial Records

In this work we used a dataset of financial records from a Brazilian health plan provider. The last five years of data from their customers were used. There is around $7,000,000$ unique individuals with 327 million records.

The records follow a national standard from Brazilian National Regulatory Agency for Private Health Insurance and Plans (ANS), which is called TUSS (*Terminologia Unificada em Saúde Suplementar* - Supplementary Healthcare Unified Terminology). The TUSS terminology was created in 2010 by Brazilian government to standardize the payment to healthcare companies, based on what services will be performed or have already been, and the exchange of information between health plan providers. The terminology have unique codes for: medical procedures, hospital and clinics rates, materials, medicines and special materials like orthoses and prostheses. Some examples are on Table 1.

Table 1: TUSS interoperability example

Terminology	Hemoglobin A1c
TUSS	40302733
CPT	83037
LOINC	4548-4
SNOMED-CT	43396009

Table 2: TUSS Codes Examples

TUSS	Description
10101012	Appointment
10101039	Emergency Room Service
40901475	Color Doppler of Aorta and Iliac
90019415	Dipyrone 500mg

TUSS encoding can be used as an interface to other procedures or medicine standards like the American Healthcare Common Procedure Coding System or the International Classification of Procedures in Medicine, providing an international interoperability.

As we are using financial records, complications were identified through related records. Some of the main diabetes complications is kidney failure and cardiovascular diseases, so we focused on three kinds of records: (i) Amputations and debridements; (ii) Revascularization and angioplasty; and (iii) Hemodialysis.

3   Proposed Model

The proposed model is mainly inspired by the self-attention sentence embedding proposed by [10]. The idea behind using self-attention is due to its capacity of select the most relevant sentences. We

believe this can be useful to select relevant records and sequence of records among a long input with other possibly irrelevant records, like basic hospital materials, as needles and serum.

The model have an embedding layer connected to a bidirectional Long Short-Term Memory (LSTM) with the self-attention mechanism followed by two fully-connected layers.

The input embedding layer is pre-trained with Word2Vec skipgrams [11] extracted from the whole dataset. As each code is unique as a word we expect to create a vector representation to deal with proprieties that the records have in common to natural language, like: synonyms, there is codes from the same drug but with different doses; antonyms, some drugs have the opposite effect; words combination, some protocols have a standard sequence of exams and drugs and it can be represented as the sum of different codes. This kind of input pre-processing was already done successfully by other works like [8, 12, 13] as a way to extract information from the whole dataset and also deal with the high number of different codes, in our case we have around $150,000$ different codes.

The main idea is to use the network to assess the likelihood that the diabetic will have a complication after a time gap.

3.1 Experiments

To asses prediction capacity we made tests with time gaps of 60, 120, 180 and 240 days as prediction windows. The network input is restricted to the last twelve months of records, with a maximum limit of 500 records due to the vanish gradient problem with LSTMs and the computational cost, and a minimum of 40 records to ensure enough data to make a decision and to exclude individuals with too few records. For each diabetic who had a complication we extracted one input sequence before the first record of complication, and for those who do not had complications we extracted one sequence randomly selected.

As we do not have any diagnosis to confirm who is a diabetic or not and the company do not requires the healthcare providers to inform a diagnosis or an ICD (International Classification of Diseases) code, we define a basic filter to find the greatest possible number of diabetics. As the glycated haemoglobin is define by WHO as the standard diabetes diagnosis test [14], all individuals who had at least two glycated haemoglobin test in less then a year were considered a possible diabetic. The filter used selected $100,000$ individuals which is coherent with the diabetes statistics in Brazil (8.9% of the country's population according to government data), since we have around two millions individuals with more than one year of data we have 5% considered diabetic.

The low prevalence of complications makes the dataset imbalanced. Among the possible diabetics we found around 1900 samples of complications. To overcome such problem we oversampled the positives for each complication class in the training set for each experiment to achieve a better balance and avoid overfitting.

For each time gap we run a 5-fold cross-validation due to the small number of positive samples. The training of all models was implemented on PyTorch and took about 24 hours with a Nvidia Tesla P100.

4   Results

As baseline, we compared with a LSTM network without self-attention, but with the same pre-trained input embeddings. Table 3 report the average area under ROC curve (AUC) over folds with the self-attentive model (LSTM+SA) and with standard LSTM.

Table 3: Mean AUC results of cross-validation procedure.

Time Gap	AUC	
	LSTM+SA	LSTM
60	0.83	0.72
120	0.82	0.72
180	0.81	0.72
240	0.81	0.71

As the diabetic situation becomes more critical there is an increase in specialized exams and visits to the hospital\emergency, so the performance decrease a bit as the prediction windows increase. Long term prediction seems to be a harder problem due to the lack of data.

Comparing self-attentive model with standard LSTM we can see a better performance. This is due to the ability to deal with longer inputs with the self-attention (SA) mechanism. Figure 1 shows a part of neural network attention mapping scores for a particular individual. SA allowing us to see which parts of the input are attended to predict diabetes disease evolution.
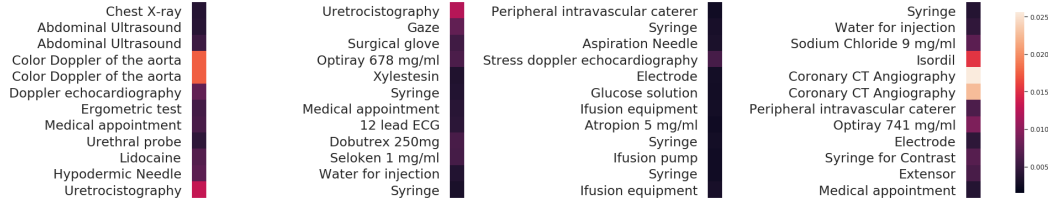
Figure 1: Mapping of the attention scores for an individual. Record descriptions was simplified and the latest records were organized from left to right. The individual had a angioplasty three months later.

The average ROC and Precision-Recall curves over folds at Figures 2 and 3 report the model indicates good balance between positive and negative cases and also point a limitation of how hard is to set a threshold to ensure a low false positive rate. We believe this can be related to the lack of exams results in the financial records. We inferred it is possible to detect patterns as a high frequency of relevant events, but not possible to understand the outcome of it.
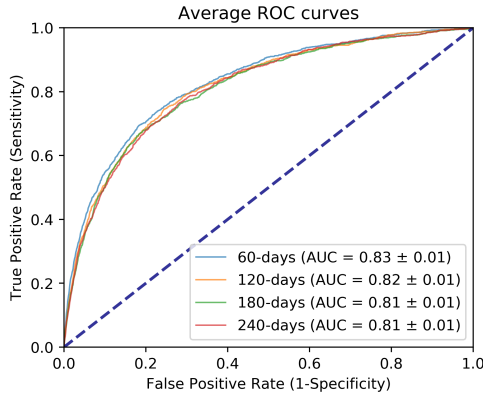

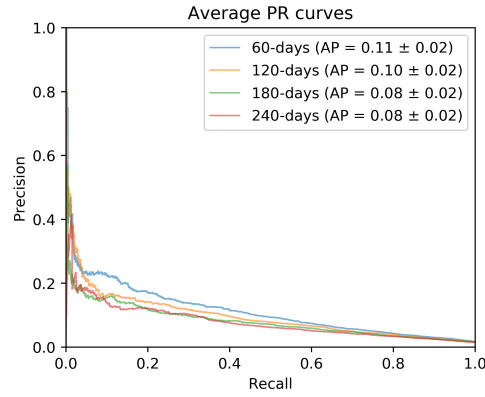Figure 2: Average ROC curves over time gaps

Figure 3: Average PR curves over time gaps

Observing the data in some cases of false negative and false positive, some of those have a lack of data to do the prediction. We assumed that in some of these cases the diabetes silent progress or people refused to be treated. These indicate a challenge to increase the network specificity. One way to minimize this problem is to include these individuals in monitoring programs to give them proper treatment and then assess those with real high risk.

5   Conclusions

We propose the use of self-attentive recurrent neural network to predict diabetes complications using financial records. Our results demonstrated a promising methodology to predict complications and asses the high risk diabetics efficiently with an average AUC of 0.82. Despite of false positives it can still be used as an assessment tool to integrate individuals into a medium-high risk monitoring program.

At the time we are evaluating these results and studying with health plan providers the effectiveness to prevent or minimize diabetes complications.

Using a novel method to embed dates we improved the model performance to an AUC of 0.94. We are still validating this result and expect to publish the proposed method with a full paper report.

References

[1] World Health Organization WHO et al. *Global report on diabetes*. World Health Organization, 2016.

[2] Colin D Mathers and Dejan Loncar. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS medicine*, 3(11):e442, 2006.

[3] Andréa D Bertoldi, Panos Kanavos, Giovanny VA França, André Carraro, Cesar Augusto Ovieda Tejada, Pedro C Hallal, Alessandra Ferrario, and Maria Inês Schmidt. Epidemiology, management, complications and costs associated with type 2 diabetes in brazil: a comprehensive literature review. *Globalization and health*, 9(1):62, 2013.

[4] Amine Farias Costa, Luísa Sorio Flor, Mônica Rodrigues Campos, Andreia Ferreira de Oliveira, Maria de Fátima dos Santos Costa, Raulino Sabino da Silva, Luiz Cláudio da Paixão Lobato, and Joyce Mendes de Andrade Schramm. Burden of type 2 diabetes mellitus in brazil. *Cadernos de saude publica*, 33(2), 2017.

[5] IDF et al. International diabetes federation diabetes atlas - 8th edition, 2017.

[6] Chandra Pandey, Zina Ibrahim, Honghan Wu, Ehtesham Iqbal, and Richard Dobson. Improving rnn with attention and embedding for adverse drug reactions. In *Proceedings of the 2017 International Conference on Digital Health*, DH '17, pages 67–71, New York, NY, USA, 2017. ACM.

[7] B Jin, C Che, Z Liu, Shulong Zhang, Xiaomeng Yin, and XP Wei. Predicting the risk of heart failure with ehr sequential data modeling. *IEEE Access*, 2018.

[8] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pages 301–318, 2016.

[9] Arianna Dagliati, Simone Marini, Lucia Sacchi, Giulia Cogni, Marsida Teliti, Valentina Tibollo, Pasquale De Cata, Luca Chiovato, and Riccardo Bellazzi. Machine learning methods to predict diabetes complications. *Journal of diabetes science and technology*, page 1932296817706375, 2017.

[10] Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *CoRR*, abs/1703.03130, 2017.

[11] 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.

[12] Jacek M Bajor and Thomas A Lasko. Predicting medications from diagnostic codes with recurrent neural networks. *ICLR Workshop 2017*, 2016.

[13] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

[14] World Health Organization WHO. *Use of Glycated Haemoglobin (HbA1c) in the Diagnosis of Diabetes Mellitus*. World Health Organization, 2011.
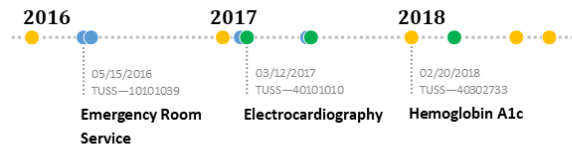
Appendices

1

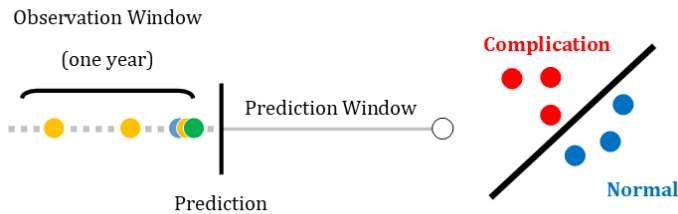Health plan companies store financial data from healthcare providers based on costumers usage.

2

These records describe individually exams, services, procedures and medications consumed.

3

The observation window is one year. Complication prediction is made as a binary classification to asses diabetes evolution.

Target Complications: **Amputations**, **Revascularization** and **Hemodialysis**
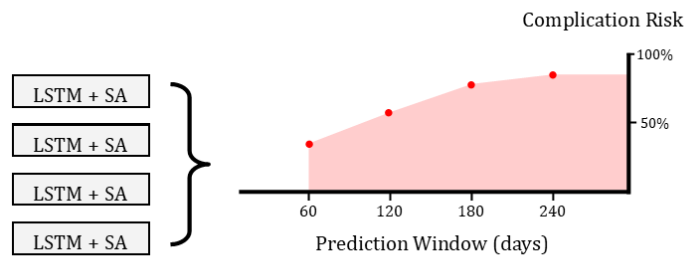
4

A Self-Attentive RNN is trained for each prediction window from 60 to 240 days.

Figure 4: Model's workflow.