Predicting Unplanned Readmissions with Highly Unstructured Data

Constanza Fierro,[*] **Jorge Prez**
Department of Computer Science, Universidad de Chile &
Millennium Institute for Foundational Research on Data
{cfierro,jperez}@dcc.uchile.cl

Javier Mora
Clinica Las Condes, Santiago, Chile &
Surgery Department, School of Medicine, Universidad de Chile
jmorap@clc.cl

Abstract

Deep learning techniques have been successfully applied to predict unplanned readmissions of patients in medical centers. The training data for these models is usually based on historical medical records that contain a significant amount of free-text from admission reports, referrals, exam notes, etc. Most of the models proposed so far are tailored to English text data and assume that electronic medical records follow standards common in developed countries. These two characteristics make them difficult to apply in developing countries that do not necessarily follow international standards for registering patient information, or that store text information in languages other than English.

In this paper we propose a deep learning architecture for predicting unplanned readmissions that consumes data that is significantly less structured compared with previous models in the literature. We use it to present the first results for this task in a large clinical dataset that mainly contains Spanish text data. The dataset is composed of almost 10 years of records in a Chilean medical center. On this dataset, our model achieves results that are comparable to some of the most recent results obtained in US medical centers for the same task (0.76 AUROC).

1 Introduction

Unplanned hospital readmissions are frequent and expensive. The Centers for Medicare & Medicaid Services (CMS) (2017) showed that between 2013 and 2016, 15.5% of the annual hospitalizations correspond to unplanned readmissions within 30 days after discharge. Another study showed that US unplanned re-hospitalizations in 2004 had a cost of $17.6 billion (Jencks et al., 2009). Moreover, unplanned readmissions are used as a measure of quality; To encourage an improvement in the quality of health services, the US and other countries have set penalties to early readmissions (Kristensen et al., 2015). For these reasons, it is expected that hospitals are interested in methods that can accurately predict patients that have a high risk of being readmitted, so they can focus on that group and conduct follow up calls, schedule recurrent visits or give more healthcare education.

Deep learning (Goodfellow et al., 2016) has emerged as one successful machine learning technique for predicting several health care measures. In particular, one of the best results reported so far for predicting unplanned readmissions was produced by Rajkomar et al. (2018) with a deep learning architecture. The most successful deep learning models use free-text as one of their main sources for training data, and most of the models have been trained for the English language. Moreover, besides text data, these models rely on Electronic Health Records (EHR) that follow standards such as the Fast Healthcare Interoperability Resources (FHIR) or the National Center for Biomedical Ontology (NCBO) BioPortal. Thus, they cannot be directly applied to medical centers that do not necessarily

---

[*]Currently at Google (constanzam@google.com).

Table 1: Number of data points for each (sub)entity in CLC data.

Patient	Problem	Encounter	Diagnosis	Order	Form	Clinical note
1,475,224	943,024	7,955,974	6,188,676	19,937,072	21,508,633	18,448,316

follow international standards for registering health-care records, or in which text information is stored in languages other than English. One possible partial solution would be to perform a data migration to a standard EHR format. But for public hospitals in developing countries, such a solution is totally prohibited in terms of the amount of resources it may cost. This motivates the search for a more affordable solution.

In this paper we present a deep learning model that predicts unplanned readmissions consuming data that is considerably less structured compared with the models proposed in the literature. For training our model we use almost 10 years of patient records in a Chilean medical center. In this dataset, a big part of the records are structured in ad-hoc formats designed by the medical center itself, and all the text data is in the Spanish language.

Our preliminary results are encouraging: for the task of predicting unplanned re-admisions after 30 days of discharge, we achieve 0.76 AUROC (Area Under the Receiver Operating Characteristic curve). This result is comparable to the one by Rajkomar et al. (2018) for two US medical centers for the same task in which they obtained 0.76 and 0.77 AUROC. It is also comparable to the AUROC obtained by Jamei et al. (2017) of 0.78. Our results outperform also the results by the Length-Admission-Comorbidity-Emergency (LACE) method (van Walraven et al., 2010) that has an estimated AUROC of 0.72 (Jamei et al., 2017). To the best of our knowledge, our reported results are the first results for our task in a large clinical dataset that mainly contains Spanish text data.

2 RELATED WORK

Deepr (Nguyen et al., 2017) is an end-to-end system based on a convolutional network (Goodfellow et al., 2016) that predicts future risk for patients. It is validated predicting unplanned readmissions after discharge. Deepcare (Pham et al., 2016) is a model that predicts future medical outcomes based on a Long Short-Term Memory recurrent network (LSTM) (Hochreiter & Schmidhuber, 1997). Its performance is tested for disease progression modeling and readmission prediction in diabetes. Both Deepr and Deepcare are based on diagnosis codes and intervention (procedures and medications) codes. In Chile there's no regulation that forces hospitals to record intervention codes, so in our dataset these codes are not present. Thus, we cannot map EHRs to feature vectors using this approach. Deep Patient (Miotto et al., 2016) is an unsupervised deep feature learning method to derive a general-purpose patient representation from EHR. It is strongly based on the use of the NCBO BioPortal (Musen et al., 2011) to extract biomedical concepts from text in English, which we can certainly not use in Spanish.

In 2018, Rajkomar et al. proposed a different way to represent the patient history using the entire raw healthcare records (Rajkomar et al., 2018). They validated their approach in several tasks including unplanned readmissions. They assume that the data can easily be transformed to FHIR resources (Mandel et al., 2016). In our dataset there is a significant portion of the data that does not have a matching FHIR resource (see the **Form** entity in Section 3).

There are some recent works that obtain good AUROC performance by exploiting specific handcrafted features that are then fed to neural network architectures (Jamei et al., 2017; Lin et al., 2019). In our work we explicitly avoid the manual creation of features, as most of our data is highly unstructured and stored as raw text.

3 DATA

We use the historical data of Clinica Las Condes (CLC) of Chile between November 2009 and December 2018. To use this data we first obtained the approval of the respective ethics committee, and it was conditioned on the no use of personal traceable data. The EHR in Clinica Las Condes includes: free text of reason of visit, diagnosis information, clinical notes, medication orders, procedure or-
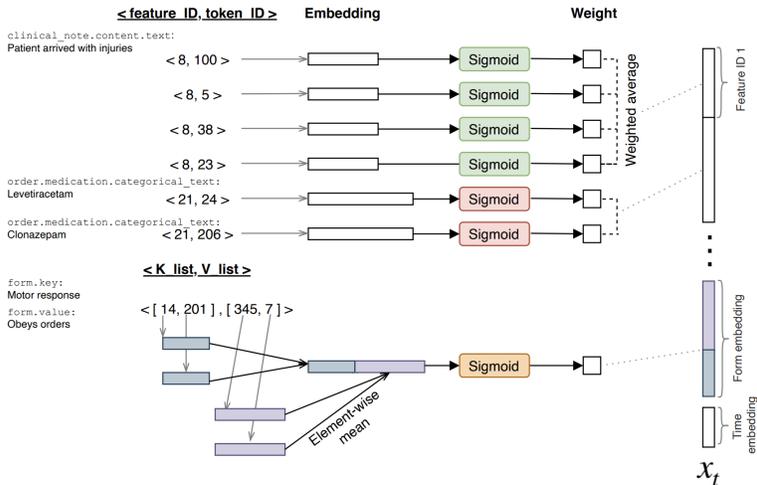
Figure 1: Model architecture to create the vector representation within a time window.

ders, and socio-demographic variables. This data was stored in relational databases and modelled with the following three main *entities*: (1) **Personal** data, (2) **Problem** and (3) **Encounter**. Each encounter (which corresponds to a one time visit) may contain several of the following sub-entities: (3.1) **Diagnosis**, (3.2) **Order**, (3.3) **Form** and (3.4) **Clinical Note**. Except for the personal data, a patient history may have zero or multiple instances of each entity. The numbers corresponding to each (sub)entity are shown in Table 1.

The (sub)entities in our data can be described as follows: **Personal** data contains information about patiens such as birth date, age and nationality. The **Problem** entity describes chronic illnesses or medical conditions that last for a long period of time (e.g.: asthma). The **Encounter** entity contains one time visit information consisting of: beginning and end date, reason of the visit and type of the visit. As we said, each encounter can contain zero or several sub-entities. The **Diagnosis** sub-entity, describes the details of the diseases diagnosed in that encounter. The **Order** corresponds to medical exams orders or prescription drugs. The **Clinical Note** corresponds to free text for any type of information that a doctor wants to add. Finally, the **Form** sub-entity, which is the most numerous in our dataset (21.5M data points) was created at CLC as a very flexible way of storing different types of information. The data in a **Form** entity is a list of key-value pairs, where each key and value are strings that may represent any type of data: free text, categorical classes, numbers, measurements, dates, etc. For example, there are forms to describe the patient state when he/she arrived at the emergency room, or to register the vital signs of an inpatient.

4 METHODOLOGY

To define valid admissions for our task we followed as much as possible the definition of the CMS (for Medicare & Medicaid Services et al., 2014), so we just consider admissions where: the patient didn't die, the patient wasn't discharged against medical advice, and the patient was not transfered to another hospital. We were not able to totally follow the CMS definition for unplanned readmission primarily because there was no treatment or procedures codes on the dataset. In cooperation with nurses and computer technicians of the hospital, we decided to use the "pre-registration" records as a proxy. Everytime an admission is scheduled in advance a pre-registration record is created. So every admission that does not have a pre-regristraion was considered as unplanned. We also considered as unplanned every admission that has a pre-registration during the 24 hours before the hospitalization is created.

With the definition of valid admission and unplanned readmission settled, we created our training dataset considering each example as all the available patient history until a valid admission. We labeled this example as **1** if the patient has an unplanned re-admision after 30 days of discharge.

Table 2: Labels distribution by set.

	Number of examples	Percentage of the total	Percentage of positives examples	Percentage of negative examples
Total	185.323	–	6.18%	93.81%
Train	148.510	80.13%	6.23%	93.76%
Validation	18.318	9.88%	5.81%	94.18%
Test	18.495	9.97%	6.11%	93.88%

Otherwise we labeled that example as **0**. Our cohort then has 6.18% (11,459) positive examples, and 93.82% (173,864) negative examples. It is important to note that each example corresponds to one valid admission, so one patient that has more than one valid hospitalization can produce multiple examples. Inspired by Arango et al. (2019), we split the train (80%), validation (10%) and test (10%) sets by patient and not by example. Since this was made randomly, the overall rates of positive/negatives examples were preserved in each set, as we can see in Table 2.

4.1 EHR REPRESENTATION

As we said in the introduction, we cannot completely rely on EHR representations used by the models proposed so far in the literature. So we decided to follow an idea similar to Rajkomar et al. (2018) in order to use all types of data (text, codes, names, etc.). We consider each entity data and each of their subentities and labels as **features**. For example in our dataset we have the feature "`order.medication.categorical_text`". Each of these features can be associated with several different **tokens**. For example, for the above mentioned feature we can have "`levetiracetam`", "`clonazepam`" and "`enalapril`" as possible tokens. For the case of free-text data, we follow a very simple approach. As an example, consider the feature "`clinical_note.content.text`" and assume that it is associated with the text "`patient arrived with injuries`". In this case we just consider it as four pieces of data for each of the words in the text. That is, the feature is associated with the tokens "`patient`", "`arrived`" "`with`", and "`injuries`". After this process we can represent each piece of atomic data as a pair (`feature_ID, token_ID`).

Given that Forms are really ad-hoc resources, we use a different representation for them. As we have said, Forms are just a set of key-value pairs. Since both keys and values can be defined with arbitrary text, we use sequences of token ids to represent both. Thus, a key-value pair is represented as (`K_list, V_list`). At the end, we have two types of data representation, one as pairs (`feature_ID, token_ID`) to represent non-Form data, and another as pairs (`K_list, V_list`) to represent Form data. Finally, since every data can be associated with a timestamp, we can represent the patient history as a list of time-ordered set of pairs.

4.2 MODEL

We use embeddings to represent each feature and token. Following Rajkomar et al. (2018) we consider 12 hours windows. For each feature we calculate a weighted average between all the token embeddings for that feature that occur in the time window. In our case we computed the weights for each token by using a neural network (NN) with one hidden layer and a single output neuron with a sigmoid activation. We have one NN for each feature, and each NN receives as input the embedding of a token (upper-left in Figure 1).

We consider Form embeddings but computed in a different way. Recall that each key-value pair in a Form is represented as a pair of lists of tokens (`K_list, V_list`). For each (`K_list, V_list`) pair occurring in the time window, we first average all the token embeddings of `K_list` to obtain an embedding k, and separately we averaged the token embeddings of `V_list` to obtain the embedding v. We then use the concatenation $[k; v]$ as input for a NN (one hidden layer, sigmoid output) to compute a weight. Finally, we use these weights to compute a weighted average of all the $[k; v]$ vectors representing every key-value pair in the Forms of the time window (right side in Figure 1).

Table 3: AUROC results

Model	AUROC
Baseline	0.5807
LSTM	**0.7626**
+oversamp.	0.7499

Table 4: Precision and recall by class

Model	Prec. 0	Rec. 0	Prec. 1	Rec. 1
Baseline	**0.95**	0.96	0.26	0.23
LSTM	**0.95**	**0.99**	**0.52**	0.13
+oversamp.	**0.95**	0.97	0.36	**0.29**

To obtain the representation of the time window, we just concatenate the embeddings representing each feature and the embedding representing the Forms, plus an embedding that represents a relative time w.r.t. the time of discharge (x_t in Figure 1). Finally, the complete history is a sequence of those 12 hours window vectors.

To make a prediction, we use the above described representation of the history of a patient to feed a LSTM. We use the last hidden state of the LSTM to perform a logistic regression to predict the probability for each class (1 and 0). We optimized everything jointly (LSTM, embeddings and weighted averages) to minimize the log-loss of the logistic regression using the Adam optimizer (Kingma & Ba, 2014). We used embedding dropout and LSTM hidden dropout.

Baseline model   As in others studies that predict unplanned readmissions (Nguyen et al., 2017; Huang et al., 2019) we compared our model to a Bag of Words (BOW) with Term-Frequency Inverse-Document Frequency (TF-IDF) and Logistic Regression (LR) output. We used the EHR representation (Section 4.1) but we selected the 5,000 more relevant tokens according to its TF-IDF value. Each token TF-IDF value was calculated as the average of its TF-IDF value in each document, and we use each patient timeline as a document.

5   RESULTS

We trained the LSTM model for 6 epochs. The baseline was trained for 8 epochs. On each training we test the model after one epoch on the validation set and we chose the one that performed the best according to the AUROC value. Tables 3 and 4 show the performance of the chosen models on the test set. We also tried to do oversampling of the positive values till we reach 20% on the training set but it obtained worse results compared with no oversampling. Although our AUROC (0.76) is comparable with that of similar work (Rajkomar et al., 2018), our recall is significantly low. We note that Rajkomar et al. (2018) did not report precision nor recall, so we cannot compare with that work on those metrics. On the other hand, the AUROC achieved by our model is above the LACE approximated AUROC performance of 0.72 (Jamei et al., 2017). The precision and recall of LACE are 0.21 and 0.49, respectively (Jamei et al., 2017). Improving the recall of our method is part of our future work.

6   CONCLUSION

Our proposed model achieves good results on the Chilean clinical dataset, and it presents the first experience of its kind on Spanish data. Our trained model can be directly used to deploy a system to be used by CLC. We are currently working with the CLC professionals on reaching that goal.

As future work we would like to visualize the weights used for averaging in every time window to try to understand where the model is focusing to make the predictions. We would also like to improve the recall of our model possibly by modifying the loss function to weight more on the minority class. This should be done taking into account the hospital goals and resources, because it would probably imply to have more false positives.

REFERENCES

Aymé Arango, Jorge Pérez, and Barbara Poblete. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 45–54. ACM, 2019.

Centers for Medicare & Medicaid Services (CMS). Trend in hospital-wide 30-day readmission rates (2017 chartbook). https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/OutcomeMeasures, 2017. [Online; accessed 26-January-2020].

Centers for Medicare & Medicaid Services et al. Yale new haven health services corporation. *Center for Outcomes Research & Evaluation*, 2014.

Ian J. Goodfellow, Yoshua Bengio, and Aaron C. Courville. *Deep Learning*. Adaptive computation and machine learning. MIT Press, 2016. ISBN 978-0-262-03561-3. URL http://www.deeplearningbook.org/.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

Mehdi Jamei, Aleksandr Nisnevich, Everett Wetchler, Sylvia Sudat, and Eric Liu. Predicting all-cause risk of 30-day hospital readmission using artificial neural networks. *PloS one*, 12(7), 2017.

Stephen F Jencks, Mark V Williams, and Eric A Coleman. Rehospitalizations among patients in the medicare fee-for-service program. *New England Journal of Medicine*, 360(14):1418–1428, 2009.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Søren Rud Kristensen, Mickael Bech, and Wilm Quentin. A roadmap for comparing readmission policies with application to denmark, england, germany and the united states. *Health policy*, 119 (3):264–273, 2015.

Yu-Wei Lin, Yuqian Zhou, Faraz Faghri, Michael J Shaw, and Roy H Campbell. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. *PloS one*, 14(7), 2019.

Joshua C Mandel, David A Kreda, Kenneth D Mandl, Isaac S Kohane, and Rachel B Ramoni. Smart on fhir: a standards-based, interoperable apps platform for electronic health records. *Journal of the American Medical Informatics Association*, 23(5):899–908, 2016.

Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6: 26094, 2016.

Mark A Musen, Natalya F Noy, Nigam H Shah, Patricia L Whetzel, Christopher G Chute, Margaret-Anne Story, Barry Smith, and NCBO team. The national center for biomedical ontology. *Journal of the American Medical Informatics Association*, 19(2):190–195, 2011.

Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. Deepr: A convolutional net for medical records. *IEEE journal of biomedical and health informatics*, 21(1):22–30, 2017.

Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. Deepcare: A deep dynamic memory model for predictive medicine. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 30–41. Springer, 2016.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):18, 2018.

Carl van Walraven, Irfan A Dhalla, Chaim Bell, Edward Etchells, Ian G Stiell, Kelly Zarnke, Peter C Austin, and Alan J Forster. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Cmaj*, 182(6):551–557, 2010.