Advances in Informatics, Management and Technology in Healthcare J. Mantas et al. (Eds.) © 2022 The authors and IOS Press. This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0). doi:10.3233/SHTI220732

# Hospital Readmission Prediction via Keyword Extraction and Sentiment Analysis on Clinical Notes

Aitor ZUBILLAGA<sup>a,1</sup>, Paula LACCOURREYE<sup>b</sup>, Jon KEREXETA<sup>a,b</sup>, Nekane LARBURU<sup>a,b</sup>, Eduardo ALONSO<sup>a</sup>, Dr. D. Jesús GÓMEZ<sup>c</sup>, Francisco MARTÍNEZ<sup>c</sup> and Maykel ALONSO-ARCE<sup>d</sup>

<sup>a</sup>Vicomtech Foundation, Basque Research and Technology Alliance, (BRTA), 20009 Donostia, Spain

<sup>b</sup>Biodonostia Health Research Institute, 20014 San Sebastián, Spain <sup>c</sup> Clinica Asunción, Tolosa, Spain <sup>d</sup> STT Systems, Donostia, Spain

**Abstract.** Unplanned hospital readmission is a problem that affects hospitals worldwide and is due to different factors. The identification of those factors can help determine which patients are at greater risk of hospital readmission for early intervention. Our end goal is to predict and identify patterns to (i) feed a decision support system for efficient management of patients and resources and (ii) detect patients at high risk of 30-days readmission enabling preventive actions to improve management of hospital discharges. This study aims to analyze whether natural language processing and specifically keyword extractions tools and sentiment analysis can support 30-days readmission prediction. Features extracted from medical history notes and discharge reports were used to train a Logistic Regression model. The resulting model obtains an AUC of 0.63 indicating that the sentiment polarity score of the discharge report and several of the extracted keywords are representative features to consider.

Keywords. Hospital readmissions, machine learning, natural language processing, unstructured data, sentiment analysis

## 1. Introduction

Reducing hospital readmission rates has been one of the top priorities for healthcare providers and institutions in the last few decades [1], as the ratio of unplanned readmissions (re-hospitalization) in less than 30 days after discharge is a quality indicator of the efficiency of hospital care [2].

The academic community has extensively explored prediction models with different algorithms [3] for the identification of patients at risk of hospital readmission. Prior research generally confirms that machine learning (ML) algorithms improve the prediction of readmission for different diseases and outperform the Medicare and Medicaid Services models or the LACE index [4]. Furthermore, little research has been

<sup>&</sup>lt;sup>1</sup> Corresponding Author, Aitor Zubillaga, Vicomtech Foundation, Basque Research and Technology Alliance, (BRTA), 20009 Donostia, Spain; E-mail: azubillaga@vicomtech.org.

conducted to explore the value of using Natural Language Processing (NLP) techniques to develop 30-day hospital readmission predictive models.

To fill this gap in the literature, our approach successfully extracts clinical knowledge through retrospective analysis of unstructured clinical records from hospital discharge reports and medical history information using NLP and ML.

In order to discriminate between relevant and irrelevant features for 30-day readmission prediction, feature importance will be performed to select the NLP features that have the greatest impact on the results in order to narrow down the feature space of a future extended model, which will include structured and unstructured data.

#### 2. Materials and Methods

This study is based on retrospective unstructured data derived from electronic health records (EHR) of the hospital Clínica Asunción (Tolosa, Spain) between 2004 and 2020. The cohort contains 82,355 admissions (24,610 patients), of which 8,736 were 30 days readmissions. All the admissions had a discharge report, and 47,086 had medical history notes. The inclusion criteria for this study was to have both a discharge report and a medical history note.

Two NLP techniques were used to obtain features from unstructured data: (i) Keyword Extraction (KE) from medical history notes (e.g. "diagnosed with COPD ten years ago", "monthly episodes of acute heart failure syndrome") and (ii) sentiment analysis of the discharge reports (e.g. "at the moment of discharge symptoms have been improved and the patient is stable").

KE was performed using the tool presented by Pérez et al. [6],[7] to normalize biomedical terms of EHRs with the Unified Medical Language System (UMLS) Metathesaurus, a large, multi-lingual compendium of biomedical and health-related terminologies. The 'medical history' field for each patient was selected as the input text to be processed with this tool. From the total of 47,086 medical history notes, 11,680 different biomedical terms (e.g. C0184666) were identified. The less frequent terms (frequency<100) were discarded and a total of 672 terms were used for downstream analysis.

Sentiment analysis was performed on the discharge reports using the sentiment\_by function from the sentimentr R library [8]. This function returns a numeric value indicating the positive/negative degree of the discharge report, extracting the subjectivity of the text written by the clinician who discharged the patient.

Being the objective of the study is to assess the relevance of unstructured clinical notes alone to predict readmission risk, only features obtained by NLP were used. Finally, once the KE and sentiment analysis information was obtained, a Logistic Regression (LR) algorithm was used to predict binary outcome (readmission/no readmission) and analyse the impact of the used features.

#### 3. Results

The ML model was trained by splitting the dataset into train/test by 80/20. In addition, 10-fold cross-validation was performed to avoid overfitting in the training set. The training was iteratively performed while discarding/adding features according to their importance and performing grid search to get optimal hyperparameters. In figure 1.a the boxplot distribution of the results for the ten folders can be seen. The AUC median is 0.63 with a deviation standard of 0.002. The green dot represents the result obtained in the same training set (AUC=0.643) and the red dot the AUC obtained in the test set (AUC=0.616, Figure 1.b), with the 20% partition.



Figure 1 a) AUCs of cross-validation splits b) ROC curve of the testing set.

Since the study's main objective is to identify the most important NLP features, we focus on the features that the model has not discarded after the iterations. From the 672 keywords that have been introduced to the model, after feature selection according to the LR model coefficients, the model keeps 52. Table 1 shows the three most influential features, both positively and negatively.

**Table 1** The three features that have the largest impact on the results (positively and negatively). The description of UMLS concepts with its corresponding code in brackets is presented for features extracted with UMLS mapper. Feature importance relative to the prediction ("impact") is estimated according to the coefficient of the LR model. A higher score means that the specific feature will have a larger effect (positively or negatively) on the model in predicting readmissions.

Feature	Positive impact	Feature	Negative impact
hospital admission C0184666)	0.09	without (C0332288)	-0.12
illness (C0221423)	0.07	Sentiment polarity score	-0.09
complicated (C0231242)	0.06	varicose vein (C0042345)	-0.06

On the one hand, it is very noticeable that the sentiment feature has one of the highest negative impacts. That is to say, the more positive the value extracted by the model, the less likely it is that the patient will be readmitted. On the other hand, UMLS concepts with the most significant impact were consulted with clinicians (Asunción Clinic) for relevant domain knowledge interpretation and several conclusions were drawn: the term C0332288 which indicates "without" has the highest negative impact on patient risk. This result was expected as this term tends to indicate that the patient does not have any significant disease registered in their medical history (e.g. "without a history of significant/chronic diseases") so it is less likely to be readmitted. The opposite occurs with the term C0184666, meaning "hospital admission", which denotes that if included in the report (e.g. "several hospital admissions in the last year") the risk of readmission

increases. This result is also plausible as it indicates that a patient with a previous history of hospital admissions is more prone to readmission.

#### 4. Discussion and Conclusions

Overall, the results obtained demonstrate how machine learning can be used to predict hospital readmissions within 30 days after discharge. Results achieved a 0.63 AUC, similarly to other state-of-the-art studies. However, considering that only data from unstructured clinical notes has been used, a significant improvement in this model's predictive performance is expected once the rest of the clinical, demographic, and laboratory data of patients is included. This study presents the first approach to test whether the use of sentiment analysis and KE of UMLS concepts from medical history can be useful for readmission prediction. The research approach presented differs from other risk prediction models available in the following aspects: (i) applies keyword extraction techniques, (ii) uses sentiment analysis on discharge reports.

In summary, it is concluded that processing clinical notes such as the medical history and discharge report with UMLS concepts extraction and sentiment analysis respectively could be a potential way to further improve the discriminatory ability of a 30-days readmission risk prediction model.

Future work contemplates the combination of the identified significant features with clinical data extracted from EHR such as laboratory results, in order to improve the results and develop a decision support system to aid clinicians during the discharge of patients, key to preventing readmissions.

### Acknowledgments

This project has received funding from the Basque Government's HAZITEK 2020 program under the REINGRES project, with reference number ZL-2021/00604.

## References

- [1] Gai Y, Pachamanova D. Impact of the Medicare hospital readmissions reduction program on vulnerable populations. BMC Health Services Research 2019;19:837.
- [2] Benbassat J, Taragin M. Hospital Readmissions as a Measure of Quality of Health Care: Advantages and Limitations. Archives of Internal Medicine 2000;160:1074–81.
- [3] Safdar S, Zafar S, Zafar N, Khan NF. Machine learning based decision support systems (DSS) for heart disease diagnosis: a review. Artificial Intelligence Review. 2018 Dec;50(4):597-623.
- [4] Garcia-Arce A, Rico F, Zayas-Castro JL. Comparison of Machine Learning Algorithms for the Prediction of Preventable Hospital Readmissions. The Journal for Healthcare Quality (JHQ) 2018;40:129–38.
- [5] Agarwal A, Baechle C, Behara R, Zhu X. A natural language processing framework for assessing hospital readmissions for patients with COPD. IEEE journal of biomedical and health informatics. 2017 Mar 17;22(2):588-96.
- [6] Perez N, Cuadros M, Rigau G. Biomedical term normalization of EHRs with UMLS. arXiv preprint arXiv:1802.02870. 2018 Feb 8.
- [7] Perez N, Accuosto P, Bravo À, Cuadros M, Martínez-Garcia E, Saggion H, Rigau G. Cross-lingual semantic annotation of biomedical literature: experiments in Spanish and English. Bioinformatics. 2020 Mar 1;36(6):1872-80.
- [8] W. Rinker T. sentimentr: Calculate Text Polarity Sentiment. Buffalo, New York: 2021.