proceedings

MDPI

Proceedings

Electronic Health Records Exploitation Using Artificial Intelligence Techniques †

Carla Guerra Tort [1,*], Vanessa Aguiar Pulido [2], Victoria Suárez Ulloa [3], Francisco Docampo Boedo [4], José Manuel López Gestal [4] and Javier Pereira Loureiro [1]

1 CITIC-Research Center of Information and Communication Technologies, University of A Coruña, 15071 A Coruña, Spain; javier.pereira@udc.es
2 Department of Computer Science, University of Miami, Coral Gables, FL 33146, USA; vanessa@cs.miami.edu
3 Institute for Biomedical Research of A Coruña (INIBIC)-Fundación Profesor Novoa Santos, 15006 A Coruña, Spain; Victoria.Suarez.Ulloa@sergas.es
4 Instituto Médico Quirúrgico San Rafael, 15009 A Coruña, Spain; fdocampo@imqsanrafael.es (F.D.B.); jlopez@imqsanrafael.es (J.M.L.G.)
* Correspondence: c.gtort@udc.es
† Presented at the 3rd XoveTIC Conference, A Coruña, Spain, 8–9 October 2020.

Published: 9 September 2020

Abstract: The exploitation of electronic health records (EHRs) has multiple utilities, from predictive tasks and clinical decision support to pattern recognition. Artificial Intelligence (AI) allows to extract knowledge from EHR data in a practical way. In this study, we aim to construct a Machine Learning model from EHR data to make predictions about patients. Specifically, we will focus our analysis on patients suffering from respiratory problems. Then, we will try to predict whether those patients will have a relapse in less than 6, 12 or 18 months. The main objective is to identify the characteristics that seem to increase the relapse risk. At the same time, we propose an exploratory analysis in search of hidden patterns among data. These patterns will help us to classify patients according to their specific conditions for some clinical variables.

Keywords: electronic health record (EHR); Artificial Intelligence (AI); relapse; respiratory diseases

1. Introduction

The electronic health record (EHR) represents the digital version of a patient's medical history. In an EHR system, data is stored in a collection of tables where each record corresponds to a patient's healthcare episode. EHRs constitute a rich source of information, including demographic data (age, gender, address, ...), administrative data and a wide range of clinical information (clinical notes, diagnoses, procedure-treatments, lab test, medical imaging...) [1–3]. The knowledge extracted from EHRs can be used in clinical decision support, epidemiological and predictive tasks, population care improvement and pattern recognition [2,4]. For this reason, the exploitation of EHRs has aroused interest of researchers in the last years [5,6]. Nevertheless, EHRs have some characteristics that make this goal hard to achieve. Heterogeneity, noise, incompleteness, redundancy or the inconsistent representation of data are some of the challenges to cope with. In this context, exploratory analysis and preprocessing steps play a fundamental role [7,8].

Artificial Intelligence (AI) has become a key tool for EHR exploitation. Machine Learning and Deep Learning have been successfully used to identify new risk factors, patterns and medical associations [6,9]. In addition, recent studies show the potential of these modern techniques to make predictions better than the traditional existing methods [9–11].

In this project, we propose the use of AI to exploit and extract value from EHR data. More concretely, we focus our study on the analysis of relapse rates in patients suffering from the most

prevalent diagnoses in our data set. We consider as a relapse the return of a disease time after its apparent overcoming. We will construct a Machine Learning model to predict whether a patient will have a recurrence in less than 6, 12 or 18 months (depending on diagnosis). This model will allow us to identify the characteristics that seem to increase the relapse risk in those patients. At the same time, we will carry out exploratory analysis in search of hidden patterns among data. We hope the results help us to classify patients according to their specific conditions.

2. Data Set Description

Anonymous patient data were extracted from the San Rafael Hospital database. Records range from January 2000 to January 2020. Main diagnoses and procedures are encoded in both ICD-9 and ICD-10, so the data is divided in two codification sets. ICD-9 set consists of 156,362 records and 89,211 patients. ICD-10 set consists of 32,069 records and 25,013 patients. More information about the sets is given in Table 1. Demographic and clinical features acts as predictive variables.

Table 1. Numeric description of ICD-9 and ICD-10 sets.

	Records	Patients	Diagnoses	Procedures
ICD-9	156,362	89,211	4147	1581
ICD-10	32,069	25,013	2691	2555

3. Present Work

Currently, the study is in the preprocessing phase. The most frequent diagnoses have been identified by a descriptive study of the data set. Table 2 shows these main diagnoses and the associated recounts. Among all the most prevalent diagnoses, we have selected those related to respiratory problems. We discarded the diagnoses of traumatology and varicose veins because they were not considered relevant to this specific research.

After selecting the ICD-9 and ICD-10 codes of interest, the sets will be unified in order to procure a larger and completed data set. Null and missing values will be removed to ensure data quality. In addition, the Machine Learning models will be defined. Once we obtain a clean data set, the next steps will allow us to recognize the most explanatory predictive variables for the chosen diagnoses. For this task, we will apply a Principal Component Analysis (PCA) [12].

Table 2. Most prevalent diagnoses in the data set.

	Records	Patients	Relapses
Dorsopathies	10,177	8228	1250
Varicose veins	9700	7981	1568
Arthropathies	9437	8137	1116
Respiratory infections	7722	4349	1466
Rheumatism	6601	5943	534

Author Contributions: All the authors have contributed to the conceptualization of the paper and the design of the research; methodology and AI models definition, V.A.P. and V.S.U.; data acquisition, F.D.B. and J.M.L.G.; writing—original draft preparation, C.G.T.; writing—review and editing, J.P.; All authors have read and agreed to the published version of the manuscript.

Funding: Centro de Investigación de Galicia CITIC is funded by Consellería de Educación, Universidades e Formación Profesional from Xunta de Galicia and European Union (European Regional Development Fund—FEDER Galicia 2014-2020 Program) by grant ED431G 2019/01. Partially supported by the Spanish Ministry of Science (Challenges of Society 2019) PID2019-104323RB-C33.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pham, T.; Tran, T.; Phung, D.; Venkatesh, S. Predicting healthcare trajectories from medical records: A deep learning approach. *J. Biomed. Inform.* **2017**, *69*, 218–229, doi:10.1016/j.jbi.2017.04.001.
2. Yadav, P.; Steinbach, M.; Kumar, V.; Simon, G. Mining Electronic Health Records (EHRs): A Survey. *ACM Comput. Surv.* **2018**, *50*, 85:1–85:40, doi:10.1145/3127881.
3. Martínez-Romero, M., Vázquez-Naya, J. M., Pereira, J., Pereira, M., Pazos, A., & Baños, G. The iOSC3 system: Using ontologies and SWRL rules for intelligent supervision and care of patients with acute cardiac disorders. *Comput. Math. Methods. Med.* **2013**, doi:10.1155/2013/650671.
4. Shickel, B.; Tighe, P.J.; Bihorac, A.; Rashidi, P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE J. Biomed. Health Inform.* **2018**, *22*, 1589–1604, doi:10.1109/JBHI.2017.2767063.
5. Marier, A.; Olsho, L.E.W.; Rhodes, W.; Spector W.D. Improving prediction of fall risk among nursing home residents using electronic medical records. *J. Am. Med. Inform. Assoc.* **2016**, *23*, 276–282, doi:10.1093/JAMIA.
6. Panahiazar, M.; Taslimitehrani, V.; Pereira, N.; Pathak, J. Using EHRs and Machine Learning for Heart Failure Survival Analysis. *Stud. Health Technol. Inform.* **2015**, *216*, 40–44, doi:10.3233/978-1-61499-564-7-40.
7. Yue, L.; Dongyuan, T.; Weitong, C.; Xuming, H.; Minghao, Y. Deep learning for heterogeneous medical data analysis. *World Wide Web* **2020**, 1–23, doi:10.1007/s11280-019-00764-z.
8. Miotto, R.; Li, L.; Kidd, B.A.; Dudley, J.T. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci. Rep.* **2016**, *6*, 26094, doi:10.1038/srep26094.
9. Weng, S.F.; Reps, J.; Kai, J.; Garibaldi, J.M.; Qureshi, N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE* **2017**, *12*, e0174944, doi:10.1371/journal.pone.0174944.
10. Kawaler, E.; Cobian, A.; Pessig, P.; Cross, D.; Yale, S.; Craven M. Learning to Predict Post-Hospitalization VTE Risk from EHR Data. In Proceedings of the 12th AMIA Annual Symposium, Chicago, Illinois, USA, 3–7 November 2012; pp. 436–445.
11. Wong, N.C.; Lam, C.; Patterson, L.; Shayegan, B. Use of machine learning to predict early biochemical recurrence after robot-assisted prostatectomy. *BJU Int.* **2019**, *123*, 51–57, doi:10.1111/bju.14477.
12. Maćkiewicz, A.; Ratajczak, W. Principal components analysis (PCA). *Comput. Geosci.* **1993**, *19*, 303–342, doi:10.1016/0098-3004(93)90090-R.