MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220273

Predicting Readmission Following Hospital Treatment for Patients with Alcohol Related Diagnoses in an Australian Regional Health District

Jingxiang Zhang[a], Siyu Qian[a,b], Guoxin Su[c], Chao Deng[d,e] , Ping Yu[a,e*]

[a] *Centre for Digital Transformation, School of Computing and Information Technology, Faculty of Engineering and Information Sciences, University of Wollongong, Australia*
[b] *Drug and Alcohol Service, Illawarra Shoalhaven Local Health District, Australia*
[c] *School of Computing and Information Technology, Faculty of Engineering and Information Sciences, University of Wollongong, Australia*
[d] *School of Medicine and Molecular Horizons, University of Wollongong, Wollongong, Australia*
[e] *Illawarra Health and Medical Research Institute, Wollongong, Australia*

Abstract

This study aims to investigate the prediction of hospital readmission of alcohol use disorder patients within 28 days of discharge and compare the performance of six machine learning methods i.e., random forest (RF), logistics regression, linear support vector machine (SVM), polynomial SVM, radial SVM, and sigmoid SVM.

Keywords:
Alcohol use disorder, Machine learning, Hospital readmission

Introduction

Unplanned hospital readmission within 28 days after discharge is an indicator of medical quality[1]. Since 2006, the Australian government has been monitoring the 28-day readmission rate to gain more insights into the quality of care[2]. Reducing the rate of unplanned hospital readmission is a way to improve the quality of care and reduce costs[3]. Alcohol use is thought to be associated with a generally higher rate of hospital readmission[4]. Identifying the predictors of readmission in patients with alcohol use disorder is an important step to prevent readmission[5]. Knowing which patients are at increased risk of readmission at the beginning of treatment may also help clinicians provide treatments that match the needs of the patients and reduce potentially preventable readmissions[6]. Although there have been many studies on the factors that affect patients' readmission[5-7], to the best of our knowledge, no existing study has specifically focused on hospital readmission for patients with alcohol use disorder. Therefore, this study aims to identify the predictors of hospital readmission in 28 days for patients with alcohol use disorder, and to predict the patient's hospital readmission status (i.e., binary results of yes or no) within these 28 days by conducting machine learning on data extracted from electronic medical records.

Machine learning algorithms

Logistic regression (LR) is a generalised linear model used to solve binary classification problems[8]. Using a given set of features which can be continuous, discrete or a mixture of the two types and a binary target, LR calculates the linear combination of the input and passes it through the sigmoid (or logical) function. Therefore, the output value of LR is between 0 and 1, which can be interpreted as classification probabilities. The competitive advantage of this method is easy to use a, thus is often used[9]. García-Laencina et al. used LR to predict the 5-year survival of breast cancer patients[8].

Random forest (RF) is an effective prediction tool. Based on the theory of ensemble learning, it completes the learning task by constructing and combining multiple learners, thereby improving the generalisation ability of the classifier, and enabling the algorithm to accurately learn simple and complex classification functions[10]. RF can achieve good performance on various data sets, thus is advantageous than the other technologies for processing highly nonlinear biological data and noise resistance[10].

Classical Support Vector Machines (SVMs) are binary classifiers[9]. SVMs are among the best off-the-shelf supervised learning models that are capable of effectively dealing with high-dimensional data sets[9]. Linear SVM is usually used to handle large amount of data vectors, such as text categorization. Polynomial SVM is used to process images, and radial SVM be used when they have no prior information about the data. Sigmoid SVM is mainly used as a proxy for neural networks[11].

Methods

Data collection

Data were obtained from the Illawarra Health Information Platform (IHIP) - a non-identifiable health databank with data sourced from the Illawarra Shoalhaven Local Health District and other sources[12]. It is a large data set managed by the Centre for Health Research Illawarra Shoalhaven Population (CHRISP). We integrated four de-identified datasets: emergency department (ED), hospital admitted patient data (AP), community-based drug and alcohol service data (DA) and mental health data (MH) from December 2011 to January 2019 (See Table 1). Unique patient identifiers were used across the four datasets, allowing us to determine whether a patient had presented to ED, been admitted to a hospital ward, or had some form of contact (e.g., intake or assessment) with the drug and alcohol service or mental health service.

Variable selection

Since we were interested in factors that led to hospital readmission within 28 days, we defined the readmission in 28 days as the output variable.

We chose a set of predictor variables for examination based on the variables that had been studied in previous research[5-7], and a discussion with the health information manager. The selected variables were from the hospital admitted patient dataset.

They include patient characteristics (i.e., age group, gender, race, marital status) and hospitalisation related information (i.e., emergency status, source of referral, length of stay). We summarised the existing data and manually identified new variables (primary diagnosis, ED visit, DA visit, MH visit, first hospitalisation, number of diagnoses, number of specialties, number of historical admissions before current admission).

Table 1. Introduction of the four datasets

	ED	AP	DA	MH
No. of patients	10,447	9,392	7,068	3,911
No. of records	70,172	49,654	22,6534	50,6065
No. of variables	47	127	18	66

Data pre-processing

The discharge diagnoses in the ED and AP data were used to identify patients with alcohol use disorder. And all the admission records of these patients from the AP data set were extracted. The data was further processed following the rules below:

1. Records of patients who were admitted only once were excluded from the analysis.
2. Planned readmission records were removed to minimise bias in prediction results.
3. Admissions that resulted in either discharge to a hospice or patient death were removed.
4. For admission with multi-episode, the information of the first episode was kept representing the entire admission.
5. Records of each patient were sorted according to the admission date and time.
6. Variables with more than 30% missing values or variables with the same value in all records were removed, the missing observations of medical data were usually difficult to estimate.
7. For categorical variables, merged the categories with percentages less than 1% as "other".

Data analysis

The output variable “whether a patient was readmitted within 28 days” is dichotomous. Therefore, binary logistic regression will be used to determine whether an input variable has a relationship with the output variable[7]. All analyses will be performed with IBM SPSS for Windows version 26.

We plan to apply the RF model to the variables existing in the data set as a benchmark and compare it with the model after adding new variables to determine whether the addition of new variables improves the prediction accuracy. We will apply six machine learning algorithms: RF, LR, linear SVM, polynomial SVM, radial SVM, and sigmoid SVM to build predictive models and compare them with the benchmark. The R statistical package will be used to run these machine learning algorithms. We plan to use ten-fold cross-validation to validate each prediction model. Models will be measured with six parameters: accuracy, sensitivity, specificity, positive predictor values, negative predictor values, and receiver-operating characteristic curve. We will assign the accuracy measure the highest importance because the main goal of the study is to find a prediction model that best predicted the variables.

Results

We expect to define the factors affecting hospital readmission within 28 days of patients with alcohol use disorder through this study. Identifying these predictors are of great clinical importance as they are beneficial for reducing readmission rates going forward. In addition, we plan to establish a high-performance predictive model based on these predictors to accurately predict the patients who will be re-admitted in 28 days. This will help to better match patients and treatment methods or establish a social support network to reduce the burden on hospitals.

References

[1] J. Considine, K. Fox, D. Plunkett, M. Mecner, M. O Reilly, and P. Darzins, Factors associated with unplanned readmissions in a major Australian health service, *Australian health review : a publication of the Australian Hospital Association* **43** (2019), 1-9.

[2] C. Fischer, H.F. Lingsma, P.J. Marang-van de Mheen, D.S. Kringos, N.S. Klazinga, and E.W. Steyerberg, Is the readmission rate a valid quality indicator? A review of the evidence, *PLoS One* **9** (2014), e112282-e112282.

[3] K. Sebelius, US Department of Health and Human Services Strategic Plan; Fiscal Years 2010–2015. Washington, DC: Department of Health and Human Services; 2010, in, 2015.

[4] A. Mason, E. Daly, and M. Goldacre, Hospital readmission rates: literature review, *National Centre for Health Outcomes Development. University of Oxford, Report MR* **3** (2000), 2-61.

[5] E.M. Hansen, A. Mejldal, and A.S. Nielsen, Predictors of Readmission Following Outpatient Treatment for Alcohol Use Disorder, *Alcohol and alcoholism (Oxford)* **55** (2020), 291-298.

[6] J.Y.Z. Li, T.Y. Yong, P. Hakendorf, D.I. Ben-Tovim, and C.H. Thompson, Identifying risk factors and patterns for unplanned readmission to a general medical service, *Australian health review* **39** (2015), 56-62.

[7] A.R. Hoy, Which young people in England are most at risk of an alcohol-related revolving-door readmission career?, *BMC Public Health* **17** (2017), 185-185.

[8] P.J. García-Laencina, P.H. Abreu, M.H. Abreu, and N. Afonoso, Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values, *Computers in Biology and Medicine* **59** (2015), 125-133.

[9] C.-T. Su, C.-T. Su, P.-C. Wang, P.-C. Wang, Y.-C. Chen, Y.-C. Chen, L.-F. Chen, and L.-F. Chen, Data Mining Techniques for Assisting the Diagnosis of Pressure Ulcer Development in Surgical Patients, *Journal of medical systems* **36** (2012), 2387-2399.

[10] X. Zhu, X. Du, M. Kerich, F.W. Lohoff, and R. Momenan, Random forest based classification of alcohol dependence patients and healthy controls using resting state MRI, *Neuroscience letters* **676** (2018), 27-33.

[11] S.-K. Lee, J. Ahn, J.H. Shin, and J.Y. Lee, Application of Machine Learning Methods in Nursing Home Research, *International Journal Of Environmental Research And Public Health* **17** (2020), 6234.

[12] Illawarra Shoalhaven Local Health District *https://www.islhd.health.nsw.gov.au/*.

Address for correspondence

Associate Professor Ping Yu

Centre for Digital Transformation

School of Computing and Information Technology

Faculty of Engineering and Information Sciences

University of Wollongong

Wollongong, New South Wales, 2522, Australia

Tel +61 2 4221 5412

Fax +61 2 4221 4045

Email ping@uow.edu.au