Chapter 22
Trying to Predict in Real Time the Risk of Unplanned Hospital Readmissions

Nilmini Wickramasinghe
https://orcid.org/0000-0002-1314-8843
Swinburne University of Technology, Australia & Epworth HealthCare, Australia

ABSTRACT

This study aims to identify predictors for patients likely to be readmitted to a hospital within 28 days of discharge and to develop and validate a prediction model for identifying patients at a high risk of readmission. Numerous attempts have been made to build similar predictive models. However, the majority of existing models suffer from at least one of the following shortcomings: the model is not based on Australian Health Data; the model uses insurance claim data, which would not be available in a real-time clinical setting; the model does not consider socio-demographic determinants of health, which have been demonstrated to be predictive of readmission risk; or the model is limited to a particular medical condition and is thus limited in scope.

INTRODUCTION

Like all OECD countries, Australia is also facing cost pressure regarding delivering high quality care. In the private healthcare sector in Australia unplanned readmissions are 3.1 typically requires the joint analysis of multiple sources of data (Sherman, 1984). However, this can be challenging as data is often incomplete, fragmented and/or consists of misaligned information (Buhl, Röglinger, Moser, & Heidemann, 2013). This limitation in data quality in turn has hindered epidemiologists to extrapolate demographic information to within plausible limits (Murray, Lopez, & Organization, 1996). Additionally, fragmented data spread across multiple sources makes it difficult for policymakers to compare the relative cost-effectiveness of different interventions (Chechulin, Nazerian, Rais, & Malikov, 2014). Thus, measuring, gauging and creating benchmarks for unplanned readmission is difficult and yet trying to solve this problem, as is the goal of this research, will have many far reaching consequences.

DOI: 10.4018/978-1-7998-1371-2.ch022

LITERATURE REVIEW AND BACKGROUND

Recent developments in the fields of data warehousing and data science have enabled researchers to contribute to a growing body of knowledge in predictive analytics (Buhl et al., 2013). In particular, the building, training and application of predictive models to stratify patients into various risk groups based on information from administrative, insurance, clinical, and government registry sources is becoming a key focus (Chechulin et al., 2014). Such studies are aimed at first aligning complex and sensitive information across multiple sources (Blumenthal, Chernof, Fulmer, Lumpkin, & Selberg, 2016). This information is then used to identify patients in need of additional healthcare resources by means of various intervention methods (Blumenthal et al., 2016).

The preponderance of research on predicting unplanned readmissions applies logistic regression models using dichotomous dependent variables (Blumenthal et al., 2016; Chechulin et al., 2014; Cunningham, 2017; Fleishman & Cohen, 2010; Hartmann, Jacobs, Eberhard, von Lengerke, & Amelung, 2016; Leininger, Saloner, & Wherry, 2015; Li, Cairns, Fotheringham, Ravanan, & Group, 2016; Rodriguez, Munevar, Delaney, Yang, & Tumlinson, 2014) and occasionally linear regressions (Fleishman & Cohen, 2010; Rodriguez et al., 2014). Although the variable to be explained is dichotomous, logistic regression can additionally determine the probability of belonging to a certain group, for example, whether a patient is cost intensive (i.e. a likely unplanned readmission or high risk patient) or not (a relatively healthy patient unlikely to have complications) (Snider et al., 2014). Compared to logistic regression, the scale level of the dependent variable in linear regression is metric (Snider et al., 2014). On the one hand, the use a dichotomous dependent variable with a well-defined threshold allows for a better comparability. However, the dichotomous dependent variable has the disadvantage that potential cost savings can not directly be assigned (Cunningham, 2017). In addition to regression models, classification models such as Support Vector Machine (SVM) and Decision Tree (DT) methods can be applied (Behnke, 2014; Bertakis & Azari, 2010; Koukouvinos, 2016). Classification is the assignment of data objects to a suitable class, whereby, for example, the minimization of the classification error or the maximization of the degree of affiliation are used as performance evaluation criteria (Moturu, Johnson, & Liu, 2007). In SVMs, data objects are represented as vectors in a d-dimensional data space. An SVM looks for a boundary where the objects with different class affiliation are separated as distinctively as possible. This limit is represented by so-called support vectors. In case of more than two attributes, the separating boundary corresponds to a hyperplane (Moturu et al., 2007). Drosou and Koukouvinos (Bertakis & Azari, 2010) use SVM to find an optimal hyperplane that separates cost-intensive from “regular” patients. However, comparing different classification and predictive models, Moturu, Johnson, and Liu (Behnke, 2014) show that SVM have the lowest performance. In their study, Bertsimas et al. (Bertsimas et al., 2008) utilize DT to classify high-cost patients. The advantage of decision trees lies in the ability to be easily interpreted, where the importance of an attribute is reflected by its proximity to the root node. However, especially for data sets with many attributes, the danger of overfitting occurs (Moturu et al., 2007). In this case, very large decision trees are created. Although a large decision tree leads to a high classification accuracy on the training data, it does not necessarily lead to a high classification accuracy on the test data (Moturu et al., 2007). Since the mentioned classification models have not shown a sufficient performance in literature and logistic regression has the advantage of generating probabilities as well, this method is chosen for the predictive analysis. In order to evaluate whether overfitting occurs when learning a classifier, cross-validation of the models is applied.