journal homepage: www.ijmijournal.com

Prediction of hospitalization due to heart diseases by supervised learning methods

Wuyang Dai[a], *Theodora S. Brisimi*[a], *William G. Adams*[b], *Theofanie Mela*[c], *Venkatesh Saligrama*[a], *Ioannis Ch. Paschalidis*[a,*]

[a] *Department of Electrical & Computer Engineering, and Division of Systems Engineering, Boston University, 8 Saint Mary's Street, Boston, MA 02215, United States*

[b] *Department of Pediatrics, Boston University School of Medicine and Boston Medical Center, 88 East Concord Street, Boston, MA 02118, United States*

[c] *Electrophysiology Lab/Arrhythmia Service, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114, United States*

ARTICLE INFO

Article history:
Received in revised form
24 September 2014
Accepted 3 October 2014

Keywords:
Prevention
Predictive models
Hospitalization
Heart diseases
Machine learning
Electronic Health Records (EHRs)

ABSTRACT

Background: In 2008, the United States spent $2.2 trillion for healthcare, which was 15.5% of its GDP. 31% of this expenditure is attributed to hospital care. Evidently, even modest reductions in hospital care costs matter. A 2009 study showed that nearly $30.8 billion in hospital care cost during 2006 was potentially preventable, with heart diseases being responsible for about 31% of that amount.

Methods: Our goal is to accurately and efficiently predict heart-related hospitalizations based on the available patient-specific medical history. To the best of our knowledge, the approaches we introduce are novel for this problem. The prediction of hospitalization is formulated as a supervised classification problem. We use de-identified *Electronic Health Record* (*EHR*) data from a large urban hospital in Boston to identify patients with heart diseases. Patients are labeled and randomly partitioned into a training and a test set. We apply five machine learning algorithms, namely Support Vector Machines (SVM), AdaBoost using trees as the weak learner, logistic regression, a naïve Bayes event classifier, and a variation of a Likelihood Ratio Test adapted to the specific problem. Each model is trained on the training set and then tested on the test set.

Results: All five models show consistent results, which could, to some extent, indicate the limit of the achievable prediction accuracy. Our results show that with under 30% false alarm rate, the detection rate could be as high as 82%. These accuracy rates translate to a considerable amount of potential savings, if used in practice.

[*] *Corresponding author at*: Department of Electrical and Computer Engineering, Boston University, 8 Saint Mary's Street, Boston, MA 02215, USA. Tel.: +1 617 353 0434; fax: +1 617 353 6440.

E-mail address: yannisp@bu.edu (I. Ch. Paschalidis).

http://dx.doi.org/10.1016/j.ijmedinf.2014.10.002
1386-5056/© 2014 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

The US health care system is considered costly and highly inefficient, devoting substantial resources to the treatment of acute conditions in a hospital setting rather than focusing on prevention and keeping patients out of the hospital. According to a recent study [1], nearly $30.8 billion in hospital care cost during 2006 was preventable. Leading contributors were heart-related diseases accounting for more than $9 billion, or about 31%. Clearly, even modest percentage reductions in these amounts matter. This motivates our research to predict heart-related hospitalization. Two key enablers to such research are: the availability of patient EHRs and the existence of sophisticated (machine learning) algorithms that can process and learn from the data.

The adoption of EHRs into medical practices started more than two decades ago and EHRs have found diverse uses [20] e.g., in assisting the quality management in hospitals [2], in detecting adverse drug reactions [3], and in general primary care [4]. These early applications use EHRs for record keeping and information sharing and merely scratch the surface of what may be possible. Our belief is that the *true potential of EHRs* lies in their predictive ability of future acute health episodes and in guiding decision making. Foreseeing future hospitalizations for a large population of patients can drive preventive actions, such as scheduling a visit to the doctor, more frequent and exhaustive screening, calls by case nurses to assure medication adherence, or other mild interventions. All of these actions are much less costly than a hospitalization and, if successful, can drastically reduce hospital care costs. To that end, machine learning methods seem to be promising tools and we extensively explore them for our problem.

Machine learning techniques have recently found use in various health-care applications. Vaithianathan et al. [5] uses multivariate logistic regression, a supervised learning method, to predict re-admissions in the 12 months following the date of discharge. Kim et al. [6] also employs two supervised learning algorithms and additionally incorporates into consideration the interpretability of the models. This interpretability of results is also what we emphasize as an important criterion of method evaluation. Based on insurance claims data, Bertsimas et al. [7] combine spectral clustering (unsupervised method) with classification trees (supervised method) to first group similar patients into clusters and then make more accurate predictions about the near-future health-care cost. More closely related to our work are the prediction of re-admissions [8,9] and the prediction of either death or hospitalization due to congestive heart failure [10,11]. However, we differ from this line of work in that we do not limit our study to patients who are already admitted or to patients with a specific heart ailment. This makes our setting novel and broader.

Our algorithms consider the history of a patient's records and predict whether each individual patient will be hospitalized in the following year, thereby, alerting the health care system and potentially triggering preventive actions. An obvious advantage of our algorithmic approach is that it can easily scale to a very large number of monitored patients; such scale is not possible with human monitors. Our results suggest that with about 30% false positives, 82% of heart-related hospitalizations can be accurately predicted. An important contribution is that these accuracy rates surpass what is possible with more empirical but well accepted risk metrics, such as a heart disease risk factor that emerged out of the Framingham study [12]. We show that even a more sophisticated use of the features used in the Framingham risk factor, still leads to results inferior to our approaches. This suggests that the *entirety of a patient's EHR is useful in the prediction and this can only be achieved with a systematic algorithmic approach*.

The remainder of the paper is organized as follows. Section 2 contains a detailed description of the data set, the preprocessing steps, the methods we propose for hospitalization prediction, and the criteria we apply for evaluating the performance of the methods. Section 3 contains our experimental results. A discussion of the results is in Section 4. We end with some concluding remarks in Section 5.

2. Data and methods

2.1. Detailed data description and objective

The data we used are from the Boston Medical Center (BMC) – the largest safety-net hospital in New England. The study is focused on patients with at least one heart-related diagnosis or procedure record in the period 01/01/2005–12/31/2010. For each patient in the above set, we extract the medical history (demographics, visit history, problems, medications, labs, procedures and limited clinical observations) for the period 01/01/2001–12/31/2010, which contains relevant **medical factors** and from which the features of the dataset will be formed. Data were available from the hospital EHR and billing systems (which record admissions or visits and the primary diagnosis/reason). The various categories of medical factors, along with the number of factors and some examples corresponding to each, are shown in Table 1. We note that some of the Diagnoses and Admissions are not directly heart-related, but may be good indicators of a heart problem. Overall, our data set contains 45,579 patients. 60% of that set forms our *training set* – used for training algorithms – and the remaining 40% is designated as the *test set* and used exclusive for evaluating the performance of the algorithms.

Our objective is to leverage past medical factors for each patient to predict whether she/he will be hospitalized or not during a **target** year which could be different for each patient.

In order to organize all the available information in some uniform way for all patients, some preprocessing of the data is needed to summarize the information over a time interval. Details will be discussed in the next subsection. We will refer to the summarized information of the medical factors over a specific time interval as **features**.

Each feature related to Diagnoses, Procedures CPT, Procedures ICD9 and Visits to the Emergency Room is an integer count of such records for a specific patient during the specific time interval. Zero indicates absence of any record. Blood pressure and lab tests features are continuous-valued. Missing values are replaced by the average of values of patients with a record at the same time interval. Features related to tobacco use are indicators of current- or past-smoker in the specific time interval. Admission features contain the total