Computers in Biology and Medicine 53 (2014) 9–18

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Identifying high-cost patients using data mining techniques and a small set of non-trivial attributes

Seyed Abdolmotalleb Izad Shenas [a], Bijan Raahemi [a], Mohammad Hossein Tekieh [b], Craig Kuziemsky [a],*

[a] University of Ottawa, Telfer School of Management, 55 Laurier Avenue East, Ottawa, ON, Canada K1N 6N5
[b] Electronic Health Information Laboratory, CHEO Research Institute, 401 Smyth Road, Ottawa, ON, Canada

ARTICLE INFO

Article history:
Received 25 March 2014
Accepted 1 July 2014

Keywords:
Healthcare expenditures
Predictive models
Data mining
Decision tree
Medical Expenditure Panel Survey

ABSTRACT

In this paper, we use data mining techniques, namely *neural networks* and *decision trees*, to build predictive models to identify *very high-cost* patients in the top 5 percentile among the general population. A large empirical dataset from the *Medical Expenditure Panel Survey* with 98,175 records was used in our study. After pre-processing, partitioning and balancing the data, the refined dataset of 31,704 records was modeled by Decision Trees (including C5.0 and CHAID), and Neural Networks. The performances of the models are analyzed using various measures including *accuracy*, *G-mean*, and *Area under ROC curve*. We concluded that the CHAID classifier returns the best *G*-mean and AUC measures for top performing predictive models ranging from 76% to 85%, and 0.812 to 0.942 units, respectively. We also identify a small set of 5 non-trivial attributes among a primary set of 66 attributes to identify the top 5% of the high cost population. The attributes are the individual's overall health perception, age, history of blood cholesterol check, history of physical/sensory/mental limitations, and history of colonic prevention measures. The small set of attributes are what we call non-trivial and does not include visits to care providers, doctors or hospitals, which are highly correlated with expenditures and does not offer new insight to the data. The results of this study can be used by healthcare data analysts, policy makers, insurer, and healthcare planners to improve the delivery of health services.

1. Background

The continued growth of health care spending and the widespread implementation of quality performance initiatives have created a growing need for tools to identify high-cost populations. Several countries including Canada (11.2%), the United States (17.7%), Netherlands (11.9%), France (11.6%) and Germany (11.3%) spent more of its GDP on healthcare in 2011 than the OECD average of 9.3% [1]. In the U.S., healthcare spending reached 14–17% of the nation's GDP in the 2005–2009 period and amounted to $2.5 Trillion in 2009 [2,3]. Canada has seen similar trends as health spending reached $192 billion in 2010, growing an estimated $9.5 billion or 5.2% since 2009. This represents an increase of $216 per Canadian, bringing total health expenditure per capita to an estimated $5614 [4]. Other developed countries are showing similar expenditures and increasing rates of healthcare spending.

While there is an overarching desire to reduce healthcare spending, doing so becomes more complicated in presence of the skewed nature of healthcare costs. A small portion of the population is responsible for a majority of healthcare costs. In the US, chronic disease including diabetes, heart attacks, cancer, and stroke cause approximately 70% of all deaths and over 75% of all healthcare costs [2,5]. In Canada, chronic diseases are the leading causes of death [6]. The estimated total cost in Canada of illness, disability and death attributable to chronic diseases amounts to over $80 billion, annually [7].

Healthcare has been referred to as data rich but knowledge poor in that we collect large amounts of data but have difficulty using the data to support tasks such as decision making and policy development [8]. The abundance of healthcare data that is currently available often leads to information overload which severely limits our ability to analyze and apply data meaningfully [9]. A variety of data mining approaches including decision trees, neural networks, and Bayesian nets have been used to analyze and model healthcare data. These data mining approaches have

* Corresponding author. Tel.: +1 613 562 5800x4792.
E-mail addresses: sizad096@uottawa.ca (S.A. Izad Shenas), braahemi@uottawa.ca (B. Raahemi), mtekieh@ehealthinformation.ca (M. Hossein Tekieh), kuziemsky@telfer.uottawa.ca (C. Kuziemsky).

http://dx.doi.org/10.1016/j.compbiomed.2014.07.005
0010-4825/© 2014 Elsevier Ltd. All rights reserved.

supported healthcare decision making about interventions [10], to diagnose disease conditions [11–14], to track major epidemics and outbreaks [15–17], and to support management decisions such as resource allocation and capacity decisions [18–20]. We will expand on existing studies conducted on healthcare cost modeling that have applied regression analysis on the MEPS data [27] and data mining techniques [26].

We have also seen the use of data mining algorithms in healthcare for predictive model building. Data mining is a multidisciplinary field of science and technology which includes machine learning, information retrieval, algorithm development, and statistical analysis, among others, and focuses on the individual units of analysis and predicting its final assignment to a specific class; follows a bottom-up approach which is not concerned with hypothesis formation and testing; are not affected by multi-collinearities among numerous predictors; and effectively handle multiple independent variables in a large dataset with exhaustive details [21]. Researchers have used data mining algorithms to examine healthcare costs by focusing on high-cost profiles among patients diagnosed with specific medical conditions including cardiovascular diseases [22], diabetes [23] and asthma [24]. However, few studies have examined healthcare costs among the general population, irrespective of an individual's disease background [25,26].

A significant shortcoming in existing data mining research in healthcare is the use of trivial measures such as diagnostic disease category [27] or visit counts [22,26] for determining contribution of a given factor in predicting higher health cost. By trivial we are not implying negative connotations of the data but rather a shortcoming of the data because while disease categories, visits, and access to services can provide insight on cost prediction, a shortcoming with that approach is that it only allows us to predict costs after the fact. A better approach is to identify non-trivial and proactive factors of health system expenditures to allow early identification of high cost patients. Doing that could help reduce healthcare expenditures by developing policies to better manage care for these patients.

We summarize the above literature on data mining approaches in healthcare by identifying two areas needing further research. First, is to study the performance of different machine learning predictive models in order to identify which model should be used for different tasks. Second, is to identify data management approaches to enable better incorporation of healthcare data into decision making for clinical and administrative decisions. More specifically, we need to support identification of high cost patients by identifying non-trivial and easy to survey data elements that could enable better proactive identification of high cost patients.

This paper addresses the two above research shortcomings. First, we build and compare the performance of two predictive models to estimate high-cost patients in the general population. Second, we introduce a *small set of attributes* from the Medical Expenditure Panel Survey (MEPS) database to predict high and low-cost patients in order to better estimate healthcare costs.

2. Methodology

We used data mining techniques to build a set of predictive models based on the Medical Expenditure Panel Survey (MEPS) dataset. The research methodology follows the data mining process model which consists of 3 consecutive steps (Fig. 1). The first step is *preprocessing* that includes raw data extraction, attribute selection, and preparation of different versions of the final dataset with a select set of pertinent attributes that are used in decision tree and neural networks classifiers. The second step is *modeling* in which we build, train, and run multiple models on the test sets. The third step is evaluation which deals with analyzing the model's performance using relevant measures to compare different models based on common performance measures.

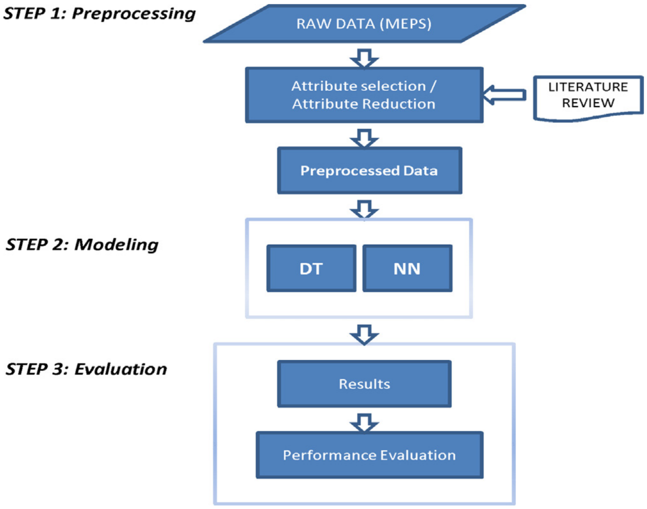

Fig. 1. The research methodology.

The performances of classifiers are analyzed by five performance measures. The first three measures are derived from the confusion matrix: *sensitivity*, *specificity*, and *correctness accuracy*. The sensitivity of a classifier is defined as its ability to correctly identify actual cases true positives (TP). In our study it measures the proportion of high-cost instances which are correctly identified as such. The specificity of a classifier is defined as its ability to correctly identify negative cases true negatives (TN). In our study it measures the proportion of low-cost instances which are correctly identified as such. The correctness accuracy for a data mining classifier is defined as the degree of closeness of its prediction to the actual values, either true or false. In our study, it measures the true results (both true positives or high-costs, and true negatives or low-costs) among all the test population. The confusion matrix measures are summarized as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \quad \text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}}$$

$$\text{Correctness accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FN}+\text{TN}+\text{FP}}$$

The correctness accuracy is reported as an absolute value between 0% and 100% and is used to differentiate models based on their accuracies. However, Kubat et al. [28] claimed that this measure is not adequate when the absolute count of actual negative cases is much larger than actual positives. This is the case in our study, where, for example, we define the high-cost population as the top 5% of the test population, the proportion of high-costs vs. low-costs proportionate is 1:19. This biases the correctness accuracy toward specificity, not sensitivity.

G-mean [28] is a geometric mean of sensitivity and specificity and is only the highest when both of these measures are high:

$$G-\text{mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}}$$

TPr is the percentage of positive examples correctly recognized, and TNr is the percentage of negative examples correctly recognized.

The area under the ROC curve [29] is a similar measure to compare different classification models, which takes into account the trade-off between sensitivity and specificity of a model. In our study, the misclassification costs have been weighed low, but the AUC measure is still reported. The AUC measure for all decision trees (C5.0 and CHAID) and neural networks models are