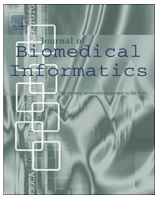
Journal of Biomedical Informatics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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



Coronary artery disease risk assessment from unstructured electronic health records using text mining

Jitendra Jonnagaddala [a,b,c], Siaw-Teng Liaw [a,*], Pradeep Ray [b], Manish Kumar [c], Nai-Wen Chang [d,e], Hong-Jie Dai [e,*]

[a] School of Public Health and Community Medicine, University of New South Wales, Australia
[b] Asia-Pacific Ubiquitous Healthcare Research Centre, University of New South Wales, Australia
[c] Prince of Wales Clinical School, University of New South Wales, Australia
[d] Institution of Information Science, Academia Sinica, Taiwan
[e] Department of Computer Science and Information Engineering, National Taitung University, Taiwan

ARTICLE INFO

Article history:
Received 16 February 2015
Revised 30 July 2015
Accepted 3 August 2015
Available online xxxx

Keywords:
Coronary artery disease
Text mining
Framingham risk score
Temporal data
EHR

ABSTRACT

Coronary artery disease (CAD) often leads to myocardial infraction, which may be fatal. Risk factors can be used to predict CAD, which may subsequently lead to prevention or early intervention. Patient data such as co-morbidities, medication history, social history and family history are required to determine the risk factors for a disease. However, risk factor data are usually embedded in unstructured clinical narratives if the data is not collected specifically for risk assessment purposes. Clinical text mining can be used to extract data related to risk factors from unstructured clinical notes. This study presents methods to extract Framingham risk factors from unstructured electronic health records using clinical text mining and to calculate 10-year coronary artery disease risk scores in a cohort of diabetic patients. We developed a rule-based system to extract risk factors: age, gender, total cholesterol, HDL-C, blood pressure, diabetes history and smoking history. The results showed that the output from the text mining system was reliable, but there was a significant amount of missing data to calculate the Framingham risk score. A systematic approach for understanding missing data was followed by implementation of imputation strategies. An analysis of the 10-year Framingham risk scores for coronary artery disease in this cohort has shown that the majority of the diabetic patients are at moderate risk of CAD.

1. Introduction

Coronary artery disease (CAD), also known as coronary heart disease, is a leading cause of death worldwide [1,2]. CAD is caused by accumulation of plaque in the coronary arteries. Severe blockage of the coronary arteries by plaque can lead to myocardial infarction, which can be fatal. CAD is the most common type of heart disease observed in the general population and the incidence of CAD is rising globally [3]. The costs involved in managing CAD are significantly high, creating an enormous burden on healthcare systems worldwide. Thus, it is important to predict patients at risk of CAD. CAD prediction can assist clinicians to provide early intervention and consequently prevent the development of CAD [4]. CAD risk assessment is part of various national and international clinical guidelines [3,5,6]. The risk assessment is usually done with the help of scoring systems. There are various CAD risk scoring systems available and some of them are specifically modeled for a particular group of patients. The Framingham risk score (FRS) is one of the most popular and well-accepted risk scores to predict CAD. FRS was developed as part of the Framingham heart study. One of the aims of this study is to develop predictive models to estimate probabilities of developing various cardiovascular and/or cerebrovascular diseases [7]. The FRS for CAD provides the probabilities of individuals, aged 30–74 years, to develop CAD. The prediction made with this particular model is valid for 4–12 years. FRS for CAD is calculated using risk factors: age, gender, total cholesterol, or low-density lipoproteins cholesterol (LDL-C), high-density lipoproteins cholesterol (HDL-C), blood pressure (BP), diabetes history and smoking history [8].

With the rapid adoption of electronic health record (EHR) systems, most of the data from patients are stored in electronic format. Necessary data is difficult to obtain in retrospective research studies because the data is scattered across various systems in different formats. Often the risk factor data required for determining FRS are buried in unstructured discharge summary clinical notes.

* Corresponding authors.
E-mail addresses: siaw@unsw.edu.au (S.-T. Liaw), hjdai@nttu.edu.tw (H.-J. Dai).

http://dx.doi.org/10.1016/j.jbi.2015.08.003
1532-0464/© 2015 Elsevier Inc. All rights reserved.

This leads to a problem since most FRS calculators available online require manual input of structured data [9,10]. Entering data manually for a single patient is no trouble but it could be time consuming when this is done for thousands of patients. Extracting the required risk factor data and calculating FRS manually from unstructured EHRs, can be very expensive and resource intensive. Clinical text mining can be used to extract relevant unstructured data and convert it into structured data which can then be used to calculate the FRS.

In this study, we present methods to calculate the FRS from unstructured EHRs using clinical text mining. We retrospectively calculated the 10-year CAD risk scores for a cohort of diabetic patients. Similar studies reporting CAD or cardiovascular risk assessment using EHR data can be found in the literature [11–14]. While these studies provide comprehensive risk models for identifying diabetic patients at risk of CAD, most of the studies used structured data collected specifically for CAD risk assessment. On the other hand, CAD risk assessment using unstructured EHR data has not been well discussed and, to the best of our knowledge, there have not been any studies for calculating FRS for CAD using text mining. The main objective of this study is to demonstrate the feasibility of assessing the risk of CAD from unstructured EHRs using clinical text mining. Specifically, this work presents a system to extract necessary information from unstructured EHRs needed to calculate the FRS. Additionally, the study aims to understand the distribution of calculated FRS in diabetic patients. It is hypothesized that patients who develop CAD will have higher FRS as compared to the ones who do not develop CAD.

2. Materials and methods

2.1. Data

Unstructured EHRs (from here on referred to as corpus) were obtained from the i2b2 2014 shared task 2 which deals with identifying risk factors for heart disease over a period of time [15]. The corpus is de-identified and specifically annotated for heart disease risk factors [16]. The corpus also contains valuable temporal information (up to five encounters) like demographics, medical history, medication and allergies, immunization status, laboratory test results, radiology images, vital signs, personal statistics such as age and weight, and billing details. The corpus includes 1304 unstructured EHRs from 296 diabetic patients. The 296 diabetic patients are stratified into three groups based on when they developed coronary artery disease (CAD). The three groups are: (i) patients who develop CAD over a period of time, (ii) patients who do not develop CAD and (iii) patients who have already been diagnosed with CAD. Although the EHRs were de-identified, the time progression was maintained in the form of adjusted dates. The time between the first and last record for patients was calculated to understand the length of their medical history (Appendix A).

2.2. Workflow

Fig. 1 demonstrates the steps carried out to calculate the 10-year FRS for CAD. All the risk factors required for calculation of 10-year CAD FRS were extracted using a text mining system specifically developed for this study. An error analysis was conducted to understand the output obtained from the text mining system. Cohort selection was performed to determine patients eligible for calculating FRS. Systematic assessment was carried out to understand the quality of data. Various imputation strategies were employed to address missing data. Following the imputations, the 10-year CAD FRS was calculated for eligible patients. Finally, analysis was performed on eligible patients by stratifying them according to their CAD status. The key steps involved in the workflow are explained in the following sections.

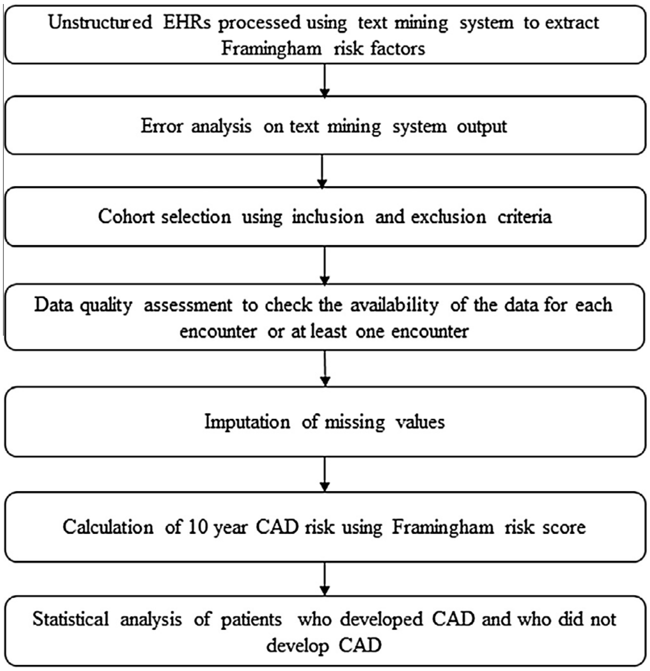

Unstructured EHRs processed using text mining system to extract Framingham risk factors
Error analysis on text mining system output
Cohort selection using inclusion and exclusion criteria
Data quality assessment to check the availability of the data for each encounter or at least one encounter
Imputation of missing values
Calculation of 10 year CAD risk using Framingham risk score
Statistical analysis of patients who developed CAD and who did not develop CAD

Fig. 1. Overall workflow of the study.

2.3. Text mining system

The system developed for determining coronary artery disease risk scores (Fig. 2) is an extension to our work for i2b2 2014 shared task track 2 [17,18]. The developed system consists of three major components, namely FRS risk factor extraction component I, FRS risk factor extraction component II and post-processing component. The first two components include sub-components. FRS risk factor extraction component I and post-processing component were specifically developed for this study. Unstructured EHRs were processed through FRS risk factor extraction component I and II at the same time. After which the output was passed to the post-processing component for further processing.

FRS risk factor extraction component I and its sub-components were developed using Apache Ruta, a scripting language based rules engine. Rules were implemented to recognize mentions, abbreviations, punctuations and specific terms that imply age, gender, total cholesterol and HDL-C. For example, in a record with phrase ‘63 yo ’, the value 63 is extracted and implied as age based on abbreviation ‘yo’ which stands for ‘years old’. Rule-based FRS risk factor extraction component II and its sub-components were also based on Apache Ruta. This component extracts information regarding patient smoking history and BP values. A custom-built dictionary of smoking terms was used to identify smoking history. Similarly, to extract systolic blood pressure (SBP) and diastolic blood pressure (DBP) values pattern-matching rules were used. Post-processing involved filtering records based on rules. For example, rules were developed to remove records, which do not contain age and gender information. This component also assigns diabetes history for the patients in cohort. Since the corpus only includes those patients who are diagnosed with diabetes, the component assigned diabetes history as present for all patients. This component was also responsible to identify patients in the corpus eligible for the 10-year CAD FRS calculation. The output from this component are values for Framingham risk factors (age, gender,

Please cite this article in press as: J. Jonnagaddala et al., Coronary artery disease risk assessment from unstructured electronic health records using text mining, J Biomed Inform (2015), http://dx.doi.org/10.1016/j.jbi.2015.08.003