Journal of Biomedical Informatics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A hybrid model for automatic identification of risk factors for heart disease

Hui Yang *, Jonathan M. Garibaldi

School of Computer Science, University of Nottingham, Nottingham, UK
Advanced Data Analysis Centre, University of Nottingham, Nottingham, UK

ARTICLE INFO

Article history:
Received 2 February 2015
Revised 3 September 2015
Accepted 4 September 2015
Available online xxxx

Keywords:
Risk factors
Heart disease
Machine learning
Rule-based approach
Hybrid model
Natural language processing
Clinical text mining

ABSTRACT

Coronary artery disease (CAD) is the leading cause of death in both the UK and worldwide. The detection of related risk factors and tracking their progress over time is of great importance for early prevention and treatment of CAD. This paper describes an information extraction system that was developed to automatically identify risk factors for heart disease in medical records while the authors participated in the 2014 i2b2/UTHealth NLP Challenge. Our approaches rely on several nature language processing (NLP) techniques such as machine learning, rule-based methods, and dictionary-based keyword spotting to cope with complicated clinical contexts inherent in a wide variety of risk factors. Our system achieved encouraging performance on the challenge test data with an overall micro-averaged *F*-measure of 0.915, which was competitive to the best system (*F*-measure of 0.927) of this challenge task.

1. Introduction

Coronary artery disease (CAD), i.e. Coronary heart disease (CHD), is the leading cause of death in both the UK and worldwide. It is responsible for more than 73,000 deaths in the UK each year. About 1 in 6 men and 1 in 10 women die from CAD. Extensive clinical and statistical studies have identified several factors that increase the risk of CAD. The traditional risk factors for CAD are high LDL cholesterol, low HDL cholesterol, high blood pressure, family history, diabetes, smoking and obesity. The detection of related risk factors and tracking their progress over time is of great importance for early prevention and treatment of CAD.

The rapid adoption of Electronic Health Records (EHRs) in recent years has been shown to be a promising avenue for improving clinical research [13]. Despite structured information in EHRs – diagnosis codes, medications, and laboratory test results – a significant amount of medical information is still stored in narrative text format, principally in clinical notes from primary care patients. Unstructured clinical texts are widely recognized barriers for the application of clinical tools to clinical data. Natural language processing (NLP) technologies provide a solution to convert free text into structured representations that will be further re-used and re-purposed by clinical research [3].

Manual detection of heart disease risk factors from large scale medical records is prohibitively expensive, time-consuming and prone to error. Large-scale accurate risk identification therefore requires automated software that is fine-tuned to the structure of the text, the content of the medical records, and the specific requirements of a particular project. To facilitate the application of the NLP tools to the studies of heart disease, 2014 i2b2/UTHealth NLP Challenge[1] Track 2 [19] was organized to comprehensively investigate the identification of related risk factors for heart disease in diabetic patients. The objective of this challenge task is to find clinical evidence from medical records, which indicates the presence and progression of diseases such as DIABETES (DM) and CORONARY ARTERY DISEASE (CAD), and associated risk factors like HYPERTENSION, HYPERLIPIDEMIA, SMOKING STATUS, OBESITY STATUS, and FAMILY HISTORY OF CAD. In addition, different categories of medications prescribed for individual diseases or risk factors are required to be recognized from the text.

The prediction of heart disease risk factors using clinical and statistical methods has been receiving much attention in the recent decade [6,17,28,33]. But few published studies employed NLP techniques to investigate this research issue on the basis of textual

* Corresponding author at: School of Computer Science, University of Nottingham, Jubilee Campus, Nottingham NB8 1BB, UK. Tel.: +44 (0) 115 95 14212; fax: +44 (0) 115 95 14254.
E-mail address: Hui.Yang@nottingham.ac.uk (H. Yang).

[1] http://www.i2b2.org/NLP/HeartDisease.

http://dx.doi.org/10.1016/j.jbi.2015.09.006
1532-0464/

medical records. There have been some related studies that were targeted for a number of text mining tasks such as obesity identification [24], smoking status identification [25], and medication extraction [26]. The most related work was conducted by Byrda et al. [1] where a hybrid NLP pipeline was proposed for the identification of heart failure diagnostic criteria.

This paper is the extension of our i2b2 workshop paper [31], which details our efforts to the 2014 i2b2 risk factor challenge task. A hybrid model was developed, which integrates a variety of methodological approaches, such as dictionary-based keyword spotting, rules and supervised learning, for the detection of a variety of heart disease risk factors. Our developed system achieved promising performance with an overall micro-averaged *F*-measure of 0.915.

The rest of the paper is organized as follows: Section 2 provides the details about the dataset used for the risk factor detection. In Section 3 we discuss the research issues in risk factor detection. Section 4 details the methods that we employ to deal with the complexity in risk detection. System performance and error analysis is reported in Section 5. Section 6 reflects related work, and our conclusions are given in Section 7.

2. Dataset

2.1. The i2b2 corpus

The dataset used in the challenge includes discharge summaries, clinical notes and letters obtained from Partners HealthCare.[2] For the challenge task, a total of 1,304 medical reports for 296 patients were released to challenge participants. All records have been fully de-identified and manually annotated for heart disease risk factors. 790 annotated medical records (178 patients) are used as a training set, and the remaining 514 records (118 patients) are used as a test set to evaluate the performance of the participating systems.

Fig. 1 gives the excerpt of a medical record with clinical evidence to denote the heart disease risk factors that a patient probably has. The annotated clinical text in Fig. 1 is visualised using the Brat Annotation tool[3] [18]. The text that indicates the presence of a particular risk factor (RF) is extracted as relevant evidence. Eight main risk factor categories are required to be identified from text. The distributions of eight main risk factor categories with 38 associated indicators in both training and test data are shown in Table 1. More details of the description of individual risk factors with associated indicators can be found in the i2b2 challenge annotation guideline [20].

Each risk factor category has its own set of indicators that are used to identify whether or not the disease or risk factor is present for that patient. For example, in Fig. 1 the risk factor HYPERLIPIDEMIA has two indicators: (a) '*hyperlipidemia*' → <HYPERLIPIDEMIA indicator="mention"/> (b) '*LDL 118*' → <HYPERLIPIDEMIA indicator="high LDL"/>.

Moreover, each risk factor (except for SMOKING STATUS and FAMILY HISTORY) is associated with a '*time*' attribute, i.e. when it is present, *before/during/after DCT* (Document Creation Time). For each medical record, the system will output a list of document-level risk factor annotations as shown in Fig. 2, which are used for the final evaluation of system performance. Each annotation consists of three parts, i.e. a risk factor, a time attribute, and an associated risk indicator or medication type. In Fig. 2, each risk factor annotation is supported by one particular clinical evidence instance detected from the excerpt of the clinical record in Fig. 1. It is noted that sometimes one evidence instance (e.g., '*atenolol*', and '*hyperlipidemia*') might refer to multiple annotations with different time attributes.

Each risk indicator should, at minimum, have one clinical instance to support its presence. It is possible that multiple instances related to a specific indicator are found in the same medical record. Furthermore, to track the progression of heart disease, each patient has 3 ∼ 5 longitudinal documents with different DCT, e.g., the file with the earlier time stamp is labelled with '*xxx-01*' whereas the latter one with '*xxx-02*', which allow a general timeline in the patient's medical history.

2.2. Refining clinical evidence provided in the training data

In the training data, the challenge organizers provided two sets of data: one is phrase-level clinical evidence found in medical records, another is document-level risk factor annotations (see Fig. 2) that are generated based on the detected evidence. Each document was annotated by three different annotators in which relevant text fragments were extracted and marked as clinical evidence shown as below:

```
<CAD text="RCA stenting" time="during DCT" indica
tor="event"/>
```

The final document-level risk factor annotations were created by combining three sets of evidence provided by different annotators.

In the challenge task, the identification of phrase-level evidence was not required, and thus annotating phrase-level evidence was not the part of the annotation task. In fact the provided evidence set was still in its raw form, i.e. the 'working notes' of the annotators in support of their decision on the document-level tags. The challenge organization provided these working notes as supporting material rather than as the ground truth. As a result, to utilize this evidence for system development, we needed to refine it as follows:

- *Inconsistent span boundaries in clinical evidence.* The boundary of supporting evidence in the same text marked by different annotators is not consistent. For example, in the sentence (E1) below, three annotators gave different text spans to depict clinical evidence, '*cut back his cigarettes*', '*cigarettes to one time per week*', '*cut back his cigarettes to one time per week*' for SMOKER STATUS.
  E1. *He has cut back his cigarettes to one time per week.*
- *Conflicted evidence.* We observed that sometimes the annotators disagree with each other in terms of risk factor, associated indicator, or time attribute due to incompatible interpretations of some unclear or ambiguous contexts. For example, in the text, 'repeat episode relived by nitro again', one annotator considered the mention 'nitro' as [CAD:mention] whereas another treated it as [MEDICATION:nitrate].
- *Missing clinical evidence.* The annotators are just required to provide, at minimum, one instance for each identified risk factor indicator. There exist some scenarios in which a document contains multiple mentions that refer to the same risk factor indicator, but the annotators only mark up one or two of them as relevant evidence.

To facilitate the detection of clinical evidence and the classification of risk factor indicators, we applied several strategies to further refine the provided annotations:

(a) For the identical evidence instances from different annotators, replace them with a single evidence annotation.

[2] http://www.partners.org.
[3] http://brat.nlplab.org/.

Please cite this article in press as: H. Yang, J.M. Garibaldi, A hybrid model for automatic identification of risk factors for heart disease, J Biomed Inform (2015), http://dx.doi.org/10.1016/j.jbi.2015.09.006