Journal of Biomedical Informatics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

An automatic system to identify heart disease risk factors in clinical texts over time

Qingcai Chen [a], Haodi Li [a], Buzhou Tang [a,*], Xiaolong Wang [a], Xin Liu [a], Zengjian Liu [a], Shu Liu [a], Weida Wang [a], Qiwen Deng [b], Suisong Zhu [b], Yangxin Chen [c], Jingfeng Wang [c]

[a] Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen 518055, China
[b] The Sixth People's Hospital of Shenzhen, Shenzhen 518052, China
[c] Department of Cardiology, Sun Yat-sen Memorial Hospital of Sun Yat-sen University, Guangzhou 510120, China

ARTICLE INFO

Article history:
Received 30 January 2015
Revised 22 August 2015
Accepted 1 September 2015
Available online xxxx

Keywords:
Risk factor identification
Clinical information extraction
Heart disease
Machine learning

ABSTRACT

Despite recent progress in prediction and prevention, heart disease remains a leading cause of death. One preliminary step in heart disease prediction and prevention is risk factor identification. Many studies have been proposed to identify risk factors associated with heart disease; however, none have attempted to identify all risk factors. In 2014, the National Center of Informatics for Integrating Biology and Beside (i2b2) issued a clinical natural language processing (NLP) challenge that involved a track (track 2) for identifying heart disease risk factors in clinical texts over time. This track aimed to identify medically relevant information related to heart disease risk and track the progression over sets of longitudinal patient medical records. Identification of tags and attributes associated with disease presence and progression, risk factors, and medications in patient medical history were required. Our participation led to development of a hybrid pipeline system based on both machine learning-based and rule-based approaches. Evaluation using the challenge corpus revealed that our system achieved an *F*1-score of 92.68%, making it the top-ranked system (without additional annotations) of the 2014 i2b2 clinical NLP challenge.

1. Introduction

Heart disease attracts much attention, given its history as the number one cause of death in both women and men throughout the world [1]. Several factors have been identified as risks related to heart disease, including hyperlipidemia, hypertension, obesity, and smoking status. In order to predict and prevent heart disease, it is necessary to first identify risk factors embedded in unstructured clinical documents. Over the last decade, many studies have been undertaken to identify these risk factors, resulting in the creation of publicly available tools, such as clinical Text Analysis and Knowledge Extraction System [2], an open-source tool capable of identifying smoking status. However, no study has investigated the identification of all risk factors associated with heart disease, possibly due to the diversity of their clinical descriptions.

Heart disease is often related to other diseases, such as diabetes, that share several observable characteristics, including obesity and smoking status, as well as some medications, such as metoprolol. All of these were regarded as heart disease risk factors for this study.

The main challenge in identifying all heart disease risk factors is that they are presented in a variety of forms in clinical texts. To comprehensively investigate the identification of all heart disease risk factors, the National Center of Informatics for Integrating Biology and Beside (i2b2) issued a risk factor identification track (track 2) in the clinical natural language processing (NLP) challenge in 2014 [3]. The goal was to identify information medically related to heart disease risk and track its progression over sets of longitudinal patient medical records. We participated in this track and developed a hybrid pipeline system based on both machine learning and rule-based approaches.

In our system, all heart disease risk factors were divided into three categories according to their descriptions, with each category identified individually. Evaluation using the challenge corpus revealed that our system achieved an *F*1-score of 92.86%, making it the top-ranked system (without additional annotations) for the 2014 i2b2 clinical NLP challenge.

* Corresponding author at: Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen 518055, China.

E-mail addresses: qingcai.chen@gmail.com (Q. Chen), Haodili.hit@gmail.com (H. Li), tangbuzhou@gmail.com (B. Tang), wangxl@insun.hit.edu.cn (X. Wang), hit.liuxin@gmail.com (X. Liu), liuzengjian.hit@gmail.com (Z. Liu), liushuhit@outlook.com (S. Liu), weida.wong@gmail.com (W. Wang), qiwendeng@hotmail.com (Q. Deng), 13809883596@163.com (S. Zhu), tjcyx1995@163.com (Y. Chen), dr_wjf@hotmail.com (J. Wang).

http://dx.doi.org/10.1016/j.jbi.2015.09.002
1532-0464/

2. Related work

The heart disease risk factor identification track of the 2014 i2b2 clinical NLP challenge consisted of two subtasks: risk factor extraction and time attribute identification. To the best of our knowledge, no study has ever been specifically designed for heart disease risk factor identification, although many related studies have been proposed. The most closely related study by Roy et al. developed a hybrid NLP pipeline system to extract Framingham heart failure criteria with time attributes from electronic health records [4].

Heart disease risk factor extraction is a typical information extraction task related to clinical concept recognition [5–9], phenotyping [10], smoking status identification [11–15], obesity identification [16,17], etc. clinical concept recognition is a named entity recognition (NER) task that extracts all problems, treatments, and tests, where problems include diseases and observable characteristics and treatments include medications. The most representative work concerning clinical concept recognition is the 2010 i2b2 clinical NLP challenge, where various machine learning-based, rule-based, and hybrid methods were proposed [18–20]. Phenotypes that include diseases and some observable characteristics have also been widely investigated. Chaitanya et al. summarized approaches for phenotyping [10]. The i2b2 clinical NLP challenges in 2006 and 2008 involved a track on smoking status identification and a track on obesity identification, respectively. The best system for smoking status identification used a method based on support vector machines (SVMs) [21], whereas the best system for obesity identification combined dictionary lookup, rule-based, and machine learning-based methods [17].

The time attribute of each heart disease risk factor represents the relationship between risk factor and the corresponding document creation time (DCT), which is similar to the temporal relationship between a clinical event and DCT in the 2012 i2b2 clinical NLP challenge [22], except that the value of the time attribute can be any combination of {"before", "during", or "after"} rather than a single variable consisting of {"before", "during", "after"}. Most state-of-the-art systems presented for the 2012 i2b2 clinical NLP challenge used machine learning-based methods to extract relationships between events and DCT [23,24]. For example, the best system proposed by Tang et al. adopted SVMs [23].

3. Material and methods

3.1. Dataset

The i2b2 challenge organizers manually annotated longitudinal records of 300 patients (1304 documents), from which 180 patients (790 documents) were used as a training set and the remaining 120 patients (514 documents) were used as a test set. The annotation guidelines defined a set of tags to indicate the presence and progression of diseases (diabetes, heart disease), associated risk factors (hyperlipidemia, hypertension, obesity status, family history, and smoking status), and associated medications. Each tag for the diseases and associated risk factors had one indicator value from its own set, while each tag for the associated medications could have two indicators (denoted by "type1" and "type2") to identify its category.

A brief description of each tag type in the challenge data is presented in Table 1. For more information, please refer to the annotation guidelines [25]. The challenge organizers released the data in two versions: complete and gold. The former provides the evidence (if any exists) for each tag and is used for system development. The latter only provides each tag itself without any evidence and is used for system evaluation.

Fig. 1 shows an example of a tag extracted from sample data (321-03.xml) in both versions, where the evidence associated with the tag is listed in the "text" field.

3.2. Overview of system

Our system identified each type of tag in the following order:

1. Extract evidence (if any exists) by type and indicator.
2. Determine attribute (i.e., time, if it exists).

By analyzing the evidence of tags, we found that the tags mainly fell into the following three categories:

1. Phrase-based tags, the evidence for which is provided explicitly in phrases.
2. Logic-based tags, the evidence for which is provided explicitly in phrases/sentences, but needs additional logical inferences, such as numerical comparisons.
3. Discourse-based tags, the evidence for which is not provided explicitly, but is embedded in clinical text fragments.

Table 2 lists these three categories of evidence-based tags, where the evidence is marked in bold and italics, followed by their tags in parenthesis. The evidence associated with phrase- and logic-based tags are very similar, with the difference being whether logical inference is further required after the phrases are located. For example, the blood pressure (BP) measurement "BP 140/80" is evidence of hypertension due to a high systolic pressure of 140 (Fig. 1). If the BP measurement of a patient is 120/80, "BP 120/80" will not qualify as evidence. Each tag listed in Table 1 may belong to multiple categories mentioned above based on the associated evidence and distinguished by its indicator(s).

The relationships between tag types listed in Table 1 and tag categories listed in Table 2 are shown in Table 3, where each item indicates to which category a tag with an indicator belongs.

Table 1
A brief description of each tag type used in the 2014 i2b2 clinical NLP challenge data.

Tag	Indicator	Attribute	Number	
			Training	Test
Diabetes	Mention, high A1c, high glucose	Time	1695	1180
CAD	Mention, event, test result, symptom	Time	1186	784
Hyperlipidemia	Mention, high cholesterol, high LDL	Time	1062	751
Hypertension	Mention, high blood pressure (high bp)	Time	1926	1293
Obesity status	Mention, BMI,[a] waist circumference	Time	433	262
Family history	Present, no present	NA[b]	790	514
Smoking status	Current, past, ever, never, unknown	NA	771	512
Medication	Metoprolol, . . . , lorquess	Time	8638	5674

[a] Body mass index.
[b] Not available.

Please cite this article in press as: Q. Chen et al., An automatic system to identify heart disease risk factors in clinical texts over time, J Biomed Inform (2015), http://dx.doi.org/10.1016/j.jbi.2015.09.002