Journal of Biomedical Informatics 63 (2016) 11–21

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A computational framework for converting textual clinical diagnostic criteria into the quality data model

Na Hong [a,b,1], Dingcheng Li [a,1], Yue Yu [c,1], Qiongying Xiu [d], Hongfang Liu [a], Guoqian Jiang [a,*]

[a] Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
[b] Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing, China
[c] Department of Medical Informatics, School of Public Health, Jilin University, Changchun, Jilin, China
[d] Computer Science, Winona State University, Rochester, MN, USA

ARTICLE INFO

Article history:
Received 1 March 2016
Revised 7 July 2016
Accepted 17 July 2016
Available online 19 July 2016

Keywords:
Diagnostic criteria
Quality data model
Natural language processing
cTAKES
Conditional random fields

ABSTRACT

Background: Constructing standard and computable clinical diagnostic criteria is an important but challenging research field in the clinical informatics community. The Quality Data Model (QDM) is emerging as a promising information model for standardizing clinical diagnostic criteria.
Objective: To develop and evaluate automated methods for converting textual clinical diagnostic criteria in a structured format using QDM.
Methods: We used a clinical Natural Language Processing (NLP) tool known as cTAKES to detect sentences and annotate events in diagnostic criteria. We developed a rule-based approach for assigning the QDM datatype(s) to an individual criterion, whereas we invoked a machine learning algorithm based on the Conditional Random Fields (CRFs) for annotating attributes belonging to each particular QDM datatype. We manually developed an annotated corpus as the gold standard and used standard measures (precision, recall and *f*-measure) for the performance evaluation.
Results: We harvested 267 individual criteria with the datatypes of Symptom and Laboratory Test from 63 textual diagnostic criteria. We manually annotated attributes and values in 142 individual Laboratory Test criteria. The average performance of our rule-based approach was 0.84 of precision, 0.86 of recall, and 0.85 of *f*-measure; the performance of CRFs-based classification was 0.95 of precision, 0.88 of recall and 0.91 of *f*-measure. We also implemented a web-based tool that automatically translates textual Laboratory Test criteria into the QDM XML template format. The results indicated that our approaches leveraging cTAKES and CRFs are effective in facilitating diagnostic criteria annotation and classification.
Conclusion: Our NLP-based computational framework is a feasible and useful solution in developing diagnostic criteria representation and computerization.

1. Introduction

The term "diagnostic criteria" designates the specific combination of signs, symptoms, and test results that clinicians used to determine correct diagnosis [1]. It is one of the most valuable sources of knowledge that can be used for supporting clinical decision making and improving patient care [2]. However, existing diagnostic criteria are scattered over different media such as medical textbooks, literature, and clinical practice guidelines, and they are usually described in unstructured free text without uniform standard. This situation hinders the efficient use of diagnostic criteria for supporting contemporary clinical decision making, which needs an integrated system with interoperable and computable processes.

One solution to better support clinical decision making is to make these diagnostic criteria computerized; however, it is costly and time-consuming for experts and clinicians to complete all of the tasks manually. To this end, on the one hand, the Natural Language Processing (NLP) technology could be used to enable automatically, or semi-automatically transforming diagnostic criteria into a computable format. On the other hand, a data model to represent diagnostic criteria is equally essential for its computerized implementation. Such a data model would enable the representation of diagnostic criteria in a structured, standard, and encoded framework to support many of the clinical applications in a scalable fashion. In order to investigate the model adaptability to diagnostic criteria, we previously evaluated the application feasibility

* Corresponding author at: Mayo Clinic, Department of Health Sciences Research, 200 First Street, SW, Rochester, MN 55905, USA.
E-mail address: Jiang.Guoqian@mayo.edu (G. Jiang).
[1] Co-first author.

http://dx.doi.org/10.1016/j.jbi.2016.07.016
1532-0464/

of the National Quality Forum (NQF) Quality Data Model (QDM) [3] through a data-driven approach, in which we manually analyzed the distribution and coverage of the data elements extracted from a collection of diagnostic criteria in QDM. The results demonstrated that the use of QDM is feasible in building a standards-based information model for representing computable diagnostic criteria [4].

The objective of the present study is to develop and evaluate automated methods for converting textual clinical diagnostic criteria into a structured format using QDM. We leverage clinical NLP tools to facilitate the computerization and standardization of diagnostic criteria. Specifically, we use a combination of the Clinical Text Analysis and Knowledge Extraction System (cTAKES)-supported and rule-based methods for extracting individual diagnostic criterion from full-text clinical diagnostic criteria. We also develop a machine learning algorithm based on the Conditional Random Fields (CRFs) to automatically annotate and classify the attributes of diagnosis events. Finally, we develop an integrated web-based system that automatically transforms textual diagnostic criteria into a standard QDM template by implementing the algorithms.

2. Background

2.1. Clinical NLP tools and information models

A number of tools and methods based on NLP technology have been reported and used in structuring free-text-based clinical text, such as clinical guidelines, clinical notes, and electronic health records (EHRs) [5,6]. Typical clinical NLP tools that could support term recognition and text annotation from clinical text include Health Information Text Extraction tool (HITex) [7], MetaMap [8], OpenNLP [9], and cTAKES [10]. Some studies compared the performance of these frequently used NLP tools, and the cTAKES shows satisfactory performance and usability [11,12]. cTAKES is an open-source Apache project and is an NLP system designed to extract information from EHR-based clinical free-text. cTAKES was built on the Unstructured Information Management Architecture (UIMA) framework, which is an open source framework designed by IBM and a series of comprehensive NLP methods [13]. Its modular architecture is composed of pipelined components combining rule-based and machine learning techniques [10]. These components exchange data using a standard data structure known as the Common Analysis System (CAS). CAS contains the original document with annotated results, and a powerful index system. The components of cTAKES are specifically trained for use in the clinical domain, and create rich linguistic and semantic annotations that can be utilized by clinical decision support systems, as well as in clinical and translational research [14].

Other than common tools in clinical NLP tasks, there are also many successful applications of machine learning algorithms [15–18] that are customized to support different scenarios and use cases. In recent years, the Conditional Random Fields (CRFs) algorithms demonstrated significant performance in the clinical NLP field in comparison with other machine learning algorithms [19–21]. In the CRFs applications, such as entity extraction and text classification, we usually wish to predict a vector $Y = [y_0; y_1; \ldots; y_m]$ of random variables given an observed feature vector $X = [x_0; x_1; \ldots; x_n]$, which requires us to label the words in a sentence with their corresponding features (i.e., contextual information) which subsequently used for training. The features can be part-of-speech (POS), neighboring words and word bigrams, prefixes and suffixes, capitalization, membership in domain-specific lexicons, semantic information of words, etc. Considering the advantage of CRFs in the contextual information understanding and decent NLP performance, we leveraged CRFs for the purpose of the attributes extraction and classification and the performance tuning in the present study.

Current efforts to develop international recommendation standard models in clinical domains have laid the foundation for modeling and representing computable diagnostic criteria. There are a number of clinical data models developed in related fields (e.g., QDM, Clinical Element Models [CEMs] [22], and HL7 Fast Health Interoperable Resources [FHIR] [23]). QDM is designed to allow EHRs and other clinical electronic systems to share a common understanding and interpretation of the clinical data. It allows quality measure developers and many clinical researchers or performers to clearly and unambiguously describe the data required to calculate the quality measure. Different from CEM and FHIR, QDM contains both a data model module and a logic module. The latter handles logic expressions elegantly with a collection of functions, logic operators and temporal operators. Therefore, we chose QDM as the information model for standard representation of diagnostic criteria in the present study.

2.2. Clinical text computerization and standardization

The related studies on clinical text computerization mainly include the following three aspects.

(1) Clinical guideline computerization and Computer Interpretable Guideline (CIG) Systems. Various computerized clinical guidelines and the decision support systems that incorporate the guidelines have been developed. Researchers have tried different approaches to computerizing clinical practice guidelines [24–27], but those guidelines cover many complex medical procedures; thus, the application of these studies in real-world clinical practice is still very limited. However, the methods used to computerize guidelines are valuable in addressing the issues in diagnostic criteria computerization.
(2) Clinical NLP technologies. Unstructured clinical text mainly exists in the form of clinical notes, eligibility criteria, and clinical guidelines. There is much precedent on the work of clinical NLP applications using machine learning, rule-based methods, and other novel methods [19,28,29]. These studies offer valuable contribution in exploring different methods to automatically process information in clinical text.
(3) Formalization method studies on clinical research data. Some previous studies investigated the eligibility criteria in clinical trial protocols, developed approaches for eligibility criteria extraction and semantic representation, and used hierarchical clustering for dynamic categorization of such criteria [28,30]. For example, EliXR provided a corpus-based knowledge acquisition framework that uses the Unified Medical Language System (UMLS) to standardize eligibility-concept encoding and to enrich eligibility-concept relations for clinical research eligibility criteria from text. QDM-based phenotyping methods used for identification of patient cohorts from EHR data also provide valuable reference on our work [31].

Although current studies on the computerization and standardization of diagnostic criteria are still immature, there are some studies that started working on the diagnostic criteria computerization and only focused on some particular diseases. Examples of such studies include the computerized diagnostic criterion of inclusion body myositis [32] or Brugada-type electrocardiograms [33]. However, few of the current studies are taken from the perspective of using a standard information model to build