INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS XXX (2013) XXX–XXX

journal homepage: www.ijmijournal.com

Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records

Roy J. Byrd[a,*], *Steven R. Steinhubl*[b], *Jimeng Sun*[a], *Shahram Ebadollahi*[a], *Walter F. Stewart*[c]

[a] *IBM T. J. Watson Research Center, Yorktown Heights, NY, United States*
[b] *Geisinger Medical Center, Center for Health Research, Danville, PA, United States*
[c] *Sutter Health, Research, Development, & Dissemination, Concord, CA, United States*

ARTICLE INFO

Article history:
Received 29 August 2012
Received in revised form
13 December 2012
Accepted 17 December 2012

Keywords:
Natural language processing
Text mining
Heart failure
Electronic health records
Diagnostic criteria

ABSTRACT

Objective: Early detection of Heart Failure (HF) could mitigate the enormous individual and societal burden from this disease. Clinical detection is based, in part, on recognition of the multiple signs and symptoms comprising the Framingham HF diagnostic criteria that are typically documented, but not necessarily synthesized, by primary care physicians well before more specific diagnostic studies are done. We developed a natural language processing (NLP) procedure to identify Framingham HF signs and symptoms among primary care patients, using electronic health record (EHR) clinical notes, as a prelude to pattern analysis and clinical decision support for early detection of HF.

Design: We developed a hybrid NLP pipeline that performs two levels of analysis: (1) At the criteria mention level, a rule-based NLP system is constructed to annotate all affirmative and negative mentions of Framingham criteria. (2) At the encounter level, we construct a system to label encounters according to whether any Framingham criterion is asserted, denied, or unknown.

Measurements: Precision, recall, and F-score are used as performance metrics for criteria mention extraction and for encounter labeling.

Results: Our criteria mention extractions achieve a precision of 0.925, a recall of 0.896, and an F-score of 0.910. Encounter labeling achieves an F-score of 0.932.

Conclusion: Our system accurately identifies and labels affirmations and denials of Framingham diagnostic criteria in primary care clinical notes and may help in the attempt to improve the early detection of HF. With adaptation and tooling, our development methodology can be repeated in new problem settings.

* *Corresponding author at*: IBM T. J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, USA. Tel.: +1 914 945 4968; fax: +1 914 945 4781.
E-mail address: roybyrd@us.ibm.com (R.J. Byrd).
1386-5056/$ – see front matter © 2013 Published by Elsevier Ireland Ltd.
http://dx.doi.org/10.1016/j.ijmedinf.2012.12.005

1. Introduction and objective

The individual and societal impact of heart failure (HF) is staggering. One in five US citizens over age 40 is expected to develop HF during their lifetimes. It is currently the leading cause of hospitalization among Medicare beneficiaries and, with an aging U.S. population, HF prevalence and related costs will only increase, as prevalence of HF is expected to double by 2030 [1]. Individual and societal burdens may be mitigated through early detection of HF and intervention with lifestyle changes and proven preventive therapies.

Identifying the early manifestations of HF in the primary care setting is not straightforward. HF is a complex pathophysiologically heterogeneous syndrome, with substantial individual variability in expression. Moreover, because the signs and symptoms are also expressed for multiple causal factors unrelated to HF (e.g., chronic obstructive pulmonary disease, venous insufficiency, kidney disease), both false positive and false negative rates of diagnosis are relatively high in primary care [2,3].

The Framingham heart failure criteria published in 1971 [4] are based on clinical data acquired in the 1950s and 1960s but are still the most common HF signs and symptoms documented by primary care physicians (PCPs) today, usually well before more specific diagnostic studies are considered. But, relatively little is known about how these criteria are documented by PCPs or the extent to which these criteria vary in their sensitivity and specificity to HF diagnosis. In fact, when originally developed, the Framingham criteria only identified approximately half of the patients who had previously been diagnosed clinically with HF [4]. While other clinical criteria for HF have been developed, the agreement among different criteria is poor to moderate at best [5].

Ambulatory care is rapidly changing, especially with regard to adoption of electronic health records (EHR). Despite the structured information in EHRs – such as diagnosis codes, medications, and lab results – large portions of EHR data are still in narrative text format, principally in clinical encounter notes and imaging notes. There are widely recognized barriers to the application of NLP tools to such data [6–8].

This paper presents results of using NLP to extract Framingham criteria from clinical notes of primary care patients with and without HF. This work is part of a larger project, called PredMED, which is focused on the early detection and management of HF [9,10]. In PredMED, the extracted criteria serve as features for various downstream statistical and machine-learning applications. To the best of our knowledge, there are no published studies of text extraction for the Framingham HF criteria as they are documented in primary care. Lin et al. [11] reported some success in using the MedLEE parser [12] on discharge summaries and radiology reports to predict ICD-9 codes for HF diagnosis. More recent work [13,14] is based on the Unstructured Information Management Architecture (UIMA) framework, as is ours. More generally, the NLP extraction work done within the i2b2 competition [15–20] is similar to work we describe herein, with the crucial difference that our EHR dataset does not have pre-existing reference standard annotations. We describe iterative annotation methods similar to those found in other NLP work on EHR entity extraction [21,23–26] that were essential to developing our reference standards.

2. Materials and methods

An NLP application was developed and validated for identifying affirmations and denials of fifteen of the seventeen Framingham criteria for HF shown in Table 1.

2.1. Source of data

Data for this study were obtained from the Geisinger Clinic (GC) primary care practice EHRs. The dataset consisted of the full encounter records for 6355 incident primary care HF patients diagnosed between 2003 and 2010, as previously described [27], and up to ten clinic-, sex-, and age-matched control patients for each HF case. There were 26,052 controls. In total, there were over 3.3 million clinical notes, comprising over 4 gigabytes of text. While there were 56 different encounter types, “Office Visit” accounted for 81% of all encounters, followed by “Case Manager” (8%) and “Radiology” (7%).

2.2. Tools

We built a text analysis pipeline (Fig. 1) to extract Framingham criteria, using LanguageWare [28] for basic text processing and the IBM LanguageWare Resource Workbench (LRW) to develop dictionaries and grammars. The resulting analytics were then inserted into a UIMA [29,30] pipeline, which provides for acquisition of the clinical note texts and the other steps in Fig. 1. We also used a concordance program [31] to avoid overtraining to the development encounters, by letting us understand the behavior of our analytics on the entire encounter corpus.

2.3. Methods

Our development and evaluation process comprised the following steps:

- A cardiologist and a linguist analyzed a development dataset of 65 encounter documents rich in Framingham criteria, to learn the linguistics of criteria mentions.
- The linguist used the NLP tools to build initial extractors for assertions and denials of Framingham criteria (this section).
- The clinical expert and linguist incrementally measured and improved the performance of the extractors on the development documents (Section 2.4).
- The clinical expert used annotation guidelines he developed to train coders, who manually created gold standard annotations on an evaluation dataset of 400 randomly selected encounter documents (Section 3).
- The linguist used the gold standard to measure the performance of the final extractors on two tasks, criteria extraction and encounter labeling (Section 4).

Text analysis involved the following tasks:

Please cite this article in press as: R.J. Byrd, et al., Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records, Int. J. Med. Inform. (2013), http://dx.doi.org/10.1016/j.ijmedinf.2012.12.005