International Journal of Medical Informatics 111 (2018) 83–89

Contents lists available at ScienceDirect

International Journal of Medical Informatics

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

Natural language processing of clinical notes for identification of critical limb ischemia

Naveed Afzal[a], Vishnu Priya Mallipeddi[b], Sunghwan Sohn[a], Hongfang Liu[a], Rajeev Chaudhry[c], Christopher G. Scott[a], Iftikhar J. Kullo[b], Adelaide M. Arruda-Olson[b,*]

[a] Department of Health Sciences Research, Mayo Clinic and Mayo Foundation, Rochester, MN, United States
[b] Department of Cardiovascular Diseases, Mayo Clinic and Mayo Foundation, Rochester, MN, United States
[c] Division of Primary Care Medicine, Knowledge Delivery Center and Center for Innovation, Mayo Clinic and Mayo Foundation, Rochester, MN, United States

ARTICLE INFO

Keywords:
Natural language processing
Electronic health records
Peripheral artery disease
Critical limb ischemia
Subphenotyping

ABSTRACT

Background: Critical limb ischemia (CLI) is a complication of advanced peripheral artery disease (PAD) with diagnosis based on the presence of clinical signs and symptoms. However, automated identification of cases from electronic health records (EHRs) is challenging due to absence of a single definitive International Classification of Diseases (ICD-9 or ICD-10) code for CLI.

Methods and results: In this study, we extend a previously validated natural language processing (NLP) algorithm for PAD identification to develop and validate a subphenotyping NLP algorithm (CLI-NLP) for identification of CLI cases from clinical notes. We compared performance of the CLI-NLP algorithm with CLI-related ICD-9 billing codes. The gold standard for validation was human abstraction of clinical notes from EHRs. Compared to billing codes the CLI-NLP algorithm had higher positive predictive value (PPV) (CLI-NLP 96%, billing codes 67%, $p < 0.001$), specificity (CLI-NLP 98%, billing codes 74%, $p < 0.001$) and F1-score (CLI-NLP 90%, billing codes 76%, $p < 0.001$). The sensitivity of these two methods was similar (CLI-NLP 84%; billing codes 88%; $p < 0.12$).

Conclusions: The CLI-NLP algorithm for identification of CLI from narrative clinical notes in an EHR had excellent PPV and has potential for translation to patient care as it will enable automated identification of CLI cases for quality projects, clinical decision support tools and support a learning healthcare system.

1. Introduction

Lower extremity peripheral artery disease (PAD) affects millions of people worldwide [1]. Advanced cases of PAD may manifest as critical limb ischemia (CLI) which is associated with considerable morbidity, mortality and high risk of major cardiovascular events [2]. Within one year of CLI diagnosis, 30% of patients undergo limb amputation while 25% die [3–5]. Despite the availability of state of the art revascularization procedures recommended by practice guidelines for treatment of CLI, high proportions of CLI patients undergo amputation without vascular evaluation in the previous year [6]. Due to population ageing and high prevalence of diabetes which are risk factors for CLI, it has been estimated that the consequent number of CLI patients is likely to increase in both developing and developed countries [7,8]. Moreover, CLI has been associated with significant health care resource utilization. The estimate of aggregate annual US national costs associated with CLI hospitalizations was approximately $4.2 billion in 2013–2014 while the 30-day readmission rates for CLI contributed to over $624 million in healthcare costs [9].

Electronic health records (EHRs) have been widely heralded for potential to improve the quality of patient care and as a source for rapid automated identification of patients for research studies [10]. However, the electronic ascertainment of CLI from EHRs has proved challenging due to absence of a single definitive ICD-9 or ICD-10 code. For this reason, prior studies have developed and validated billing code algorithms for ascertainment of CLI cases [11] using combinations of ICD-9 codes. Sensitivity of these billing code algorithms has varied by practice setting [11]. Importantly, the clinical diagnosis of CLI is based on the presence of signs and symptoms as recorded in clinical narrative [12] while billing codes are used primarily for administrative purposes. Billing codes are used to mine structured information while natural language processing (NLP) is used to extract meaningful information from unstructured data. ICD billing codes are used primarily for administrative transactions and reimbursements. Additionally, ICD codes are used for diverse secondary

* Corresponding author at: Cardiovascular Diseases, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, United States.
E-mail address: ArrudaOlson.Adelaide@mayo.edu (A.M. Arruda-Olson).

https://doi.org/10.1016/j.ijmedinf.2017.12.024
Received 5 September 2017; Received in revised form 17 December 2017; Accepted 27 December 2017

Table 1
CLI related keywords for CLI-NLP algorithm.

Diagnostic Keywords	Location Keywords
ischemia; ischemic ulcer; ischemic ulcers; ischemic wound; ischemic wounds; ischemic pressure wound; ischemic pressure wounds; gangrene; neuropathic ischemic wound; neuropathic ischemic wounds	limb; limbs; lower extremity; lower extremities; right lower extremity; left lower extremity; right lower extremities; left lower extremities; rle; lle; leg; legs; foot; feet; toe; toes; ankle; aorto bi-iliac; aorto bi-femoral; aorto iliac; aorta femoral; sfa; plantar; heel

purposes including epidemiology studies, cohort identification and for health services research [13]. Prior studies comparing administrative and clinical approaches indicated that administrative data may be less accurate for identification of certain patient characteristics [14–16]. Notably, the use of billing code algorithms for identification of other phenotypes from EHRs has also had disappointing accuracy, positive predictive value and/or sensitivity [17–19].

NLP applied to clinical narrative may overcome the limitations of billing code algorithms for identification of CLI by recognition of text which describes signs and symptoms used to establish a diagnosis. Indeed, previous studies have demonstrated that NLP methods outperform billing code algorithms for phenotype identification from narrative clinical notes from the EHR. Specifically, NLP-based phenotyping algorithms have been used for automated case identification for a variety of diseases including inflammatory bowel disease, multiple sclerosis, rheumatoid arthritis, asthma and pancreatic cancer [20–22]. In the present study, we developed an NLP-based algorithm for ascertainment of CLI from narrative clinical notes of a community-based PAD cohort and we compared the performance of the NLP algorithm with CLI-related billing codes. Both methods were compared to human abstraction as the gold standard. We tested the hypothesis that an NLP algorithm applied to narrative clinical notes will have superior performance compared to CLI-related billing codes for identification of CLI.

2. Methods

2.1. Study setting and population

The study was conducted at Mayo Clinic, Rochester MN and used the resources of the Rochester Epidemiology Project (REP) to assemble a community-based PAD cohort from Olmsted County [23,24]. The REP is an integrated health information system that links medical records of all residents of Olmsted County [23]. In the PAD inception cohort, all patients were diagnosed with PAD by an ankle-brachial index (ABI) test performed in the Mayo Clinic noninvasive vascular laboratory using standardized protocols [1]. The institutional review board approved the study and informed consent was attained for all subjects.

2.2. Study design

We retrieved all clinical notes through June 2015 of patients participating in this study from the Mayo clinical data warehouse. We applied the previously validated knowledge-driven NLP algorithm (PAD-NLP algorithm) to the dataset to automatically ascertain PAD status [25]. The PAD-NLP algorithm automatically ascertained cases from clinical notes using PAD-related keywords and a set of rules for classification of a PAD patient. The PAD-NLP algorithm consisted of two main components: text-processing and patient classification. The text-processing component analyzed the text of each clinical note by breaking down sentences into words using MedTagger [26], an open source clinical NLP system, and identified PAD-related concepts which were mapped to specific categories that were later used for patient classification. NLP technology was used for automated extraction and encoding of clinical information from narrative clinical notes. MedTagger is a knowledge driven clinical NLP system which enables sentence detection, word tokenization, section identification, contextual information and concept identification. After sentence detection, word tokens are identified using space between two words. As recommended by the HL-7 CDA standard [27], clinical notes are divided into sections, (e.g. “impression, report and plan” and “diagnosis”) and MedTagger recognizes these note sections. Additionally, MedTagger identifies contextual information from clinical notes including: assertion, temporality and experiencer. We used MedTagger to identify assertion, temporality and experiencer of CLI-related keywords from clinical notes.

The steps for the NLP algorithm were: first, a list of PAD-related terms was identified by cardiovascular experts from narrative clinical notes. Second, these terms were mapped to corresponding concepts and their synonyms in the unified medical language system (UMLS) Metathesaurus which were also added to this list. Third, this list was further expanded during the interactive refinement of PAD-NLP algorithm [25] when additional synonyms and other lexical variants were identified in the clinical notes and added. Fourth, the PAD-NLP algorithm produced output on two levels: document and patient levels. At the document level, each clinical note was processed to find PAD-related keywords and if found produced output in the form of PAD-related keywords along with ± 2 sentences from clinical notes. Fifth, as CLI keywords were included in the list of keywords used by the PAD-NLP algorithm (Table 1) these concepts were also identified. Their relevant category along with certainty (positive, negative or possible), temporality (current, historical) and its experiencer (patient or someone else) were also extracted during the text processing phase.

The CLI-related keywords were identified by cardiovascular experts and included in the list of keywords used in the PAD-NLP algorithm. A subphenotyping algorithm for identification of CLI cases was developed and used the document level (i.e. note level) output of the PAD-NLP algorithm and narrowed the focus of the algorithm to identify the subset of PAD cases with CLI. Hence, the CLI-NLP algorithm was derived from the PAD-NLP algorithm (Fig. 1). The performance of the billing codes for identification of CLI was compared with results obtained by the CLI-NLP algorithm. Both methods were then compared to human chart abstraction as the gold standard for validation (Fig. 1).

2.3. Rules for the CLI-NLP algorithm

The CLI-NLP algorithm identified keywords from the document level output. Keywords were categorized as diagnostic and location (Table 1). The list of diagnostic keywords included both the isolated mechanisms for wound or combined mechanisms (wounds which occurred as a consequence of a combination of co-existing mechanisms, e.g., ischemic/pressure wound or neuropathic ischemic wound – see Table 1). Importantly, “ischemia” (and the lexicon variations for this term) had to be considered as one of these mechanisms. In the absence of these criteria patients were classified as not having CLI (controls).

The rule for CLI cases was:

- One diagnostic keyword + one location keyword from Table 1 within two sentences anchored by a diagnostic keyword in the same note.

For controls (not having CLI), we used the following rule:

- If not satisfied, the rule for CLI described above.