Journal of Biomedical Informatics 54 (2015) 186–190

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Regular expression-based learning to extract bodyweight values from clinical notes

Maureen A. Murtaugh [a,b,*], Bryan Smith Gibson [a,b], Doug Redd [a,c], Qing Zeng-Treitler [a,c]

[a] IDEAS Center, Veterans Administration, Salt Lake City Health Care System, Salt Lake City, UT, United States
[b] Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT, United States
[c] Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, UT, United States

ARTICLE INFO

Article history:
Received 6 November 2014
Accepted 24 February 2015
Available online 5 March 2015

Keywords:
Natural language processing
Bodyweight
Text classification

ABSTRACT

Background: Bodyweight related measures (weight, height, BMI, abdominal circumference) are extremely important for clinical care, research and quality improvement. These and other vitals signs data are frequently missing from structured tables of electronic health records. However they are often recorded as text within clinical notes. In this project we sought to develop and validate a learning algorithm that would extract bodyweight related measures from clinical notes in the Veterans Administration (VA) Electronic Health Record to complement the structured data used in clinical research.

Methods: We developed the Regular Expression Discovery Extractor (REDEx), a supervised learning algorithm that generates regular expressions from a training set. The regular expressions generated by REDEx were then used to extract the numerical values of interest.

Methods: To train the algorithm we created a corpus of 268 outpatient primary care notes that were annotated by two annotators. This annotation served to develop the annotation process and identify terms associated with bodyweight related measures for training the supervised learning algorithm. Snippets from an additional 300 outpatient primary care notes were subsequently annotated independently by two reviewers to complete the training set. Inter-annotator agreement was calculated.

Methods: REDEx was applied to a separate test set of 3561 notes to generate a dataset of weights extracted from text. We estimated the number of unique individuals who would otherwise not have bodyweight related measures recorded in the CDW and the number of additional bodyweight related measures that would be additionally captured.

Results: REDEx's performance was: accuracy = 98.3%, precision = 98.8%, recall = 98.3%, F = 98.5%. In the dataset of weights from 3561 notes, 7.7% of notes contained bodyweight related measures that were not available as structured data. In addition 2 additional bodyweight related measures were identified per individual per year.

Conclusion: Bodyweight related measures are frequently stored as text in clinical notes. A supervised learning algorithm can be used to extract this data. Implications for clinical care, epidemiology, and quality improvement efforts are discussed.

1. Introduction

The use of Electronic Health Record (EHR) data in conjunction with data extraction and categorization tools (e.g. clinical phenotyping), holds great potential to improve clinical practice [6,10] and clinical epidemiology [2]. However, challenges related to data completeness and data quality need to be addressed to maximize the effectiveness of these efforts. For example bodyweight related measures (weight, height, abdominal circumference), are needed when clinicians calculate medication dosages based on body surface area (BSA) [11], or use body mass index (BMI) to estimate risk of cardiovascular disease, diabetes or cancer (Institute). Similarly, epidemiologists rely on bodyweight measures when determining novel associations such as the recently reported association between bodyweight and mortality due to influenza and pneumonia [4].

Despite the critical importance of bodyweight data for clinical care and research, several evaluations have pointed out that these

[*] Corresponding author at: Division of Epidemiology, University of Utah, 295 Chipeta Way, Salt Lake City, UT 84132, United States. Fax: +1 (801) 581 3623.
E-mail address: Maureen.Murtaugh@hsc.utah.edu (M.A. Murtaugh).

http://dx.doi.org/10.1016/j.jbi.2015.02.009
1532-0464/© 2015 Elsevier Inc. All rights reserved.

data are frequently unavailable as structured data. For example researchers at the group health cooperative, testing the ability to use EHR data to calculate cardiac risk, found that among the records of 122,270 individuals, 11.5% were missing data for either height weight or both [5]. Similarly, Das et al. reported that among 1.8 Million Veterans who received outpatient care at VA facilities in the year 2000, 50.4% had no height or weight recorded as structured data [3]. More recently, Littman et al. reported that 32.8% of records of 173,127 veterans in the northwestern US were missing structured data for weight or height [7]. Since anecdotal reports suggested that in many cases individuals' heights and weights were measured during these visits, but the data was recorded as text in the clinical note, our research team felt that this was an important use case for information extraction.

In this project we sought to develop and validate a learning algorithm that would extract bodyweight related measures (weight, height, BMI, abdominal circumference) recorded in clinical notes from the VA's electronic Health record. We were motivated to explore this as an example of the potential to supplement structured data with data stored in text in order to fill in gaps in repeatedly measured clinical data. Our first aim was to determine how well we could capture weight, height, BMI and/or abdominal girth from outpatient notes. Our second aim was to determine the proportion of the data in the notes that was unique data.

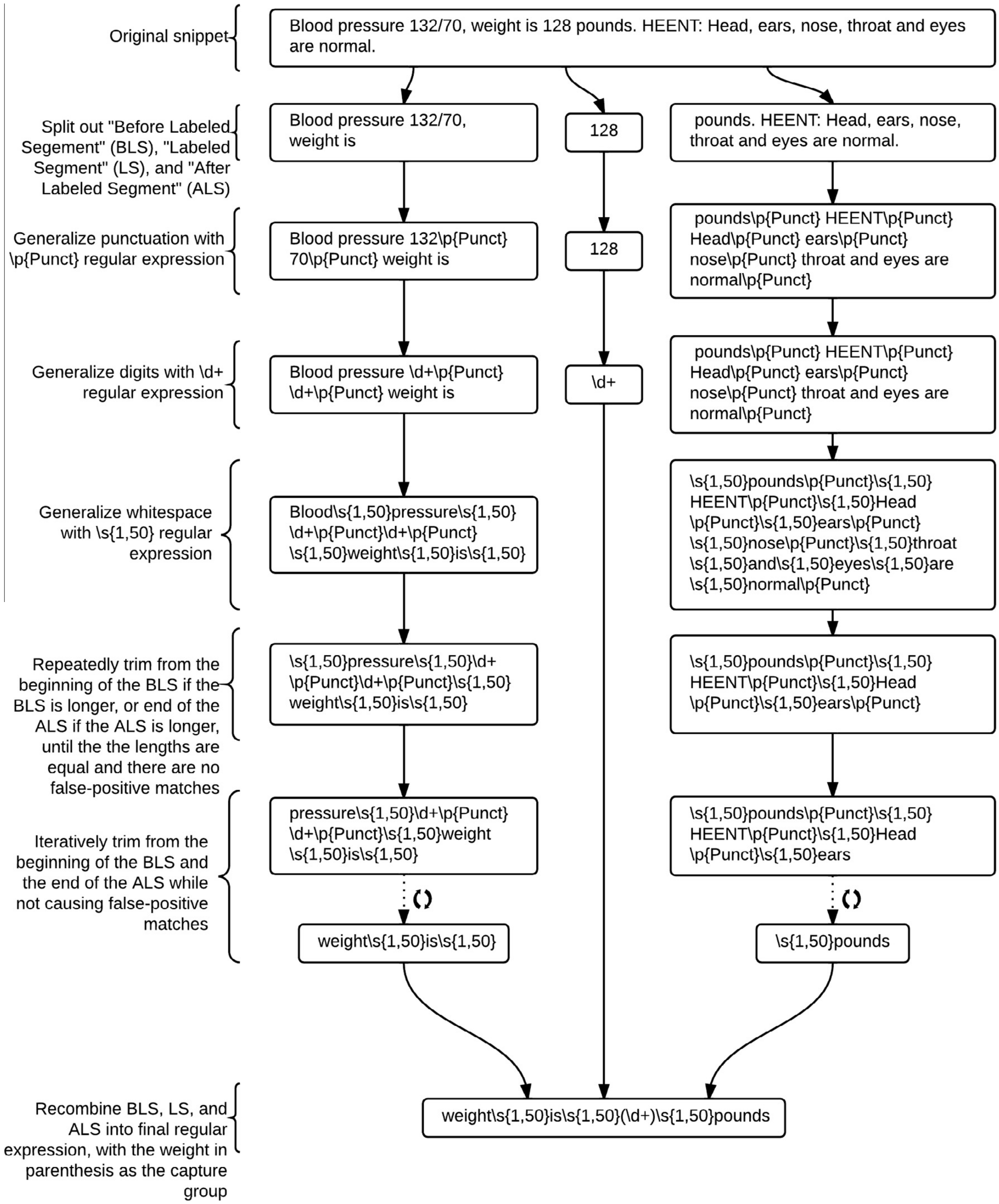
Original snippet
Blood pressure 132/70, weight is 128 pounds. HEENT: Head, ears, nose, throat and eyes are normal.
Split out "Before Labeled Segement" (BLS), "Labeled Segment" (LS), and "After Labeled Segment" (ALS)
Blood pressure 132/70, weight is
128
pounds. HEENT: Head, ears, nose, throat and eyes are normal.
Generalize punctuation with \p{Punct} regular expression
Blood pressure 132\p{Punct} 70\p{Punct} weight is
128
pounds\p{Punct} HEENT\p{Punct} Head\p{Punct} ears\p{Punct} nose\p{Punct} throat and eyes are normal\p{Punct}
Generalize digits with \d+ regular expression
Blood pressure \d+\p{Punct} \d+\p{Punct} weight is
\d+
pounds\p{Punct} HEENT\p{Punct} Head\p{Punct} ears\p{Punct} nose\p{Punct} throat and eyes are normal\p{Punct}
Generalize whitespace with \s{1,50} regular expression
Blood\s{1,50}pressure\s{1,50} \d+\p{Punct}\d+\p{Punct} \s{1,50}weight\s{1,50}is\s{1,50}
\s{1,50}pounds\p{Punct}\s{1,50} HEENT\p{Punct}\s{1,50}Head \p{Punct}\s{1,50}ears\p{Punct} \s{1,50}nose\p{Punct}\s{1,50}throat \s{1,50}and\s{1,50}eyes\s{1,50}are \s{1,50}normal\p{Punct}
Repeatedly trim from the beginning of the BLS if the BLS is longer, or end of the ALS if the ALS is longer, until the the lengths are equal and there are no false-positive matches
\s{1,50}pressure\s{1,50}\d+ \p{Punct}\d+\p{Punct}\s{1,50} weight\s{1,50}is\s{1,50}
\s{1,50}pounds\p{Punct}\s{1,50} HEENT\p{Punct}\s{1,50}Head \p{Punct}\s{1,50}ears\p{Punct}
Iteratively trim from the beginning of the BLS and the end of the ALS while not causing false-positive matches
pressure\s{1,50}\d+\p{Punct} \d+\p{Punct}\s{1,50}weight \s{1,50}is\s{1,50}
\s{1,50}pounds\p{Punct}\s{1,50} HEENT\p{Punct}\s{1,50}Head \p{Punct}\s{1,50}ears
weight\s{1,50}is\s{1,50}
\s{1,50}pounds
Recombine BLS, LS, and ALS into final regular expression, with the weight in parenthesis as the capture group
weight\s{1,50}is\s{1,50}(\d+)\s{1,50}pounds

Fig. 1. Example of the creation of a standardized regular expression by REDEx.