Journal of Biomedical Informatics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Predicting changes in systolic blood pressure using longitudinal patient records

John Wes Solomon *, Rodney D. Nielsen

University of North Texas, Denton, TX, United States

ARTICLE INFO

Article history:
Received 16 February 2015
Revised 24 June 2015
Accepted 29 June 2015
Available online xxxx

Keywords:
Prediction
Clinical
NLP
Informatics
SBP
Blood pressure

ABSTRACT

Objective: This paper introduces a model that predicts future changes in systolic blood pressure (SBP) based on structured and unstructured (text-based) information from longitudinal clinical records.
Method: For each patient, the clinical records are sorted in chronological order and SBP measurements are extracted from them. The model predicts future changes in SBP based on the preceding clinical notes. This is accomplished using least median squares regression on salient features found using a feature selection algorithm.
Results: Using the prediction model, a correlation coefficient of 0.47 is achieved on unseen test data ($p < .0001$). This is in contrast to a baseline correlation coefficient of 0.39.

1. Introduction

The primary goal of this research is to demonstrate that the output of an NLP system capable of automatically producing annotations like the ones provided by the organizers of the i2b2/UTHealth 2014 challenge [1] can be used to predict future medical events. Systems with the ability to predict future medical events based on electronic medical records (EMRs) could be used to suggest that patients and medical professionals pay closer attention to certain risk factors and events associated with the condition(s) the system is designed to predict. In this paper, a model that predicts future changes in systolic blood pressure (SBP) using time sensitive information is presented and evaluated. Such a system could help doctors recognize when to add or change medications and let patients know they should change their eating or exercise habits, among other possibilities. The presented model is novel in the sense that it is the first to predict changes in blood pressure using a feature space that explicitly takes into account the amount of time between the observation of feature values and the prediction date. Similar models that use different features, but similar feature representations may be effective for predicting other medical outcomes given appropriate training data.

2. Related work

A 1982 paper by Sparrow et al. [2] presents a model for predicting changes in blood pressure (BP). Although this work does not fall into the realm of clinical NLP, it involves the prediction of changes in BP on a patient by patient basis. The features used to inform the prediction model presented by Sparrow and colleagues could, in theory, be extracted from the electronic medical record. The prediction task presented in the paper is as follows. Given two BP measurements where the second measurement is taken after the first, and patient data collected on the day of the first BP measurement, predict the slope of the trend-line formed by the first two BP measurements and a third unseen BP measurement taken after the second one. The amount of time between observations is not taken into account when setting feature values and the model is not designed to make predictions when feature values are not known.

A 2012 paper by Fava et al. [13] investigates the predictive power of a genetic risk score (GRS) derived using 29 independent single nucleotide polymorphisms for predicting changes in blood pressure between an initial measurement and a follow-up examination when compared to other predictive variables including demographic, anthropometric, socioeconomic, and lifestyle data. The average time between the initial and follow-up examination was 23 years. Like the Sparrow paper, the amount of time between the initial and follow-up examination is not taken into account.

* Corresponding author.
E-mail addresses: John.Solomon@my.UNT.edu (J.W. Solomon), Rodney.Nielsen@UNT.edu (R.D. Nielsen).

http://dx.doi.org/10.1016/j.jbi.2015.06.024
1532-0464/

One significant way the present work differs from Sparrow's and Fava's is that the proposed model accounts for the amount of time between measurements and mentions of medical concepts. First, the model for predicting SBP changes accounts for the amount of time between measurements by using a logarithmic weighting scheme that gives more weight to recent numeric measurements. The intuition behind this is that more recent measurement values will be closer to the current value than less recent ones. The model also utilizes binary features tied to the relative time of a variable observation with respect to the prediction date. Finally, the model is capable of making predictions without all feature values being known.

These are not the only examples of work outside the realm of clinical NLP that predicts medical events on a patient by patient basis. Other notable examples include models that predict morbidity following gastrectomy [3] and respiratory morbidity among those with low birth weight [4]. Like the Sparrow paper, time is not taken into account when setting the feature values for these models and all feature values are assumed to be known.

An example of work within the realm of clinical NLP where future medical events are predicted on a patient by patient basis is a paper by Bihorac and colleagues [5]. The paper presents a model that predicts patient mortality after surgery.

Numerous clinical NLP papers refer to their work as *prediction* (e.g., Himes et al. [6] and Chen et al. [7]), but the term is being used in the sense of classification. In these papers, the classifier's predictions are determined based on features or measurements from the same time to which the prediction pertains.

3. Prediction task

The goal of the present work is to provide a patient and their physician a prediction of the patient's future SBP measurement in order to help ensure appropriate proactive health care. The formally defined prediction task is as follows. Given a list of patient records $\mathbf{R} = \langle r_1, r_2, \ldots, r_k, \ldots r_n \rangle$ sorted in chronological order according to the patient's visit date, the task is to predict the change in SBP measurement from r_k to r_n, where r_n is the chronologically latest patient record that includes an SBP measurement (and is thus best suited to act as a *future* date) and r_k is the record nearest in time to r_n that also includes an SBP measurement, while only utilizing information present in the records that predated r_n. Since utilizing any of the information from visits after r_n (i.e., $\{r_{n+1}, r_{n+2}, \ldots\}$) would obviously violate the goal of predicting a *future* outcome, these records are excluded from all further analysis or discussion. d_i denotes the visit date associated with r_i.

4. Data

The i2b2/UTHealth 2014 dataset for track 2 (identifying risk factors for heart disease over time) [14] is used to train and evaluate the system. The training portion of the data contains 790 individual records from 175 patients. The test portion consists of 542 individual records for 118 patients. Every patient in the dataset has been diagnosed with diabetes. After eliminating all patients from consideration where less than three records contain at least one SBP measurement, 514 records for 138 patients remain in the training data, and 309 records for 82 patients remain in the test data. Each record in the i2b2/UTHealth 2014 dataset has annotations associated with it. The annotations are used to set the values of the features that inform the prediction model. The system utilizes annotations that denote mentions of medical conditions, medications, and tobacco use.

5. Extracting systolic blood pressure measurements and other numeric values using regular expressions

The system uses regular expressions to extract SBP measurements as well as the values of measurements used to set all features described in Section 6.3 and some of the features described in Section 6.4. A regular expression encodes text patterns in strings. All substrings in a document that match the text pattern defined in a regular expression can be easily extracted. For example, the pattern string the system uses to extract weight measurements is "\bweight\s*(\s+|:)\s*\d+". This pattern extracts all instances of the word "weight" preceded by a word boundary on the left side followed by whitespace with an optional single colon in the middle and a numeric string containing one or more characters. This regular expression extracts the substrings, "weight 100" and "weight: 210" but does not extract "weight is 100".

BP measurement candidates are identified by extracting all substrings that match the regular expression "\d+/\d+" (a numeric string followed by a forward slash and another numeric string). The system uses a heuristic to filter out implausible BP values. Among the candidate substrings, if the numerator is between 40 and 300 and the denominator is between 30 and 200, the numerator of the fraction is identified by the system as an SBP measurement. Because a visit sometimes includes multiple SBP readings, we use the average of all SBP measurements in the patient record, $AVG_{SBP}(r)$, to approximate the true value of SBP. Regular expressions have the potential to extract incorrect values and miss values. When applied to training data and used to inform features used to predict actual changes in SBP, values extracted by regular expressions represent noisy, but potentially informative features. When applied to the BP measurements used to determine the changes in SBP upon which the performance evaluation is based, this presents a methodological problem. As such, we conducted two reviews to determine the accuracy of SBPs extracted using the aforementioned regular expression and heuristic. The first review involved annotating 65 randomly selected fraction substrings in the training data as being a BP measurement or not. Among the 65 fraction substrings 44 represented BP measurements and 21 did not. No errors were found in this review. Because our aim is to predict the change in SBP from the second most recent note containing an SBP measurement to the most recent note containing an SBP measurement it is critical that we ensure the SBP values in these notes are correct for the test set. As such, we manually reviewed these SBP measurements and found five cases where the values extracted by the regular expression and filtering heuristic did not represent SBP measurements. We manually removed these measurements from the test set. Occasionally, an SBP measurement in a note was from another point in time. From a methodological standpoint, this is problematic if it happens in the last or second to last note containing an SBP measurement in the test data, as the average SBP in those two notes are used to determine the gold standard SBP change. Fortunately, this occurs infrequently, and in cases where it does occur, the measurements were taken a short time prior to the visit.

6. Features

Each feature encodes information about a single patient. Most of the features fall into one of two categories: "Binary Time Sensitive Features" and "Real-Valued Time Adjusted Features". Features that do not fall into one of these two categories will be discussed on an individual basis. Gold standard annotations for medical events provided by the organizers of the i2b2/UTHealth 2014 challenge are used to directly set the values for some of the features that inform the prediction system. Using the gold standard

Please cite this article in press as: J.W. Solomon, R.D. Nielsen, Predicting changes in systolic blood pressure using longitudinal patient records, J Biomed Inform (2015), http://dx.doi.org/10.1016/j.jbi.2015.06.024