Journal of Biomedical Informatics 57 (2015) 219–224

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Classifying individuals based on a densely captured sequence of vital signs: An example using repeated blood pressure measurements during hemodialysis treatment

Benjamin A. Goldstein [a,b,*], Tara I. Chang [c], Wolfgang C. Winkelmayer [d]

[a] Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, United States
[b] Center for Predictive Medicine, Duke Clinical Research Institute, Durham, NC, United States
[c] Division of Nephrology, Stanford University School of Medicine, Palo Alto, CA, United States
[d] Section of Nephrology, Baylor College of Medicine, Houston, TX, United States

ARTICLE INFO

Article history:
Received 26 March 2015
Revised 18 June 2015
Accepted 7 August 2015
Available online 13 August 2015

Keywords:
Classification
Electronic health records
Functional data analysis
Hemodialysis
Splines

ABSTRACT

Electronic Health Records (EHRs) present the opportunity to observe serial measurements on patients. While potentially informative, analyzing these data can be challenging. In this work we present a means to classify individuals based on a series of measurements collected by an EHR. Using patients undergoing hemodialysis, we categorized people based on their intradialytic blood pressure. Our primary criteria were that the classifications were time dependent and independent of other subjects. We fit a curve of intradialytic blood pressure using regression splines and then calculated first and second derivatives to come up with four mutually exclusive classifications at different time points. We show that these classifications relate to near term risk of cardiac events and are moderately stable over a succeeding two-week period. This work has general application for analyzing dense EHR data.

1. Introduction

Compared to other data sources used in medical research, Electronic Health Records (EHRs) contain a very heterogeneous range of data features and structures. This diversity provides many analytic opportunities as well as challenges. One of the unique aspects of some EHRs is the ability to capture dense longitudinal information on patients. These data may consist of, for example, respiratory measurements from an intensive care unit [1] or heart rate measurements from a surgical procedure [2]. Such data present the opportunity to observe and analyze a patient's changing vital signs over time.

In our current use case, we have blood pressure measurements taken during a hemodialysis (HD) session. Maintenance HD is an outpatient treatment that individuals with end stage kidney disease (ESRD) receive on a regular basis. A typical session occurs three times a week and lasts 3–4.5 h. During an HD session, regular blood pressure measurements are captured and stored in the electronic health record. With measurements in approximately 15–30 min increments, a typical session will produce 6–12, so called "intradialytic," blood pressure readings. In recent years there has been growing interest in how intradialytic blood pressure variability relates to risk of adverse events [3].

A typical question that one may consider is: how do changes in intradialytic blood pressure relate to risk of a cardiovascular event? This raises the question of how best to analyze these data. If only one blood pressure measurement were available, one could simply regress the probability of cardiac event onto the blood pressure measurement in the form of a logistic regression. Extensions to this simple model have been proposed in the presence of multiple measurements [4–6]. The challenge with these models is that they are difficult to implement and hard to interpret [7]. Instead, work in this area has focused on summarizing these measurements either as a mean, standard deviation, or maximum blood pressure differences during dialysis [3]. Such summary measures potentially lead to loss of information. Moreover, because the summary metrics are continuous, it makes identifying high risk patients challenging – though many summary measures will get broken into percentile categories.

Instead of summarizing individual blood pressure measures, we consider the task of classifying people based on their blood pressure pattern. By grouping people into mutually exclusive categories, the intention is to identify a high risk category upon which surveillance or treatment interventions could be targeted.

* Corresponding author at: Department of Biostatistics and Bioinformatics, 2424 Erwin Road, Suite 1104, Durham, NC 27705, United States.
E-mail address: ben.goldstein@duke.edu (B.A. Goldstein).

http://dx.doi.org/10.1016/j.jbi.2015.08.010
1532-0464/

Previous work has considered clustering individuals based on curves [8]. While these methods address the classification problem, one drawback is that the clusters are not intrinsic to an individual's measurements but defined through the joint-distribution of all individuals in the sample. This means that given a different training sample, individuals may find themselves grouped into a different cluster. Moreover, with such cluster analyses it is not clear how best to classify a new individual.

In this paper, we suggest a novel approach for classifying people based on a sequence of measurements. Our goal is to create a simple classification system that can be performed both analytically as well as heuristically. We use regression splines to fit individual curves and then calculate derivatives of these curves to classify individuals based on their patterns at different time points. Our approach is able to classify individuals into mutually exclusive groups, independent of the classifications of others. We then show how these classifications relate to future risk of cardiovascular events. While our data are derived from patients undergoing HD, we emphasize that this approach is appropriate for any dense sequence of measurements.

2. Methods

2.1. Data sources

We used two data sources for the current analysis: the United States Renal Data System (USRDS) and the EHR from DaVita Inc., a large dialysis organization. The USRDS is a national registry that includes almost all persons with treated ESRD [9]. Its backbone consists of medical claims submitted to Medicare, which is mandated by law to provide coverage to the majority of patients with treated ESRD, regardless of age. DaVita Inc. is the second largest provider of outpatient dialysis in the United States. Its EHR contains detailed information including hemodynamic parameters from each dialysis session. We used an anonymous crosswalk provided by the USRDS Coordinating Center to link the two datasets under a data use agreement between the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and one of the authors (WCW).

2.2. Cohort

We selected all patients with ESRD for at least 90 days, who received maintenance HD at a DaVita facility for the 2-week period between October 5, 2008 and October 19, 2008. This was an arbitrarily chosen interval meant to represent typical HD. Since our focus was on typical sessions, we required all patients to be on thrice weekly HD and to have attended exactly 6 sessions during these two weeks. We further required that each session be at least 3 h and no more than 4.5 h in duration, and patients could not be on nocturnal HD. Additionally, to avoid temporal effects, all sessions for an individual had to start either in the morning (5 am–11 am) or afternoon (11 am–5 pm). Finally, to allow for ascertainment of medical claims information from the USRDS, patients had to have Medicare Parts A&B coverage by the initiation of the ascertainment window.

2.3. Variables

2.3.1. Intradialytic blood pressure

In the DaVita EHR, systolic blood pressure (SBP) is measured and recorded approximately every 15–30 min by an automated sphygmomanometer during each dialysis session. We removed SBP values that were <50 mmHg, >250 mmHg, or that were lower than the diastolic BP measurements. For each patient, we used SBP information from the six dialysis sessions during the two-week period specified above.

2.3.2. Additional covariates

We abstracted data on patient age, sex, and race/ethnicity (white, black, Hispanic, other), time of session start (morning *versus* afternoon), average SBP at start of HD and at different time-points during HD (see below), and time since diagnosis of ESRD (i.e. dialysis "vintage").

2.3.3. Outcomes

To assess the stability of the classifications, we abstracted SBP measurements during all dialysis sessions in the subsequent two weeks (October 20–November 1, 2008). No restrictions were placed on the timing and number of sessions during this period. We also assessed whether the SBP categorizations were associated with a composite cardiovascular outcome. Using ICD-9 and death codes we ascertained events in the 90 days following the original two-week SBP categorization period: fatal or non-fatal myocardial infarction (MI) (410.**), stroke (433.**, 434.**, 436.**, 437.1*, 437.9*), or cardiac arrest (427.5).

2.4. Statistical analysis

2.4.1. Classifying serial measurements

Our analytic challenge was to characterize a series of blood pressure measurements over a time period. As others have described, intradialytic SBP typically follows a non-linear pattern that drops sharply during the first hour of HD and levels off thereafter [10]. However as shown in Fig. 1 this pattern is not consistent across all people. In devising a classification scheme we considered 2 goals:

1. We did not want a global classification, but wanted the classifications to take into account the temporality of the measurements.
2. We wanted our classifications scheme for an individual to be independent of the classifications of anyone else.

As a first step, for each individual over the two week period, we modeled the series of SBP measurements as non-linear curves. Multiple procedures exist for curve fitting including, polynomial regression, regression splines, smoothing splines, and locally

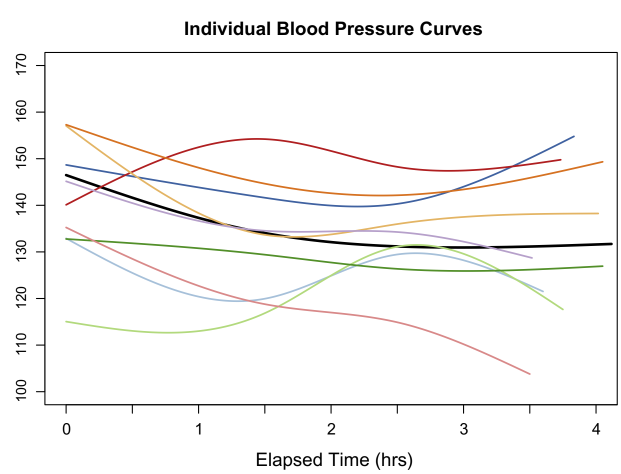

Fig. 1. Individual systolic blood pressure curve for nine random individuals. The black curve represents the average curve for the entire sample. While the average curve shows the expected blood pressure pattern – drop during the first hour and then levelling off – there is noticeable variation across people.