Journal of Biomedical Informatics 78 (2018) 87–101

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Estimating summary statistics for electronic health record laboratory data for use in high-throughput phenotyping algorithms

D.J. Albers[a,*], N. Elhadad[a], J. Claassen[b], R. Perotte[c], A. Goldstein[a], G. Hripcsak[a]

[a] Department of Biomedical Informatics, Columbia University, 622 West 168th Street, New York, NY, USA
[b] Department of Neurology, Columbia University, 710 West 168th Street, New York, NY 10032, USA
[c] Value Institute, New York Presbyterian Hospital, 601 West 168th Street New York, NY 10032, USA

ARTICLE INFO

Keywords:
Electronic health record
Kullback-Leibler divergence
Summary statistic
phenotyping
Laboratory tests

ABSTRACT

We study the question of how to represent or summarize raw laboratory data taken from an electronic health record (EHR) using parametric model selection to reduce or cope with biases induced through clinical care. It has been previously demonstrated that the health care process (Hripcsak and Albers, 2012, 2013), as defined by measurement context (Hripcsak and Albers, 2013; Albers et al., 2012) and measurement patterns (Albers and Hripcsak, 2010, 2012), can influence how EHR data are distributed statistically (Kohane and Weber, 2013; Pivovarov et al., 2014). We construct an algorithm, PopKLD, which is based on information criterion model selection (Burnham and Anderson, 2002; Claeskens and Hjort, 2008), is intended to reduce and cope with health care process biases and to produce an intuitively understandable continuous summary. The PopKLD algorithm can be automated and is designed to be applicable in high-throughput settings; for example, the output of the PopKLD algorithm can be used as input for phenotyping algorithms. Moreover, we develop the PopKLD-CAT algorithm that transforms the continuous PopKLD summary into a categorical summary useful for applications that require categorical data such as topic modeling. We evaluate our methodology in two ways. *First*, we apply the method to laboratory data collected in two different health care contexts, primary versus intensive care. We show that the PopKLD preserves known physiologic features in the data that are lost when summarizing the data using more common laboratory data summaries such as mean and standard deviation. *Second*, for three disease-laboratory measurement pairs, we perform a phenotyping task: we use the PopKLD and PopKLD-CAT algorithms to define high and low values of the laboratory variable that are used for defining a disease state. We then compare the relationship between the PopKLD-CAT summary disease predictions and the same predictions using empirically estimated mean and standard deviation to a gold standard generated by clinical review of patient records. We find that the PopKLD laboratory data summary is substantially better at predicting disease state. The PopKLD or PopKLD-CAT algorithms are not meant to be used as phenotyping algorithms, but we use the phenotyping task to show what information can be gained when using a more informative laboratory data summary. In the process of evaluation our method we show that the different clinical contexts and laboratory measurements necessitate different statistical summaries. Similarly, leveraging the principle of maximum entropy we argue that while some laboratory data only have sufficient information to estimate a mean and standard deviation, other laboratory data captured in an EHR contain substantially more information than can be captured in higher-parameter models.

1. Introduction

Electronic health record (EHR) data offer us the opportunity to carry out clinical research on a broad population relatively quickly while minimizing both the financial and human costs because the data are collected for health care. However, because these data are collected for health care and not research they actually represent our observation and actions on the patient rather than the patient him- or herself. Data tend to be collected when patients are ill, for example. We therefore must transform the raw EHR data to a form that is useful for clinical research. One approach is called phenotyping [10,1], which maps the raw data to intermediate states like inferred clinical conditions that are then used in research. Phenotyping may be done manually as a set of rules or queries that assert a state based on raw data [10–14], or it may

[*] Corresponding author.
E-mail addresses: dja2119@cumc.columbia.edu (D.J. Albers), ne60@cumc.columbia.edu (N. Elhadad), jc1439@cumc.columbia.edu (J. Claassen), rip7002@nyp.org (R. Perotte), ag3304@cumc.columbia.edu (A. Goldstein), gh13@cumc.columbia.edu (G. Hripcsak).

https://doi.org/10.1016/j.jbi.2018.01.004
Received 14 August 2017; Received in revised form 5 December 2017; Accepted 14 January 2018

be automated using machine learning [15–19]. Continuous values like creatinine levels and glucose levels are measured longitudinally, usually at irregular, sparse intervals with a very wide variation among patients in number and spacing of measurements. Providing input to phenotyping algorithms is a challenge because each of the many laboratory and other continuous measurements can be seen as multi-dimensional (one dimension for each feature) with the number and timing varying among patients. Moreover, many machine learning techniques such as topic modeling only accept ordinal or categorical variables as input, usually focusing on note content and the presence of laboratory measurements. Laboratory data, are important to include in phenotyping because they contain relatively objective information. And while the mere presence of a test has a good deal of information, the addition of a quantification of the magnitude of the test is also important because the magnitude of many laboratory tests are the diagnostics used to define many diseases. A number of simple summarization techniques have been employed, such as using the presence, last value, the median, the mean, the standard deviation, or similar variations. These summaries assume that the important information in the measurements can be conveyed in one or two parameters (e.g., mean and standard deviation). The best summary may depend upon the variable, yet it is unclear how the summaries used in phenotyping are currently selected or what should be selected. For high-throughput phenotyping the selection of a summary technique would have to be automated given the number of potential variables and phenotypes.

Our ultimate goal is to develop an algorithm that can summarize the raw, continuous, inherently noisy, outlier-ridden, biased EHR data such that it emerges as a low-dimension summary that is free of biases, outliers, and other complexities, ready to be used by current machine learning techniques. Moreover, because the point is to help advance high-throughput phenotyping, we also address the problem of scalability. For example, when a problem related to a specific continuous variable is studied, the data from normal and diseased individuals can be studied, thresholds can be extracted from clinical guidelines, and physiologic understanding can used to devise a summary of the laboratory variable. When thousands of variables or diseases are studied at once, then a more automated approach is necessary. The problem is especially challenging when we consider that the variables may be non-Gaussian, that there may be subpopulations beyond the two primary ones—normal and diseased—and that groups of patients may be measured in different clinical contexts.

Our motivation for devising a method for automatically summarizing laboratory data to be used in computational tasks such as phenotyping evolved from four directions: (i) our work on health care process and phenotyping where we observed and documented how the health care influences, confounds, and highlights features that are observable from EHR data [4,1,20,2,21,5,22]; (ii) our Bayesian approach to estimating personalized, time dependent hazard functions that predict the onset of chronic kidney disease—the functions used to model and represent the data were chosen to be Weibull rather than the more standard Gaussian distributions because of the properties of EHR data [18]; (iii) our intuition that the processes generating health care data are relatively sparse [23] and may be summarized and modeled by large contributions from a few dominant features rather than a small contributions from all possible features; and (iv) our work translating phenotypic information to clinical settings where it became clear to us that more simple representations of data, e.g., via single, parameterized families, are more understandable and hence more useful for clinicians than black box prediction [24,25]. In essence, we wanted to find a way to minimize *garbage in* for machine learning methods, to translate laboratory data to a summary that was simple, faithful, interpretable all while minimizing the amount of human effort necessary to clean and summarize the data and therefore minimizing the resources needed to use EHR data in a high throughput setting.

While we followed the above path to this paper we are certainly not the first or only people using complex medical data, or complex data generally [26–29]; there are many other data preprocessing approaches and issues that we don't address here that are important to discuss, including data transformations, preprocessing using clinical knowledge or practice, temporal information, and the use of raw EHR data for phenotyping. Transforming data to a more convenient coordinate system or distribution is one common method used to make complex data easier to handle and more likely to produce more robust results. The Box-Cox transformation [30], which is a power transform [31], is an early method for transforming non-normal data to more normal data so that statistical analysis such as linear correlation can be done more reliably and with less bias. Similarly, general linear models [32,33] depend on transforming the response variables into a space that allows for a linear model to be estimated from diverse predictor variables. In the biomedical domain some researchers have devised more complex transformations of complex medical data to concepts such as anchors [15,16] that are likely to generalize across institutions. While it is common for authors to detrend the data in relatively standard ways [34,35,21], clinical knowledge is sometimes used to preprocess data in a relatively automated way. For example, some have used clinical patterns to discover nominal values [6], while others have worked to devise methods for finding normal ranges of laboratory data [36] and used that information to transform the data into a more practically useful format [37]. Similarly, clinical insight is sometimes used to adjust and transform ICU data in a laboratory-measurement-specific manner [38]. Sometimes data preprocessing is done in a particularly disease-specific way, e.g., [39,40]. Another approach is to standardize data format and quality, e.g., OHDSI [12] represents an effort to create world-wide and standardized health care data bases. These efforts address general data quality and standards but may not address health care process biases explicitly. Time is a crucial property of laboratory data. One issue is whether or not to include time at all. Most early EHR studies to not, and its inclusion depends largely on the questions be asked, the systems generating the data, and the data being used. Another issue is how to represent or parameterize time [41,19], a preprocessing choice that can have a significant impact on what results can be found [42]. But because all EHR data have missing values in time, an ever-present issue is how to incorporate time [43], a question often addressed by framing the data through the lens of missingness [44–47] or imputation and interpolation. For example, some authors use missingness of data as a feature [48,49,7] that can be used to define phenotypes. But more often researches focus on imputation schemes, or methods for interpolate missing values [50,51,21,52–54]. And finally, some phenotyping methods just use essentially raw, unaltered EHR data [55,19,56] with the assumption that the models are flexible enough to manage and model the data complexities automatically.

Together these results point to two high-level choices when preparing EHR data for phenotyping or related applications: use preprocessed or raw data; how and whether to use time in the analysis. In this paper we address the first choice. We do come down on the side of using preprocessed data—the method developed in this paper is a time agnostic method for summarizing laboratory data automatically based on EHR data, producing a numeric or categorical summary that can then be used in phenotyping or similar applications. Our method generates a laboratory variable summary that reveals useful information about the variable despite clinical subpopulations, varying contexts, and bias due to the health care process.

2. Methods and materials

2.1. Data sources

The study was carried out using two cohorts from different contexts. The first includes EHR data collected during a stay in a neurological intensive care unit (ICU) from patients who are comatose and tube-fed. The second cohort (AIM) comprises the entire longitudinal record of patients who visit regularly the Ambulatory Internal Medicine