YJBIN 2455
20 October 2015
ARTICLE IN PRESS
No. of Pages 8, Model 5G

Journal of Biomedical Informatics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Implications of non-stationarity on predictive modeling using EHRs

Kenneth Jung [a,*], Nigam H. Shah [b]

[a] Program in Biomedical Informatics, Stanford University, Stanford, CA, United States
[b] Center for Biomedical Informatics Research, Stanford University, Stanford, CA, United States

ARTICLE INFO

Article history:
Received 28 April 2015
Revised 8 October 2015
Accepted 12 October 2015
Available online xxxx

Keywords:
Data mining
Machine learning
Predictive model
Prognostic model
Wound healing

ABSTRACT

The rapidly increasing volume of clinical information captured in Electronic Health Records (EHRs) has led to the application of increasingly sophisticated models for purposes such as disease subtype discovery and predictive modeling. However, increasing adoption of EHRs implies that in the near future, much of the data available for such purposes will be from a time period during which both the practice of medicine and the clinical use of EHRs are in flux due to historic changes in both technology and incentives. In this work, we explore the implications of this phenomenon, called *non-stationarity*, on predictive modeling. We focus on the problem of predicting delayed wound healing using data available in the EHR during the first week of care in outpatient wound care centers, using a large dataset covering over 150,000 individual wounds and 59,958 patients seen over a period of four years. We manipulate the degree of non-stationarity seen by the model development process by changing the way data is split into training and test sets. We demonstrate that non-stationarity can lead to quite different conclusions regarding the relative merits of different models with respect to predictive power and calibration of their posterior probabilities. Under the non-stationarity exhibited in this dataset, the performance advantage of complex methods such as stacking relative to the best simple classifier disappears. Ignoring non-stationarity can thus lead to sub-optimal model selection in this task.

1. Introduction

The rapid adoption of Electronic Health Records (EHRs) is a key enabler of the learning healthcare system [1–5]. One effect of EHR adoption is to vastly increase the amount of data available for tasks such as predictive modeling of clinical outcomes. This increase in data enables developers of such models to employ increasingly sophisticated models to improve performance without overfitting. For example, recent work has applied tensor factorization to discover latent disease subtypes [6]. Such models are a far cry from the logistic regression models that have long been a mainstay of clinical research, and have the potential to transform clinical care. However, the observational nature of EHR derived data raises several practical issues in the development of such models [7–9]. EHR data may be incorrect and incomplete, and the majority of such data is collected primarily for billing purposes. Furthermore, some medical interventions could lower the risk of a particular outcome of interest and the popularity of these medical interventions can change over time as practices change. These factors can affect models that treat labels as unchanging truths. Failure to take these issues into account in the development and deployment of these models could lead to high profile failures that could ultimately delay the learning healthcare system [10,11].

In this paper, we note that EHRs typically have repeated observations of a constantly evolving set of patients. Furthermore, we note that the health care system in the United States is currently, and for the foreseeable future, in a state of flux, with new systems being adopted and clinical practice evolving at a rapid pace as incentives change. Indeed, we note that this situation is in fact an explicit goal of the learning health care system [1,5]. In the spirit of Walsh and Hripcsak [12], which examined the effect of data source, cohort selection and prediction target on the performance of a logistic regression model of hospital readmissions, we explore the effects these changes have on predictive modeling using EHR data.

We focus on the development of a predictive model for delayed wound healing using a dataset previously described in Jung et al. [13], which described the development of a predictive model for delayed wound healing and its potential clinical utility. The dataset consists of wound and patient data collected over the course of care at outpatient wound care centers operated by Healogics Inc. between 2009 and 2013. In this setting, patients are seen on a weekly basis to monitor the progress of wound healing and adjust

* Corresponding author. Tel.: +1 (415) 640 5289.
E-mail address: kjung@stanford.edu (K. Jung).

http://dx.doi.org/10.1016/j.jbi.2015.10.006
1532-0464/© 2015 Published by Elsevier Inc.

care as appropriate. Quantitative and categorical descriptions of wounds are entered into an EHR during each such assessment. The objective of the model is to predict whether or not a given wound will be an outlier with respect to how long it takes to fully heal, given only information collected during the first and second wound assessments. The threshold for delayed wound healing was set to fifteen weeks based on the observations of clinical experts at Healogics. Given accurate prognostic information, it is possible to triage patients for additional care such as additional monitoring and at-home care, hyperbaric oxygen therapy (HBOT), and negative pressure wound therapy (NPWT). Thus an accurate prediction has the potential to change the course of clinical treatment.

Non-stationarity is broadly defined as occurring when the data generating process being modeled changes over time. In this study, the data generating process is the routine care of wounds, as captured by patient and wound information recorded in the EHR. This information is presumed to be informative about delayed wound healing, and so we fit predictive models that use the EHR data to predict that outcome. Changes in the wound care process over time may render the evaluation and use of these models problematic because the joint distribution of covariates and delayed wound healing giving rise to the training data may not be the same as that giving rise to test data used to evaluate the models, and to future data.

Research into classification under non-stationarity has focused on two tasks – detecting non-stationarity (referred to as *anomaly detection* or *change detection*), and learning under non-stationarity. Moreno-Torres et al. [14] provides an overview of non-stationarity and related issues, while Hoens et al. [15] summarizes current methods for dealing with non-stationarity. In brief, these methods modify the dataset or the model in response to new data such that more weight is given to the most recent data. However, these methods have for the most part been used only in domains such as online fraud detection; to the best of our knowledge, they have never been applied to clinical risk prediction. In this study, we aim to characterize the implications of non-stationarity on the development of predictive models of the sort commonly encountered in clinical informatics.

To that end, we present experiments evaluating the impact of non-stationarity on discriminative power (how well models distinguish between cases and non-cases) and on model calibration (how closely the posterior probabilities of delayed wound healing output by models match observed frequencies of delayed wound healing). We approximate different degrees of stability of the data generating process by changing the way that the data is split into training and test sets. We then examine how such change impacts model selection. To that end, we consider the use of increasingly sophisticated models, starting from regularized logistic regression, progressing through non-linear models capable of automatically modeling interactions between predictors, and ending with ensemble methods that combine the predictions of many base models. Finally, we examine the impact of non-stationarity on engineered, domain specific features.

We demonstrate that in a setting that approximates a stationary data distribution, methods such as stacking can provide significant boosts to predictive power relative to the best base models. However, this performance gain disappears when the data distribution is non-stationary. In both cases, however, there is consistent benefit from using engineered, domain specific features. We find that using non-linear models that capture feature interactions automatically is useful in this dataset but that the benefit from such models is reduced under non-stationarity. Our findings emphasize the importance of matching the model development process with the intended use of the model. If the model is intended for use on future patients, it is critical to take non-stationarity into account to obtain a reliable estimate of model performance.

2. Materials and methods

Our goal is to investigate the impact of non-stationarity on a predictive model for delayed wound healing, defined in this study as whether or not a given wound will take longer than 15 weeks to heal using information routinely collected during the first week of care. We approach this by fitting a series of increasingly complex models—with and without domain specific features—to different training and test splits of the data. We observed that the dataset exhibits substantial non-stationarity. We can, however, control the degree of non-stationarity seen by the models by changing the way we split the data. This process is summarized in Fig. 1 and explained further in Section 2.2. We evaluate the models for discriminative power and calibration under these different conditions. In the remainder of this section, we provide details about the dataset, feature construction, model development and evaluation.

2.1. Dataset

The dataset is comprised of 1,182,751 time-stamped wound assessments performed at 68 Healogics outpatient wound care centers distributed over 26 states. These wound assessments represented 180,716 unique wounds. Each wound assessment consists of both quantitative information regarding a specific wound, such as length, width, depth and area, in addition to categorical descriptors such as wound type, anatomical location, presence/absence of erythema and ICD9 codes associated with the assessment. Each assessment is also associated with unique wound and patient keys, allowing us to associate each wound with basic demographic information such as age, sex, and insurance status along with its outcome. Wound assessments were performed approximately weekly, and the dataset spans 2009 through 2013. A total of 59,958 patients are represented, and there are no restrictions on patients or wound types. Supplementary Materials Table 1 provides additional demographic details about the dataset, broken down by wound center.

We removed any wounds that were unresolved by the end of the study period unless the wound was already past the 15-week threshold for delayed healing. We also removed wounds with negative or very large values for quantitative features (>99.9th percentile) or with clearly erroneous demographic information such as negative age. This left us with 150, 277 unique wounds for use in training and testing our models. The basic features for our models are the data for each wound that is available at the time of the first wound assessment.

We performed additional pre-processing of the dataset as follows. First, ICD9 codes were aggregated to 3 digit codes. Second, wound types and locations were collapsed into 40 and 37 values from 103 and 216 values, respectively, in order to account for variation in how these variables were recorded in different wound care centers and to aggregate values that were judged to be clinically equivalent (upon manual review by kJ) for the purposes of the predictive model. For instance, the locations ‘Arm – Elbow’ and ‘Elbow’ were both mapped to the single location ‘Elbow’, and ‘Foot – 2nd Toe’ and ‘Foot – 3rd Toe’ were collapsed to ‘Toe’. Third, insurance information was collapsed into four categories – uninsured, private, Medicaid, and Medicare.

In this study, delayed wound healing is defined as taking 15 or more weeks to heal; this threshold was chosen based on the advice of domain experts from Healogics. 11.7% of wounds met this criterion in the final dataset.