Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

The effects of data sources, cohort selection, and outcome definition on a predictive model of risk of thirty-day hospital readmissions

Colin Walsh [a,b,*], George Hripcsak [a]

[a] Department of Biomedical Informatics, Columbia University, United States
[b] Department of Medicine, Columbia University, United States

ARTICLE INFO

Article history:
Received 15 February 2014
Accepted 14 August 2014
Available online xxxx

Keywords:
Readmissions
Predictive analytics
Electronic health record
Regularized Logistic Regression
Text mining
Risk modeling

ABSTRACT

Background: Hospital readmission risk prediction remains a motivated area of investigation and operations in light of the hospital readmissions reduction program through CMS. Multiple models of risk have been reported with variable discriminatory performances, and it remains unclear how design factors affect performance.
Objectives: To study the effects of varying three factors of model development in the prediction of risk based on health record data: (1) reason for readmission (primary readmission diagnosis); (2) available data and data types (e.g. visit history, laboratory results, etc); (3) cohort selection.
Methods: Regularized regression (LASSO) to generate predictions of readmissions risk using prevalence sampling. Support Vector Machine (SVM) used for comparison in cohort selection testing. Calibration by model refitting to outcome prevalence.
Results: Predicting readmission risk across multiple reasons for readmission resulted in ROC areas ranging from 0.92 for readmission for congestive heart failure to 0.71 for syncope and 0.68 for all-cause readmission. Visit history and laboratory tests contributed the most predictive value; contributions varied by readmission diagnosis. Cohort definition affected performance for both parametric and nonparametric algorithms. Compared to all patients, limiting the cohort to patients whose index admission and readmission diagnoses matched resulted in a decrease in average ROC from 0.78 to 0.55 (difference in ROC 0.23, p value 0.01). Calibration plots demonstrate good calibration with low mean squared error.
Conclusion: Targeting reason for readmission in risk prediction impacted discriminatory performance. In general, laboratory data and visit history data contributed the most to prediction; data source contributions varied by reason for readmission. Cohort selection had a large impact on model performance, and these results demonstrate the difficulty of comparing results across different studies of predictive risk modeling.

1. Introduction

Clinical, legislative, and financial drivers have elevated the significance of hospital readmissions for the multidisciplinary care team and hospital administrators. The emphasis on readmissions as a reportable quality measure and as a source of potential reimbursement penalty through the Centers for Medicare and Medicaid Services (CMS) has been well-described [1]. Consensus is forming to support the need for patient-centered interventions across care settings to prevent readmissions for particular patients [2,3]. The first step in the myriad of efforts to reduce readmissions remains identification of patients at high risk [3].

The most comprehensive review of readmissions risk prediction models to date was published in 2011 by Kansagara et al. [4]. Since then, thousands of new articles on the topic have been published. A simple OVID Medline search for “Patient readmission” in 2011 produced 5476 hits [4], while it yields 7576 results at the start of 2014. Each model has the potential to be adapted by researchers and managers in new clinical settings, but to do so appropriately, it is critical to understand the sensitivity of such models to varying the way in which they are built and deployed. While researchers

Abbreviations: D, Demographics; L, Laboratory Tests; I, Prior ICD9 Codes; S, Social/Mental Health Keywords; O, Other Clinical Keywords; V, Visit Utilization History.

* Corresponding author at: Department of Biomedical Informatics, Columbia University, 622 W 168th St, VC5, New York, NY 10032, United States. Fax: +1 212 305 3302.

E-mail address: cgw2106@columbia.edu (C. Walsh).

http://dx.doi.org/10.1016/j.jbi.2014.08.006
1532-0464/© 2014 Elsevier Inc. All rights reserved.

also must compare results across seemingly similar studies, it is poorly understood how different factors in model design affect performance. Thus, it remains unclear if comparisons are legitimate as studies may differ in a number of different aspects.

The goal of this study is to study the effect of three factors on prediction of hospital readmission risk. The first factor is the reason for readmission as defined by the primary readmission diagnosis. Early predictive models of readmissions focused on all-cause readmission and the most common diagnoses including congestive heart failure (CHF), acute myocardial infarction (acute MI), and chronic obstructive pulmonary disease (COPD), but the literature now spans multiple diagnoses and disciplines [5–15]. However, no studies have studied systematically the effects of changing readmission diagnoses being modeled while holding all else equal. This latter understanding will help interpret and compare studies of different diseases. Additionally, the ability to predict readmission as a simultaneous panel of cases may have clinical utility in that it may direct clinical interventions to causes deemed most likely for a particular patient by the predictive algorithm.

The second factor under study is data availability. Studies have included data types such as administrative and claims data, test results and clinical text [4,16–19]. One study demonstrated that readmission rates and rates of unnecessary readmissions vary by method of chart review to tally readmissions and by altering the breadth of the definition of a readmission itself [19]. This work studies the effect of varying the features in the model across multiple readmission diagnoses holding all else unchanged. We attempt to elucidate the contributions of data types included for prediction in clinically meaningful bins: laboratory tests, visit utilization, demographics, clinical narrative. While it is clear that more data and more clinically deep data should be better, it remains unclear to what extent the selection of data type is dependent on how the problem is cast.

The third factor is the cohort that is selected for study. The challenge of generalizability to new cohorts is well known; in considering external validity of predictive models, cohort selection can impact discrimination and calibration [20,21]. Prediction models generally take two forms: prediction of readmission for preselected cohorts such as known patients with chronic obstructive pulmonary disease, Medicare patients only, or those undergoing abdominal surgery [5,16,22–26]; or prediction of readmission for all patients to an institution or set of institutions. We hypothesize that this choice of cohort definition is a crucial one – that with the same input clinical data, the same prediction goal, and the same underlying population from which the cohort is selected, the criteria used to select the cohort can have large effects on the performance. This effect has not been quantified in the domain of readmissions risk to our knowledge, and there are implications to those seeking to use reported models in clinical practice. This research question has an important corollary implication: if performance is highly dependent on how the cohort is selected despite everything else being the same, then it demonstrates that comparing performance across studies must be difficult.

2. Materials and methods

2.1. Dataset

A retrospective cohort of inpatient admissions at Columbia University Medical Center (CUMC) in New York City was identified from 2005 to 2009. These years were selected as the clinical data repository at the institution is replete with clinical and administrative data over this time period and because clinical workflows with respect to electronic health record data structures were fairly static over this time. One exception is an increase in adoption of electronic documentation over the study time period. 263,859 inpatient admissions were collected. Admissions for patients aged less than 18 years were excluded. Admissions within 30 days for ICD9 650.xx, “Normal delivery”, were also excluded as were admissions to the physical medicine and rehabilitation service, which are logged as separate admissions but represent planned transfers of care.

For each unique patient identifier, a single admission was selected randomly as the index admission. The study dataset comprised this index admission, data from previous admissions or other encounters within the past year, and data for any readmission within 30 days of discharge. When necessary for admissions in 2005, visit and diagnosis data from the preceding year were collected. Similarly, follow-up data regarding readmissions were collected when necessary for admissions in December 2009. Diagnostic, laboratory, and documentation data were accessed from the clinical data repository and preprocessed in Python in preparation for importing into the open-source language for statistical computing, R [27]. Characteristics of the training dataset and readmission prevalence stratified by readmission diagnosis are described in Tables 1 and 2.

2.2. Initial feature selection

Relevant features were selected in two phases. Initially, domain expert criteria were used to choose variables based on clinical importance. Then these preselected variables were used to create

Table 1
Demographics and utilization history characteristics of training dataset (2005–2008).

Training data characteristics (total number of patients = 92,530)	Number of patients	Percentage of total number of patients
Age		
18–45	26,239	28.4
45–65	32,144	34.7
>65	34,147	36.9
Sex		
Male	43,964	47.5
Female	48,566	52.5
Insurance status		
Medicaid	12,152	13.1
Medicare	12,477	13.5
Admission service type		
Internal medicine	45,697	49.4
Surgery	13,887	15.0
Psychiatry	5391	5.8
Neurology	4380	4.7
Other	23,175	25.0
Discharge status		
To home	72,749	78.6
To skilled nursing facility	5950	6.4
With home care services	5507	6.0
Other	8324	9.0
Utilization statistics		
Number of ER visits in year preceding index admission		
0	69,778	75.4
1–4	20,861	22.5
>5	1891	2.0
Number of inpatient visits in year preceding index admission		
0	77,999	84.3
1–4	13,981	15.1
>5	550	0.6
Number of outpatient visits in year preceding index admission		
0	57,592	62.2
1–4	19,629	21.2
5–10	7,559	8.2
>10	7,750	8.4

Please cite this article in press as: Walsh C, Hripcsak G. The effects of data sources, cohort selection, and outcome definition on a predictive model of risk of thirty-day hospital readmissions. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.08.006