Predicting 30-day Hospital Readmission with Publicly Available Administrative Database*

A Conditional Logistic Regression Modeling Approach

K. Zhu[1]; Z. Lou[2]; J. Zhou[3]; N. Ballester[4]; N. Kong[5]; P. Parikh[4]

[1]Purdue University, Statistics, West Lafayette, Indiana, USA;
[2]Purdue University, Industrial Engineering, West Lafayette, Indiana, USA;
[3]Tsinghua University, Automation, Beijing, China;
[4]Wright State University, Biomedical, Industrial and Human Factors Engineering, Dayton, Ohio, USA;
[5]Purdue University, Biomedical Engineering, West Lafayette, Indiana, USA

Keywords
Hospital readmission, risk assessment, binary classification, conditional logistic regression

Summary

Introduction: This article is part of the Focus Theme of *Methods of Information in Medicine* on "Big Data and Analytics in Healthcare".

Background: Hospital readmissions raise healthcare costs and cause significant distress to providers and patients. It is, therefore, of great interest to healthcare organizations to predict what patients are at risk to be readmitted to their hospitals. However, current logistic regression based risk prediction models have limited prediction power when applied to hospital administrative data. Meanwhile, although decision trees and random forests have been applied, they tend to be too complex to understand among the hospital practitioners.

Objectives: Explore the use of conditional logistic regression to increase the prediction accuracy.

Methods: We analyzed an HCUP statewide inpatient discharge record dataset, which includes patient demographics, clinical and care utilization data from California. We extracted records of heart failure Medicare beneficiaries who had inpatient experience during an 11-month period. We corrected the data imbalance issue with under-sampling. In our study, we first applied standard logistic regression and decision tree to obtain influential variables and derive practically meaning decision rules. We then stratified the original data set accordingly and applied logistic regression on each data stratum. We further explored the effect of interacting variables in the logistic regression modeling. We conducted cross validation to assess the overall prediction performance of conditional logistic regression (CLR) and compared it with standard classification models.

Results: The developed CLR models outperformed several standard classification models (e.g., straightforward logistic regression, stepwise logistic regression, random forest, support vector machine). For example, the best CLR model improved the classification accuracy by nearly 20% over the straightforward logistic regression model. Furthermore, the developed CLR models tend to achieve better sensitivity of more than 10% over the standard classification models, which can be translated to correct labeling of additional 400–500 readmissions for heart failure patients in the state of California over a year. Lastly, several key predictor identified from the HCUP data include the disposition location from discharge, the number of chronic conditions, and the number of acute procedures.

Conclusions: It would be beneficial to apply simple decision rules obtained from the decision tree in an ad-hoc manner to guide the cohort stratification. It could be potentially beneficial to explore the effect of pairwise interactions between influential predictors when building the logistic regression models for different data strata. Judicious use of the ad-hoc CLR models developed offers insights into future development of prediction models for hospital readmissions, which can lead to better intuition in identifying high-risk patients and developing effective post-discharge care strategies. Lastly, this paper is expected to raise the awareness of collecting data on additional markers and developing necessary database infrastructure for larger-scale exploratory studies on readmission risk prediction.

Correspondence to:
Nan Kong
206 S. Martin Jischke Dr.
West Lafayette, IN 47907
USA
E-mail: nkong@purdue.edu

Methods Inf Med 2015; 54: 560–567
http://dx.doi.org/10.3414/ME14-02-0017
received: October 2, 2014
accepted: September 16, 2015
epub ahead of print: November 9, 2015

* Supplementary online material published on our website
http://dx.doi.org/10.3414/ME14-02-0017

1. Introduction

Patients in the United States are frequently readmitted to acute care hospitals with a short duration after hospital discharge [1]. Hospital readmissions cause considerable unnecessary cost to the US healthcare system. It is estimated that preventable readmissions for Medicare patients alone cost $17 billion annually [2], equivalent to more than 10% of Medicare benefit payment for hospital inpatient services [3]. Readmission reduction has been deemed critical to the U.S. public funding agencies (i.e., Medicare and Medicaid), whose spending increases rapidly in recent years with aging population and increased prevalence of chronic conditions. Furthermore, hospital readmissions present significant and unnecessary burden to care resource utilization. Regarding readmission as an important indicator of poor care quality and efficiency [4], hospitals in the U.S. have a strong impetus to reduce their readmission incidences. The U.S. Center for Medicaid and Medicare Services (CMS) provides reputational pressure to hospitals to reduce preventable readmissions [5]. In 2009, CMS began to publish 30-day risk-standardized readmission rates for heart failure, acute myocardial infarction, and pneumonia [6–8]. Recently, as part of Affordable Care Act (ACA), CMS started providing financial incentives to hospitals to lower readmission rates. Although it is evident that many readmissions are preventable with effective post-discharge care strategies (e.g., [9, 10]), there is no clear consensus on important risk factors that cause readmissions and thus no clear consensus on which discharged patients to be focused on for preventing their readmissions and by what discharge management strategy. There has been a strong need for developing accurate statistical models for predicting 30-day hospital readmissions. Good prediction models can help 1) identify high-risk patients (e.g. [11, 12]) and 2) develop an effective plan for post-discharge care (e.g. [9]).

Risk factors on hospital readmission can basically be divided into three main categories. The first category contains factors related to patient demographics and socioeconomic status. The second category is related to health care utilization, e.g., length of hospitalization, discharge population, type of insurance, etc. The third category contains mainly clinical information obtained through treatment phases of care, e.g., acute illness diagnosis, severity, comorbidity, surgery, and so on.

In most academic research, retrospective studies are conducted with hospital administrative data, including mainly above factors but with various levels of exclusion given their availability. Using administrative data helps develop population-level discharge management guidelines (e.g. [2, 13]). Moreover, this type of research can offer recommendations to individual hospitals in terms of their emergency and inpatient electronic health record acquisition. In this paper, we follow this trend of using administrative data. Nevertheless, like many other data scientists, we have only access to publicly available administrative databases. In our case, the database is a state-wide inpatient database from the Healthcare Cost and Utilization Project (HCUP), the most comprehensive source of hospital administrative data in the U.S.

As for research methodologies used in the readmission risk profiling research, most of the studies use descriptive, particularly discriminatory, analysis to decipher the casual relations between one or few selected risk factors and the readmission incidence of certain disease or disease class. For example, demographics based risk factors identified in previous studies include age [14–17], sex [18], income [19], and level of education [20]; health utilization related risk factors identified include length of hospitalization [16], frequency of hospitalization and use of emergency room within six months prior to the index hospitalization [20], frequency of hospitalization within one year from the index hospitalization [15], type of insurance [21], and type of ward [22]; and treatment and clinical factors identified include department of treatment [14, 16], specific type of comorbidity [23], number of comorbidities [18], surgery [24], clinical test results [20], change in the amount of drug dose within 48 hours of discharge [16], experience of depression [17], and state of mental health [20].

The rest of the studies in the literature are exploratory studies that identify influential risk factors from a large number of candidate predictors. In a systematic review, Kansagara et al. [11] reported 26 unique readmission prediction studies, which were developed up to year 2011. Fourteen such studies relied on retrospective administrative data. The most common outcome among these studies was 30-day readmission incidence. Among these studies, three [25–27] were conducted based on a large cohort from the U.S. Two other studies [28, 29] were conducted at multiple centers in a single state.

The systematic review in [11] reported that most current readmission risk prediction models have limited discriminative ability (i.e., the c-statistic ranges from 0.55 to 0.8 with lower values typically coming from using administrative data). There are two potential reasons for this limitation. First, while large administrative databases have been shown to contain noticeable noise, they are still the primary sources in the prediction model development. This is because detailed clinical data at the inpatient stage (e.g., daily vitals) and social/behavioral data at the post-discharge stage (e.g., whether to have informal care giver) are hard to obtain, given the current state of medical data integration and management. Additionally, administrative databases are easier and less expensive to obtain by researchers. Second, to our knowledge, almost all prediction models on 30-day readmission published in the literature (as cited earlier in this paper) are logistic regression models. Only recently, alternative statistical machine learning methods have been investigated for predicting 30-day readmission incidence. For example, Natale et al. [30] developed a decision tree model and compared it with standard logistic regression models. Lee [31] compared three methods, logistic regression, decision tree, and neural networks. Hosseinzadeh et al. [32] compared a decision tree classifier and a Naïve Bayes classifier. For other systematic reviews on the prediction modeling of readmission incidence, we refer to Desai et al. [33] and van Walraven et al. [34]. In summary, the lack of additional observations on potentially influential risk factors and use of accurate classifiers, together, has resulted in less-

than-satisfied performance on readmission risk classification/prediction.

In this research, we focus on the latter limitation and provide a comprehensive investigation of viable classifier options based on heart failure patients' health utilization data from a publicly available state-wide inpatient database. We first attempt to improve the accuracy of readmission risk prediction with application of general-use classification methods, such as random forests and support vector machines, over the entire data set. After realizing only little improvement can be made, we turn our attention to applying conditional logistic regression with comprehensive ad-hoc selection of stratification variables and rules *a priori*. Our selection approach suggests several variables with good discriminatory power. We subsequently use them to stratify the initial data set into more homogenous strata and constructed a logistic regression model for each stratum. To improve the prediction accuracy, we also explore the effect of interactions among the variables.

Coupled with careful construction of the strata, the use of conditional logistic regression offers several easily attainable predictors and derives understandable decision rules for predicting 30-day readmission risks for U.S. heart failure patients. Further, a second benefit is that this method improves upon the prediction compared to a single classifier trained on the entire dataset. With judicious use, our method can be generalized to other cohorts given that most analyzed prediction variables in this study are commonly collected when managing different acute diseases in real practice. Overall, we contribute to the growing literature of readmission risk prediction by exploring the use of conditional logistic regression and ad-hoc stratum construction. We expect our work to lead to better intuition in identifying high-risk patients and developing effective post-discharge care strategies as opposed to more standard "black-box" type classification methods.

2. Research Methodology

For our study, we examined the State Inpatient Database (SID) from the state of California. SIDs are state-wide datasets collected from participating providers throughout the U.S. These datasets contain data about nearly 90% of all U.S. hospital inpatient discharges. SIDs include a set of patient data (e.g., patient's age, gender, race, payer status), as well as information related to initial diagnosis of acute condition (e.g., ICD-9 codes) and post discharge (e.g., discharge date, readmission date, and disposition location). These patient data provide the basis of specifying readmission outcomes and offer a large set of predictors to choose from.

SIDs are part of the Healthcare Cost and Utilization Project (HCUP). HCUP, sponsored by the Agency for Healthcare Research and Quality (AHRQ), is the largest collection of nationwide and state-specific hospital administrative data in the U.S. AHRQ/HCUP databases contain encounter-level, both clinical and nonclinical information, beginning in 1988. These databases enable research on a broad range of health policy issues, including cost and quality of health services, medical practice patterns, access to health care programs, and outcomes of treatments at the national, state, and local market levels. For more information on HCUP and SID, we refer to the AHRQ webpage of HCUP (www.ahrq.gov/research/data/hcup/).

With only two labels (readmitted or not readmitted), the problem of predicting readmission incidence within 30 days after acute hospitalization falls into the category of binary classification. First, we constructed the dataset to be studied with several levels of data extraction from the original California SID. We then explored the use of various classification/regression methods, including logistic regression and random forests. Later we developed ad-hoc conditional logistic regression models and performed cross validation to compare them with alternative models explored earlier.

2.1 Data Preparation

The data preparation step comprised of data extraction and imbalance correction. We started the preparation with the data set that contains all inpatient records collected in year 2010 from the state of California. The original data set contained 3,970,921 records. We selected the cohort by several main criteria as follows: 1) age group was 65 and older; 2) primary payer was Medicare; 3) primary residence was in California; 4) primary diagnosis was heart failure (HF), as identified by validated International Classification of Disease, Ninth Revision, diagnosis codes, i.e., ICD-9 code (►Appendix A); 5) discharge occurred between January and November, 2010; 6) did not immediately transfer out after index hospitalization and hospital readmission; and 7) discharged to home self-care, home health care, or nursing home care. In addition, we removed records with missing, errand information, or very low frequency. We illustrate the entire data extraction and cleaning procedure in ►Appendix B.

Loosely speaking, we extracted relevant SID records associated with the California Medicare HF patients who were discharged from January to November of 2010. We considered only the prediction of first readmission incidence within 30 days. Note we would not be able to record 30-day readmission of patients discharged in December 2010 since we did not have data from January 2011 on. There are three main reasons of choosing HF patients. First, there seems to be noticeable differences in 30-day readmission incidence among major disease groups such as heart failure, acute myocardial infarction, pneumonia. Second, the volume of HF patients in California is high, thus helping us ensure model validity. Second, in real-world readmission reduction, HF is one of the diseases that care organizations pay attention to given its relatively high readmission risk. In regard to the sixth criterion above, it is not meaningful to consider transfer to another hospital with a very short stay, which likely indicates the studied hospital is unable to completely handle the patient's acute condition. In regard to the seventh criterion, very few records have one of the other discharge options. Thus including those records would not ensure the records to be evenly divided in terms of readmission. Similarly, we did not include records of Native American patients and patients who could not be identified whether Hispanic or not. Note that almost all patients who fell in either of the two categories

would be readmitted within 30 days. After performing all the data extraction items as above, the records of 22,410 patients remained. Among them, 17,434 were not readmitted within 30 days after discharge, which formed the majority class; the overall 30-day readmission rate (4976 cases) for the analyzed cohort was 22.2%.

We selected a comprehensive set of 36 prediction variables for the model development, which are listed in ▶Appendix C. In addition to the commonly used predictors such as the numbers of chronic conditions and acute care procedures, the variable set included 22 binary-valued AHRQ comorbidity measures, e.g., CM_LIVER – indicator of liver disease. A few other AHRQ comorbidity measures, e.g., CM_ULCER – indicator of peptic ulcer disease excluding bleeding, were not included since they would not ensure balanced dichotomy in terms of readmission. ▶Appendix D reports the cohort characteristics with respect to most of the variables selected. Variations in categorical and continuous variables are expressed with frequencies and mean ± SD, respectively.

In our study, the analyzed data were highly unbalanced so classifiers developed without properly balancing the data tend to ignore classification errors of positive responses. To correct this problem, we tested the use of three common remedies: under-sampling, over-sampling, and different error cost (DEC). Under-sampling discards observations belonging to the majority class so that the resultant sampling leads to equal number of observations for both classes [35]. Over-sampling can be considered the dual of under-sampling. It generates additional artificial observations so that the resultant sampling also leads to equal number of observations for both classes [36]. The DEC approach assigns different costs to errors due to false positive and false negative [37]. Because our data include many more non-admits than admits, we assigned a higher cost to negative.

Our study suggested that under-sampling worked best in our setting because it is computationally more efficient than the other two techniques and at least as good as the other techniques in terms of classification accuracy. It also avoided the difficulties of creating artificial observations for over-sampling or choosing error rates for the DEC approach. In our implementation of under-sampling, to ensure the total sample size, we kept all positive responses (i.e., readmitted) and matched the number of negatives response accordingly. In the process of selecting the samples with negative response, we repeatedly checked the representative validity in key prediction variables such as gender, number of chronic condition, among others, i.e., the sampled distributions were roughly identical to the distributions from the original data set with negative response. After correcting the inherent data imbalance with under-sampling, we obtained a total of 9952 cases in which half with positive response and half with negative response. We present evidence of balanced sample data in ▶Appendix E.

2.2 Ad-hoc Conditional Logistic Regression Modeling

In our preliminary experiments, we first employed the standard logistic regression with the use of R package *glm*. We supplemented the logistic regression further with stepwise variable selection. For this, we used the R package *glm* with the specification of argument "step". Meanwhile, for the classification, we constructed random forest with the use of R package randomForest and support vector machine (SVM) with the use of R package *e1071*. Unfortunately, the preliminary experiments did not show much promise in readmission risk prediction. Moreover, the complexity of random forest and SVM would not easily convert the embedded decision rules – nearly "black box" – to practically meaningful analytic intelligence.

We conjectured that poor performance in the prediction may be attributed to profound heterogeneity in the patient population. Subsequently, we explored the possibility of classifying the entire patient data into several patient subpopulations with a view that each of them would be less heterogeneous and may enable classifiers to better predict readmissions. We did this sequentially by first developing a decision tree with the entire data and trimming it when it was two levels deep. This gave us the top 2–3 influential variables and the dichotomization values on them, which would be used to determine the subpopulations. A logistic regression was then developed for each of these subpopulations. Essentially, we combined the advantages of two types of approaches (i.e., regression and decision tree) with ad-hoc stratification variable selection for improved chance of knowledge translation.

By reviewing the first layer of the decision trees constructed, we observed high appearance of three prediction variables. They are DISPUniform (i.e., disposition location after discharge), NPR (i.e., number of ICD-9-CM procedures coded on the discharge record), and NCHRONIC (i.e., number of chronic conditions). Furthermore, the decision trees specified the threshold value for the classification (i.e., dichotomy) on each variable, which also tended to remain consistent. For DISPUniform, one stratum corresponded to patients who were transferred back home (i.e., DISPUniform = 1) and who were transferred to home health care facilities (i.e., DISPUniform = 6). The other stratum corresponded to patients who were discharged to self-care at home (i.e., DISPUniform = 5). For NPR, the split occurred between the value being less or equal to 4 and above 4. For NCHRONIC, the split occurred between the value being less or equal to 7 and above 7. We also noted that the above three prediction variables showed superior discriminatory ability in the logistic regression (▶Appendix F). We thus applied conditional logistic regression through stratification of the dataset into subsets based on each of the three prediction variables identified above. With each variable, we constructed two disjoint data strata whose union is identical to the original data set. We then fitted each data stratum with a logistic regression model. Note that a few other prediction variables also yielded high discriminatory ability through straightforward logistic regression, e.g., Do Not Resuscitate. But they did not appear on the top of the decision trees. Some of these variables are indicators of comorbid conditions, and in fact, some recent studies set focus on differentiating the readmission risks with respect to them (e.g. [23]). We plan to consider these variables for the stratification and subsequently conditional regression analysis in future study.

Through preliminary experiments on the above conditional logistic regression, we noticed modest improvement on the prediction accuracy. We hypothesized that sufficient stratification of the patient cohort into homogeneous cohorts was still not achieved. We thus continued our exploratory stratification selection by considering the top two layers in the developed decision tree (with the use of R function "ctree"). After the first node (or layer), the decision tree was branched on variable DISPUniform between 1, 5, and 6. Then at one node on the second layer, the decision tree was further branched on NPR between its value ≤ 4 and > 4 whereas at the other node, it was further branched on NCHRONIC between its value ≤ 7 and > 7. As a result, we constructed four data mutually exclusive data strata. Note that the left branching and right branching at the second layer are typically not on the same prediction variable (▶Appendix G).

Also from the previous exploration, we concluded that stepwise variable selection would not lead to improved classification performance. We conjectured that the poor fitting issue was rooted at missing of higher-order modeling. Thus within each of the four constructed strata, we deployed logistic regression over an enlarged variable set with additional variables capturing the pairwise interactions between original variables. The criterion for selecting the original variables for the above purpose was that the original variables were shown to be influential via the standard logistic regression (p-value < 0.05).

In summary, we developed the following classification models: 1) standard logistic regression (whole dataset); 2) stepwise logistic regression (whole dataset); 3) random forest (whole dataset); 4) support vector machine (whole dataset); 5) conditional logistic regression (with two strata based on each of the top 3 variables); 6) conditional logistic regression (with four strata derived from a decision-tree based rule set); and 7) conditional logistic regression (embellishment of (6) with incorporation of additional pairwise interacting variables). For convenience, we call them LR, SLR, RF, SVM, CLR1, CLR2, CLR3 in the remainder of the paper. For RF, SVM, CLR1, we also considered performing them on two different prediction variable sets: all original variables and selected ones through conditional logistic regression. To compare the above classification models, we performed cross validation and assessed each model's classification accuracy. We describe this comparative cross-validation study in the following section.

2.3 Classification Model Evaluation and Comparison

To evaluate the developed classifiers, we performed cross-validation, for which we split the original data set into two subsets, i.e., training set (70% of the original data) and test set (30%). To ensure the randomness, when we created the split between the training set and the test set, we checked whether the mean response value from either set of the split was roughly equal to the mean response value from the original dataset and whether the distributions of several key prediction variables, such as ASOURCE, DISPUniform, NPR, NCHRONIC, DNR, are similar between the two subsets of each split. We performed the splitting repeatedly until satisfactory results were obtained from the above comparisons. We present evidence on the satisfactory splitting in ▶Appendix H.

For CLR2, we report the influential variables in each data stratum, where they were identified through a logistic regression over all original variables (▶Table 1). With more than one influential variable identified, we could pair them to create additional interacting variables. For example, in Stratum 1, variables CM_COAG, CM_LYTE, DNR, and FEMALE appeared to be influential. We paired them to create six additional interacting variables such as DNR_FEMALE. Note that in Stratum 2 (i.e., DISPUniform = 5 and NPR > 4), only one category of RACE presented significant discrimination on the prediction. So we did not include any additional interacting variables when using CLR3 for that stratum. We also report the number of records in each stratum and reiterate the decision rules.

With CLR3, we included interacting variables based on the influential variables identified in CLR2. We performed the standard logistic regression and noticed the potential overfitting issue in Strata 1 and 4 as none of the interacting variables appeared to be significant with logistic regression. We thus decided to keep the logistic regression models from the two strata in CLR2. On the other hand, the re-

Table 1 Influential variables identified through logistic regression

	Decision Rules	Total # of Subjects (# of admitted, # of not admitted)	Influential Variables identified for interactions from CLR2 (p ≤ 0.05)	
			Variable Name	*p*-value
Stratum 1	DISPUniform = 5; NPR ≤ 4	1660 (1039, 621)	CM_COAG	0.04674
			CM_LYTE	0.02462
			DNR	0.00139
			FEMALE	0.01807
Stratum 2	DISPUniform = 5; NPR > 4	646 (462, 184)	RACE3[a]	0.0455
Stratum 3	DISPUniform = 1 or 6; NCHRONIC ≤ 7	3278 (1309, 1969)	CM_LYTE	0.0245
			NPR	0.0173
			RACE2[b]	0.0251
Stratum 4	DISUniform = 1 or 6; NCHRONIC > 7	4368 (2166, 2202)	CM_METS	0.00046
			CM_TUMOR	0.0207
			CM_WGTLOSS	0.0185
			DNR	0.0348
			MEDINCSTQ3[c]	0.0458

[a] Hispanic; [b] black; [c] third-quartile of median household income for the state

sults in Stratum 3 seemed to be encouraging with improved classification accuracy. We thus only included the 3 interacting variables for Stratum 3 in CLR3.

To ensure fair comparison among the classification models, we conducted parameter tuning, which included identifying optimal thresholds for logistic regression modeling and determining the optimal tree size for the random forest. First, for each regression model, we varied the threshold from 0 to 1 to identify the optimal one that yielded the smallest classification error based on the training set. We then applied this optimal threshold to the test set for assessing the predicted classification error. For conditional logistic regression models, since the data strata would not normally have equal numbers of positive and negative responses, we identified the optimal threshold value for each data stratum. In addition, we conducted parameter tuning for the random forest classification method. We varied the parameter specifying the number of trees in the random forest from 250 to 500 and identified an optimal number based on the training set. We then applied this number to the test set for assessing the predicted classification error. We present the parameter tuning results in ▶Appendix I.

In ▶Table 2, we report the comparative study results. Our cross-validation results suggested that conditional logistic regression made modest improvement in classification accuracy over more straightforward classification methods. This justified the need of carefully investigating the use of conditional logistic regression. Our results also suggested that among different ideas on conditional logistic regression, CLR2 made modest improvement in classification accuracy over CLR1 and CLR3 further made slight improvement in classification accuracy over CLR2. This suggested that it is beneficial to judiciously explore the use of decision tree modeling to guide the cohort stratification and it is possibly beneficial to investigate the inclusion of interacting variables to improve the prediction accuracy as well. Furthermore, the conditional logistic regression achieved improved sensitivity over the standard logistic regression. This improvement exceeded 10% with both CLR2 and CLR3, especially in certain strata. However, the improved sensitivity might be associated with inferior specificity.

3. Conclusions and Future Research

The primary objective of this paper was exploring the use of conditional logistic regression, an alternative, yet easy to deploy statistic modeling approach, to improve the prediction of 30-day readmissions over HCUP, a publicly available large administrative database. Meanwhile, we expect that our proposed approach can more conveniently derive decision rules that appeal to the practitioners, as opposed to standard classifiers (e.g., standard logistic regression, random forest, and SVM). For this objective, we tested the applicability of conditional logistic regression and compared it with the standard classifiers, through cross validation.

The main finding from our comparative study was that incorporation of easily attainable decision rules can help improve the predicted classification accuracy especially the accuracy on patients who would truly be readmitted. We, thereby, argue that even though the improvement resulting from the conditional regression models is still somewhat modest, our research demonstrates the value of such ad-hoc decision rule construction and high-order polynomial regression modeling. As a result, our study is expected to help effective readmission risk management based on practically meaningful risk factors with potentially-improved decision making. Furthermore, the fact that these identified features are related to clinical condition and post-discharge management suggests that we make effort in collecting the relevant data (potentially extracting data from electronic health records) and integrating them with administrative data. Additionally, the state-wide data we used ensures general applicability of the proposed modeling approach and derived insights from the actual models developed. Lastly, it is worth noting that while advanced data analysis approaches have some potential when dealing with administrative databases (albeit not fully realized in this study), they tend to be overly complex at times. To physicians, such a complex model may appear to be a black box and would ultimately be of little practical importance, unless their complete trust is gained ahead of time; traditionally, physicians have relied on rather simple, fairly intuitive, models that are easy to understand, explain to their peers, implemented, and sustained. So unless such advanced data analysis approaches can lead to a consider-

Table 2 Classification model comparison

Classification Model		Prediction Accuracy*
LR: standard logistic regression		0.547
SLR: stepwise logistic regression		0.539
RF: random forests	all original variable	0.577
	only variables selected via SLR	0.574
SVM: support vector machine		0.560
CLR1: conditional logistic regression with 3 influence prediction variables	DISPUniform	0.548**
	NCHRONIC	0.564**
	NPR	0.576**
CLR2: conditional logistic regression with 4 data strata based on the first two layers of the decision tree		0.608**
CLR3: CLR2 + consideration of interacting variables based on identified influence ones in CLR2		0.615**

*Prediction accuracy = (true positive + true negative)/total # of subjects based on the test set.
**An optimal threshold was identified for each data stratum and the prediction accuracy is the combined measure over the multiple strata.

ably large improvement in the quality of prediction (e.g., high sensitivity and specificity), logistic regression type models may still remain the method of choice, On the other hand, our study demonstrated modest improvements through incremental updates within logistic regression. Thus, it shows a promising direction of balancing model sophistication and its usability for human decision making.

The main limitation of our study is the lack of availability of inpatient clinical and service information, as well as post-discharge care management information, in any publicly available administrative dataset. Although a set of comorbidity indicators and a number of utilization related measures were included as prediction variables, it was difficult to directly identify the disease severity and the quality of care delivery. As a result, the models developed in our study, using publicly administrative data, could not improve classification accuracy substantially any further compared to the models in the literature. In addition, our model development relied on statewide retrospective administrative data and did not include real-time individual-level data (e.g., detailed healthcare utilization data and data from patient surveys conducted during the hospitalization). Nevertheless, we can at least expect to use this study to raise the awareness of collecting data on additional markers and developing necessary database infrastructure for larger-scale exploratory studies on readmission risk prediction. Lastly, it is worth noting that publicly available administrative data can be quite noisy – having coding errors (e.g., upcoding and downcoding issues with ICD9), and entry inconsistencies (e.g., CM-diabetes was not consistently 1 for all subsequent admissions of a patient after it was recorded as such previously), etc. Hence, the models developed must be used judiciously for analyzing SID data for other diseases, diagnosis, and states.

Our future research aims at developing readmission risk prediction models for clinical intervention purposes. This aim requires us to make improvements along three directions. The first one is to investigate the influence of additional predictors that can be extracted from the HCUP data. For example, Smith et al. [20] reported that readmission is affected by the use of emergency room within six months prior to the index hospitalization. It is clear that more information on the influence of each predictor would help us refine the risk prediction model. This might also help us offer more generalizable and operationally feasible insights for care organization administrators. Studies in the past have suggested a diverse set of variables associated with disease severity, comorbid condition, care utilization and system, and social/behavioral determinants of health. The second direction is to explore alternative binary classification techniques, such as Bayesian networks, to further refine the classification model. The third one is to study more detailed prediction on hospital readmission, which has greater potential in designing efficient and effective post-discharge interventions. For example, it is interesting to predict the likelihood that a patient is re-hospitalized by a certain time point or interval post discharge, and investigate how the likelihood may be affected by the previous hospitalization incidences (e.g. [38, 39]).

Acknowledgments

The authors gratefully acknowledge the support of the U.S. National Science Foundation in supporting this study with two Grant Awards (#1405265 and #1405357).

References

1. Lindenauer PK, Bernheim SM, Grady JN, Lin Z, Wang Y, Wang Y, Merrill AR, Han LF, Rapp MT, Drye EE, Normand SL, Krumholz HM. The performance of US hospitals as reflected in risk-standardized 30-day mortality and readmission rates for Medicare beneficiaries with pneumonia. J Hosp Med 2010; 5 (6): E12–E18.
2. Jencks SF, Williams MV, Coleman EA. Rehospitalizations among patients in the Medicare fee-for-service program. N Engl J Med 2009; 360: 1418 –1428.
3. Centers for Medicare and Medicaid Services. The Medicare and Medicaid Statistical Supplement. 2013 Edition. Available at http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/MedicareMedicaidStatSupp/2013.html.
4. Ashton CM, Del Junco DJ, Souchek J, Wray NP, Mansyur CL. The association between the quality of inpatient care and early readmission: a meta-analysis of the evidence. Med Care 1997; 35 (10): 1044 –1059.
5. Kocher RP, Adashi EY. Hospital readmissions and the Affordable Care Act: paying for coordinated quality care. JAMA 2011; 306 (16): 1794 –1795.
6. Keenan PS, Normand SL, Lin Z, Drye EE, Bhat KR, Ross JS, Schuur JD, Stauffer BD, Bernheim SM, Epstein AJ, Wang Y, Herrin J, Chen J, Federer JJ, Mattera JA, Wang Y, Krumholz HM. An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure. Circ Cardiovasc Qual Outcomes 2008; 1 (1): 29 –37.
7. Krumholz HM, Lin Z, Drye EE, Desai MM, Han LF, Rapp MT, Mattera JA, Normand SL. An administrative claims measure suitable for profiling hospital performance based on 30-day all-cause readmission rates among patients with acute myocardial infarction. Circ Cardiovasc Qual Outcomes 2011; 4 (2): 243 –252.
8. Lindenauer PK, Normand SL, Drye EE, Lin Z, Goodrich K, Desai MM, Bratzler DW, O'Donnell WJ, Metersky ML, Krumholz HM. Development, validation, and results of a measure of 30-day readmission following hospitalization for pneumonia. J Hosp Med 2011; 6 (3): 142 –150.
9. Jack B, Chetty VK, Anthony D, Greenwald JL, Sanchez GM, Johnson AE, Forsythe SR, O'Donnell JK, Paasche-Orlow MK, Manasseh C, Martin S, Culpepper L. A reengineered hospital discharge program to decrease rehospitalization: A randomized trial. Ann Intern Med 2009; 150 (3): 178 –187.
10. Phillips C, Wright SM, Kern DE, Singa RM, Shepperd S, Rubin HR. Comprehensive discharge planning with post-discharge support for older patients with congestive heart failure: A meta-analysis. JAMA 2004; 291 (11): 1358 –1367.
11. Kansagra D. Risk prediction models for hospital readmission: A systematic review. Evidence-based Synthesis Program. Department of Veterans Affairs Health Services Research & Development Service. October 2011.
12. Wallman R, Llorca J, Gómez-Acebo I, Ortega AC, Roldan FR, Dierssen-Sotos T. Prediction of 30-day cardiac-related emergency readmissions using simple administrative hospital data. Int J Cardiol 2013; 164 (2): 193 –200.
13. Dharmarajan K, Hsieh AF, Lin Z, Bueno H, Ross JS, Horwitz LI, Barreto-Filho JA, Kim N, Bernheim SM, Suter LG, Drye EE, Krumholz HM. Diagnoses and timing of 30-day readmissions after hospitalization for heart failure, acute myocardial infarction, or pneumonia. JAMA 2013; 309 (4): 355–363.
14. Silverstein MD, Qin H, Mercer SQ, Fong J, Haydar Z. Risk factors for 30-day hospital readmission in patients ≥ 65 years of age. Baylor University Medical Center Proc 2008; 21 (4): 363 –372.
15. Reed RL, Pearlman RA, Buchner DM. Risk factors for early unplanned hospital readmission in the elderly. J Gen Intern Med 1991; 6 (3): 223 –228.
16. Corrigan JM, Martin JB. Identification of factors associated with hospital readmission and development of a predictive model. Health Serv Res 1992; 27 (1): 81–101.
17. Marcantonio ER, McKean S, Goldfinger M, Kleefield S, Yurkofsky M, Brennan TA. Factors associated with unplanned hospital readmission among patients 65 years of age and older in a Medicare managed care plan. Am J Med 1990; 107 (1): 13 –17.

18. Chu LW, Pei CK. Risk factors for early emergency hospital readmission in elderly medical patients. Gerontology 1999; 45 (4): 220 –226.
19. Jasti H, Mortensen EM, Obrosky DS, Kapoor WN, Fine MJ. Causes and risk factors for rehospitalization of patients hospitalized with community-acquired pneumonia. Clin Infect Dis 2008; 46 (4): 550 –556.
20. Smith DM, Giobbie-Hurder A, Weinberger M, Oddone EZ, Henderson WG, Asch DA, et al. Predicting non-elective hospital readmissions: a multi-site study. Department of Veterans Affairs Cooperative Study Group on Primary Care and Readmissions. J Clin Epidemiol 2000; 53 (11): 1113 –1118.
21. Runball-Smith J, Hider P, Graham P. The readmission rate as an indicator of the quality of elective surgical inpatient care for the elderly in New Zealand. N Z Med J 2009; 122 (1289): 32–39.
22. Oh HJ, Yu SH. A case-control study of unexpected readmission in a university hospital. Korean J Prev Med 1999; 32 (3): 289 –296 (Korean).
23. Thakar CV, Parikh PJ, Liu Y. Acute kidney injury (AKI) and risk of readmissions in patients with heart failure. Am J Cardiol 2012; 109 (10): 1482–1486.
24. Holloway JJ, Thomas JW, Shapiro L. Clinical and sociodemographic risk factors for readmission of Medicare beneficiaries. Health Care Financ Rev 1998; 10 (1): 27– 36.
25. Krumholz H, Normand S, Keenan P, Lin Z, Drye EE, Bhat KR, Wang Y, Ross J, Schuur J, Stauffer B, Bernheim S, Epstein A, Herrin J, Federer J, Mattera J, Wang Y, Mulvey G, Schreiner GC. Hospital 30-Day Heart Failure Readmission Measure: Methodology. Report prepared for Centers for Medicare & Medicaid Services. 2008.
26. Krumholz HM, Normand ST, Keenan PS, Desai MM, Lin Z, Drye EE, Curtis JP, Bhat KR, Schreiner GC. Hospital 30-Day Acute Myocardial Infarction Readmission Measure: Methodology. A report prepared for the Centers for Medicare & Medicaid Services. 2008.
27. Krumholz HM, Normand ST, Keenan PS, Desai MM, Lin Z, Drye EE, Bhat KR, Schreiner GC. Hospital 30-Day Pneumonia Readmission Risk Measure: Methodology. A report prepared for the Centers for Medicare & Medicaid Services. 2008.
28. Philbin EF, DiSalvo TG. Prediction of hospital readmission for heart failure: development of a simple risk score based on administrative data. J Am Coll Cardiol 1999; 33 (6): 1560 –1566.
29. Thomas JW. Does risk-adjusted readmission rate provide valid information on hospital quality? Inquiry 1996; 33 (3): 258–270.
30. Natale J, Wang S, Taylor J. A decision tree model for predicting heart failure patient readmissions. In: Krishnamurthy A, Chan WKV (eds.). Proceedings of the 2013 Industrial and Systems Engineering Research Conference, May 18 –22, San Juan, Puerto Rico.
31. Lee EW. Selecting the best prediction model for readmission. J Prev Med Public Health 2012; 45: 259 –266.
32. Hosseinzadeh A, Izadi M, Verma A, Precup D, Buckeridge D. Assessing the predictability of hospital readmission using machine learning. In: Munoz-Avila H, Stracuzzi D (eds.). Proceedings of the Twenty-Fifth Innovative Applications of Artificial Intelligence Conference, July 14 –18, 2013, Bellevue, Washington. Published by The AAAI Press, Palo Alto, California.
33. Desai MM, Stauffer BD, Feringa H, Schreiner GC. Statistical models and patient predictors of readmission for acute myocardial infarction a systematic review. Circ Cardiovasc Qual Outcomes 2009; 2 (5): 500 – 507.
34. van Walraven C, Dhalla IA, Bell C, Etchells E, Stiell IG, Zarnke K, Austin PC, Forster AJ. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. Can Med Assoc J 2010; 182 (6): 551– 557.
35. Hastie TJ, Pregibon D. Generalized linear models. Chapter 6 In Chambers S, Hastie TJ (eds.). Statistical Models. Wadsworth & Brooks/Cole; 1992.
36. Bishop CM. Pattern Recognition and Machine Learning. New York: Springer; 2006.
37. Vittinghoff E, Gildden DV, Shiboski SC, McCulloch CE. Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models. New York: Springer; 2005.
38. Yu S, van Esbroeck A, Farooq F, Fung G, Anand V, Krishnapuram B. Predicting readmission risk with institution specific prediction models. Proceedings of 2013 IEEE International Conference on Healthcare Informatics (ICHI), IEEE, Philadelphia, PA, 2013. pp 415– 420.
39. Helm J, Alaeddini A, Stauffer JM, Bretthauser KM, Skolarus TA. Reducing hospital readmissions by integrating empirical prediction with resource optimization. Production and Operations Management 2015. Published online.