Draft Proposal: Building a Predictive Model for Hospital Readmission Using LASSO

David K. Kirui

January 30, 2020

1 Executive Summary

Diabetes mellitus is a group of chronic conditions that drastically impacts the quality of life of millions of people in the United States and around the world. With proper management, afflicted individuals can lead relatively fulfilling lives. Without proper management, however, people risk short-term hyper-utilization of the heath care system. Moreover, it is extremely costly for hospitals to accommodate the short-term readmission of hypoglycemic patients. For this reason, the Centers for Medicare and Medicaid Services recently announced that they would discontinue reimbursement of hospitals for services administered to diabetic patients who were re-admitted to the hospital with complications within 30 days of their original discharge. Because of this, identifying patients most at risk of short-term readmission has become an imperative for hospitals.

To address this question, I use a dataset representing 10 years of diabetic patient care at 130 hospitals throughout the United States that was assembled through the HeathFacts national warehouse of comprehensive clinical patient records. I employ machine-learning techniques to identify the correlates best able to predict short-term hospital readmission. I then estimate binary logistic regression models to predict short-term readmission. I find that the number and nature of past contacts with the healthcare system has a significant impact on predicting short-term readmission. I also find that certain characteristics of present patient encounters and discharge have some bearing on short-term readmission. Although instructive, these results have a fair amount of caveats, chief among them the nature in which the data were generated.

2 Data Summary

These data were drawn from the Center for Clinical and Translational Research at Virginia Commonwealth University. They consist of data on diabetes patients at approximately 130 hospitals across the United States. These data are panel - that is they consist of

repeated observations across time. Indeed, we know that the sample consist of observations ranging from 1998 to 2008. Although there are over 100,000 observations, there are about 70,000 unique patients in the sample, and 16,773 who have multiple hospital admissions documented in the data (some patients have in excess of 4). All patients in this dataset have some form of diabetes. Although additional insight could be gleaned by treating these data as longitudinal, for the purposes of these analyses, we will treat them as cross-sectional mostly because not all patients have repeated observations which would lead to a large number of missing data points if we were to consider the data truly longitudinal.

Before I could begin my analysis, I had to prepare the data in a number of ways. Fortunately, the data had already been mostly clean but some additional preparation work was necessary to get it into a form conducive to analysis. First, and perhaps most importantly, I had to recode the dependent variable that measured whether and when an individual was readmitted to the hospital. Since our event of interest is readmission within 30 days or less, I dichotomized this variable so that individuals were assigned a value of 0 if they were either not readmitted to the hospital or readmitted after 30 days and a value of 1 if they were readmitted to the hospital within 30 days. See Figure 2 for a visual representation of cases on this variable (recode_readmitted). Just 11.2% of the cases in the sample experienced hospital readmissions within 30 days, while the large majority did not. Since this is our event of interest, the fact that there were relatively few cases of it occurring has implications on the robustness and generalizability of our results.

There were a number of additional measures included in these data, all of which were considered potential predictors during my model selection/regularization processes. These included variables that measured Demographic characteristics such as Race (Caucasian, Asian, African-American, Hispanic, and Other), Gender (Male, Female), and Age (grouped from 0-19, 20-59, 60-79, and 80+). There were also missing values on the race variable, which were dropped during the data cleaning process. The sample is mostly White, with African-American patients being the second-largest racial/ethnic group (for a visualization of the distribution see Figure 1). The variable that measured gender was binary, with slightly more women in the sample than men (see Figure 1). Patients aged between 60 and 79 made up the majority of observations, followed by patients aged 20-59 and patients 80 years or older; patients below 20 made up a very small proportion of observations (See Figure 1) There was also a variable that measured the time in days that the patient spent in the hospital during each encounter; observations were right skewed on this variable. In addition, there were variables that measured the number of lab (left skewed) and non-lab procedures (right skewed) performed during the encounter (num_lab_procedures and num_procedures respectively), as well as the number of generic medications administered during the encounter (right skewed), the number of outpatient visits the patient had in the year preceding the encounter (right skewed), the number of emergency visits the patient had in the year preceding the encounter (right skewed), and the number of inpatient visits the patient had in the year preceding the encounter (right skewed). Descriptive statistics for this set of quantitative variables can be found in Table 1. Most of the quantitative

variables in the dataset were decidedly right skewed, indicating that most observations have relatively few health system encounters, while a minority had a substantial number. This suggests that patients who had substantial/sustained contact with the healthcare system made up a small number of patients/observations in our sample. The dataset also included information about the number of diagnoses in the system (left skewed), as well as the results of glucose serum and A1C tests (values were "none", "normal", ">200" and ">300" for glucose serum and "none","normal", ">7" and ">8" for A1c. Distribution of A1c test results can be found in Figure 1; glucose serum results can be found in Figure 2.

In both cases, the majority of patients were not tested for either glucose serum or A1c, which seems a bit strange given that all of the observations in this dataset are of patients with discernable diabetes. There were also a number of categorical measures related to medications. This included a categorical measure that indicated whether insulin was prescribed to the patient or if their was a change in the dosage (Values: Up if increased, Down if Decreased, Steady if maintained, No if not taking/prescribed) (A visualization of the distribution of cases on this variable can be found in Figure 1. In addition to insulin there were six additional variables that measured diabetes-related medication. These variables were initially measured on the same scale as the insulin measure reported above, but I recoded them as binary (Taking the drug/Not taking the drug) because there was little variability in any of the categories on the original variable, and the large majority of observations did not report taking the drug. These drugs were Metformin, Glimepiride, Glipizide, Glyburide, Pioglitazone, and Rosiglitazone. Visualizations of the distribution of observations on each of these recoded variables can be found in Figure 3. Moreover, there was a variable that measured whether or not there was a change in any diabetic medications during the encounter (see Figure 1). There was also a variable that measured whether or not any new diabetic medications were prescribed during the encounter (see Figure 2). There were also variables that measured 1) where the patient was discharged to after their encounter, 2) who referred the patient for admission to the hospital, 3) and what type of hospital admission it was (see Figure 2 for visualizations of the distributions on these variables). Most patients were discharged to their homes, and patients were most often referred from/admitted from the Emergency Department (ED). There were also variables that measured what each of three possible diagnoses made during the encounter were: 1) the patients primary diagnosis, 2) the patients secondary diagnosis (if any), and 3) the patients tertiary diagnosis (if any). Each of these three diagnosis-related variables were coded as first three digits of ICD9 classification code, but for ease of interpretation I recoded them to correspond to the ailment associated with each code. Frequencies of cases that fell into each diagnostic category on for each of these three variables can be found in Tables 2, 3 and 4 respectively. I re-leveled a number of the categorical variables to ensure that they all had the same direction and their base category was no (when applicable); this was done to aid interpretation of subsequent analyses.

3 Preliminary Analyses

In order to estimate the impact of the covariates in my dataset on the probability of short-term hospital readmission (within 30 days), I estimated a series of binomial logistic regressions predicting the probability of short-term readmission. However, before I could estimate my predictive models, I first had to employ a number of techniques to increase the predictive power of them. Moreover, I had to cut down the number of potential predictors to increase the parsimoniousness of my model and to achieve the desired level of prediction. Moreover, in order to prevent over fitting and data snooping, I split my dataset into two random subsamples, training and test data. This was also done in order to evaluate the performance of my analysis models and associated procedures using unseen and untouched data. To that end, I fit my analysis models on the training subset of my data and subsequently estimated their performance in the testing subset of my data.

After I split my data, I conducted LASSO regression analyses with each of the three possible cross-validation losses: the deviance, the AUC, and misclassification error. I used the cv.glmnet() function to accomplish this in R; the bestglm() function would not work because of the size of the dataset (it also would not work even after I subset the dataset further). Each time I set the random number seed to ensure the same k-fold (k=10) cross-validation would be enacted, allowing my results to be reproducible. First, I ran a LASSO model with the cross-validation loss set to deviance (values of the logarithm of lambda are displayed in Figure 4). Next, I evaluated the number of non-zero coefficients returned by various values of Lambda. In this instance, Lambda First (0.01551572) returned just two covariates; while Lambda Min (0.001146724) returned 23. Next, I ran a LASSO model with the cross-validation loss set to the area under the ROC curve (AOC) (displayed in Figure 5). This model yielded a model with 8 covariates when Lambda First (0.005576061) was used and 28 covariates when Lambda Min (0.001515901) was used. Finally, I ran a LASSO model with the cross-validation loss set to the misclassification error. This technique yielded the least useful LASSO models. Using Lambda First (0.05200253) yielded a model with no covariates; using Lambda Min (0.01413735) yielded a model with 2 covariates. After careful consideration, I decided to move into the next phase of variable selection with two models from my LASSO analysis: 1) the deviance model that specified lambda at its minimum value, and 2) the AUC model that specified lambda at its first value. I chose these models because the other models that my LASSO analysis yielded were either too large (with too many potential non-zero covariates) or too small (with too few or no non-zero coefficients).

After, I concluded the LASSO analyses; I conducted manual backward deletion on both candidate models. Again, I tried to use bestglm() but was unsuccessful. Because of the substantial difference in the number of covariates yielded by the two LASSO analyses, I arrived at the final model with using the LASSO AUC model much quicker than with the LASSO Deviance model. My process of backwards selection was as follows. First, I ran each of the models with all of the potential covariates included and then ran the Anova()

function on the resulting model. I would then exclude whatever covariate had the highest p value from the subsequent model. For good measure, I would also estimate an ROC curve and a corresponding value of AUC for each model at each step using the testing data. I did this for both the AUC and Deviance candidate models until all of the covariates in each of them reached statistical significance (using Anova()) at the a = .01 level. I increased the requisite alpha level from .05 here mostly to aid in increasing the parsimoniousness of the final models. Full R code and annotations for this analysis (and all of my analyses) can be found in my RMarkdown file, which I have turned in with this paper. For the AUC model, it took one iteration of this process before I arrived at a model in which all variables reached statistical significance. For the Deviance model, on the other hand, fourteen iterations were necessary to reach the same point.

The LASSO AUC model yielded a final, most parsimonious, model with seven covariates: number of emergency room visits made by the patient within the prior year, number of inpatient visits within the prior year, number of diagnoses entered into the system, insulin, whether or not diabetes medication was prescribed, discharge disposition (where the patient was discharged to), and primary diagnoses. The final LASSO Deviance model yielded the same seven covariates as well as the number of non-lab procedures performed during the encounter, and the number of generic medications administered during the encounter. After I concluded backwards selection, I then used the resulting two models that were fit with the training data (which I refer to as the Deviance model and the AUC model) and evaluated their performance on a number of metrics with the testing data. First, I estimated ROC curves and values for AUC for each model; values for AUC in each instance were very similar. The AUC model yielded an AUC of .6459 while the Deviance model yielded an AUC of .6464. If you rounded each AUC value to three decimal place accuracy, there would be no difference between the two values. Moreover, given that the AUC model has two fewer covariates than the Deviance model, this slight reduction in the AUC makes intuitive sense.

Next, knowing that it costs twice as much to mislabel a readmission than it does to mislabel a non-readmission, I came up with a cost ratio of and computed a Bayes classifier (threshold) of .333 (1/3). Using my threshold of 1/3 and the test data, I calculated two confusion matrices gauging the prediction accuracy of both of my final candidate models (these confusion matrices are not in this paper but can be computed using the code in my RMarkdown file they are labeled cm.33). The deviance model yielded a sensitivity rate of 0.01379681, a specificity rate of .9882713, a false positive rate of 0.01172874, an un-weighted misclassification error of 0.121055, a positive prediction rate of 0.1294118, a negative prediction rate of 0.8880186, and a weighted misclassification error of 0.231697. The AUC model yielded a sensitivity rate of 0.01361763, a specificity rate of 0.9883618, a false positive rate of 0.01163818, an un-weighted misclassification error of 0.1209947, a positive prediction rate of 0.1288136, a negative prediction rate of 0.8880096, and a weighted misclassification error of 0.2316568. Both models are very similar and yield very similar estimates of the covariate effects and performance estimates.

For the purposes of this draft proposal, I will select the AUC model as my final, most parsimonious model as I think the trade offs in terms of minor reduction in the AUC and in some aspects of the prediction performance are an acceptable price to pay for increased parsimony. Nevertheless, I have chosen to display the results of both final logistic regression models in this paper. Exponentiated coefficients (odds ratios) and other relevant results from each candidate model are displayed in Table 5. The covariate effects for each model are very similar, with a few minor differences. In this paper, I will only interpret the statistically significant coefficients from the AUC model but the effects in the Deviance model are largely almost identical. First, the number of previous emergency department visits was associated with an increased likelihood of short-term hospital readmission (¿30 days). Specifically, each additional ED visit within the past year was associated with a 2.8% increase in the likelihood of short-term readmission, net of over covariate effects. Moreover, the number of previous in-patient visits within the past year was associated with an increased likelihood of short-term readmission. Specifically, each additional in-patient visit was associated with a 30.8% increase in the likelihood of short-term readmission, net of other covariate effects. Furthermore, the number of diagnoses entered into the system was associated with an increased likelihood of short-term readmission. Specifically, each additional diagnosis entered into the system was associated with 5.1% increase in the likelihood of short-term readmission, net of other covariate effects. Lowering the dosage of insulin among patients who were already on insulin was associated with a 18.2% increase in the likelihood of short-term readmission, compared to patients who weren't on insulin and net of other covariate effects. Similarly, prescribing diabetes medication during patient encounters was associated with a 17.8% increase in the likelihood of short-term readmission, net of other covariate effects. Certain discharge dispositions were associated with a decreased likelihood of short-term readmission. Specifically, discharging patients to their home both without and with home health services was associated with 37.3% and 22.9% decreases in the likelihood of short-term readmission respectively, compared to patients who were discharged to other venues and net of other covariate effects. Finally, certain diagnoses were associated with a decrease in the likelihood of short-term readmission. Diagnoses associated with diseases of the respiratory system (24.5%), sense organs (24.8%), and skin and subcutaneous tissues (24.2%) were all associated with decreases in the likelihood of short-term readmission, as well patients who presented with symptoms, signs, and Ill defined conditions (23.9%), compared to patients who were diagnosed with diseases of the circulatory system and net of other covariate effects. This result could partially be a consequence of the reference category. Although I will not do so for the purposes of this paper, additional analysis where the reference category is changed to other could be both useful and instructive in teasing out the nature of these effects.

4 Preliminary Conclusion and Next Steps

These analyses have shown us that there are a number of factors that influence short-term readmission in the hospitals in this dataset. The number of past inpatient visits and ED visits that a patient has made are both associated with an increased likelihood of short-term readmission. Hospitals concerned about their short-term readmission rate should monitor these prior visits closely in an effort to reduce them without compromising the standard of care. It could also be true, however, that sicker patients are both more likely to have prior visits of these types and be readmitted in the short-term. More data could be useful here. Moreover, the number of diagnoses entered into the system is associated with an increased likelihood of short-term readmission. Reducing the dosage of insulin among patients who regularly take insulin also appears to have an impact on short-term readmission, as does prescribing diabetes medication during an encounter. These two findings, in concert with the fact that the majority of encounters in this dataset were referred from the Emergency Department suggest that many of the patients in this dataset may be using the ED as their first point of contact with the healthcare system as is common among lower-income patients. Information on patients income could be useful here. Generally, it could be advantageous to ensure that more patients are more fully integrated (e.g. expanding access to home health services, etc.) into the healthcare system so that they are more proactive about their healthcare needs and concerns dont rise to the level of the ED before they are addressed.

Moreover, discharges to patients' homes both with and without home health services appear to be associated with decreased likelihood of short-term readmission. Here, the effect of home health services is particularly striking. It could be advantageous for hospitals to invest in more preemptive measures and home health services that allow patients and practitioners the ability to continuously monitor health metrics. Certain diagnoses appear to be associated with a reduction in the likelihood of short-term readmission but the nature of these effects is unclear and may warrant additional analysis. It could be that these diagnoses are more acute in nature and thus warrant more intensive care, which decreases the likelihood of short-term admission. However, it is unclear if this is true. Going forward, I will need to explore additional options for model building using various machine learning algorithms. After I derive the most accurate and satisfactory model, then I can start working on building the application and user interface for the end user - hospital administrators and other healthcare professionals. This initial analyses will serve me well in determining the next steps to take to continue to build and tune my predictive models.

Table 1: Descriptive Statistics for Continuous Covariates

Statistic	N	Mean	St. Dev.	Min	Median	Max
time_in_hospital	99,492	4.398	2.987	1	4	14
num_lab_procedures	99,492	43.073	19.696	1	44	132
num_procedures	99,492	1.341	1.704	0	1	6
num_medications	99,492	16.027	8.120	1	15	81
number_outpatient	99,492	0.373	1.277	0	0	42
number_emergency	99,492	0.201	0.940	0	0	76
number_inpatient	99,492	0.643	1.271	0	0	21
number_diagnoses	99,492	7.439	1.926	1	8	16

Table 2: Primary Diagnosis

	summary.readmission_data2.diag1_name.
circulatory system	22359
digestive system	1032
endocrine, nut & meta, immunity	4630
genitourinary system	3060
injury and poisoning	3002
musculoskeletal system & connective tissue	2099
Other	45951
respiratory system	7786
sense organs	1648
skin & subcutaneous tissue	1996
symptoms, signs, & ill-defined conditions	5929

Table 3: Secondary Diagnosis

	summary.readmission_data2.diag2_name.
blood & blood-forming organs	1484
circulatory system	26428
endocrine, nut & meta, immunity	16062
genitourinary system	6700
Other	36612
respiratory system	7426
skin & subcutaneous tissue	3326
symptoms, signs, & ill-defined conditions	1454

Table 4: Tertiary Diagnosis

	summary.readmission_data2.diag3_name.
blood & blood-forming organs	1177
circulatory system	24508
endocrine, nut & meta, immunity	20634
external causes & supplemental	1373
genitourinary system	3856
Other	42736
respiratory system	2555
skin & subcutaneous tissue	1342
symptoms, signs, & ill-defined conditions	1311

Figure 1: Distributions of Categorical Covariates (1 of 3)

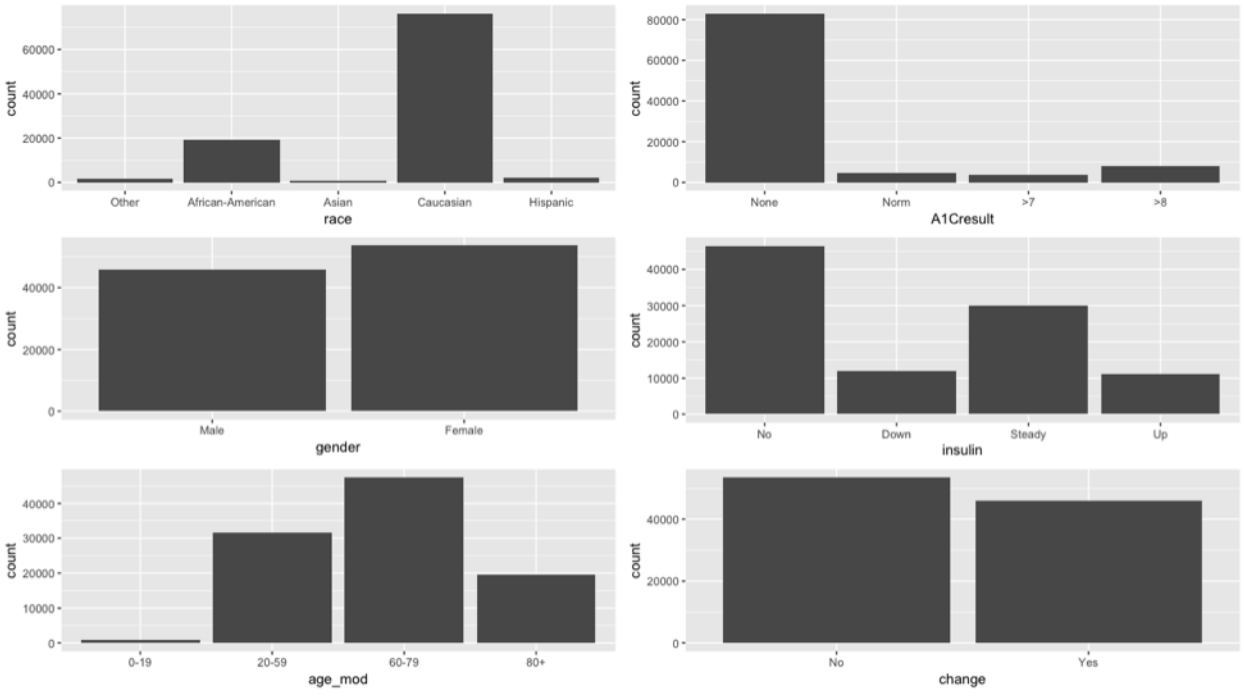

Figure 2: Distributions of Categorical Covariates (2 of 3)

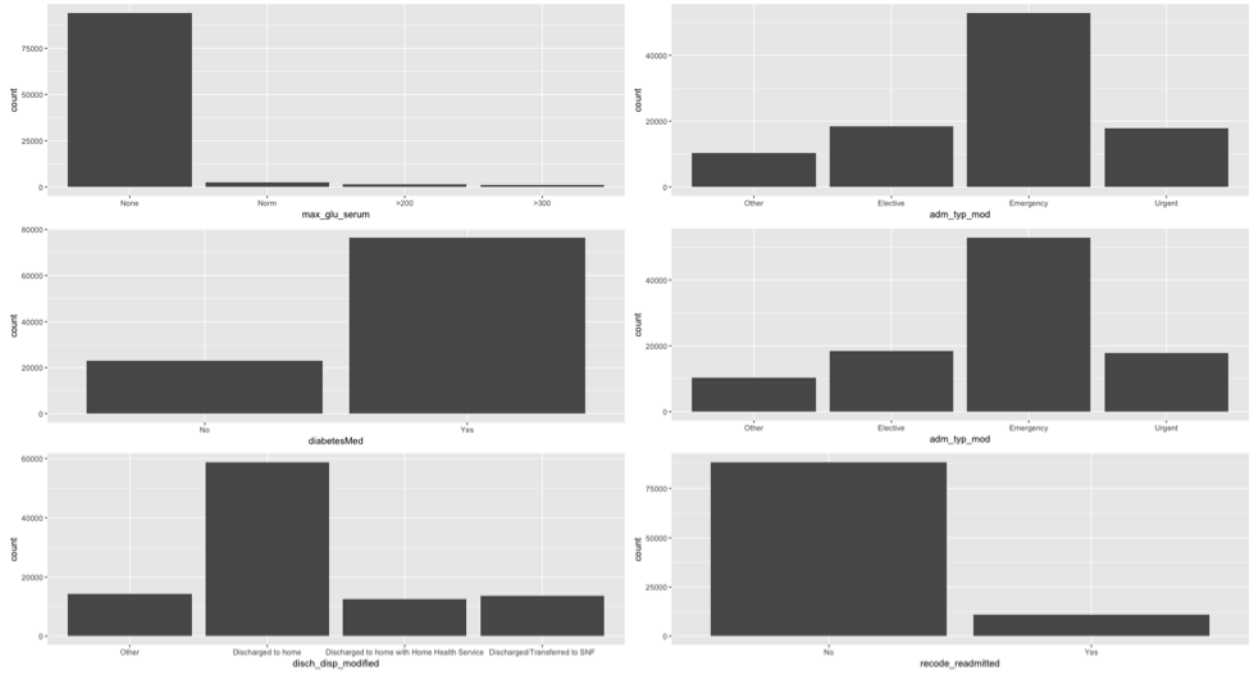

Figure 3: Distributions of Categorical Covariates (3 of 3)

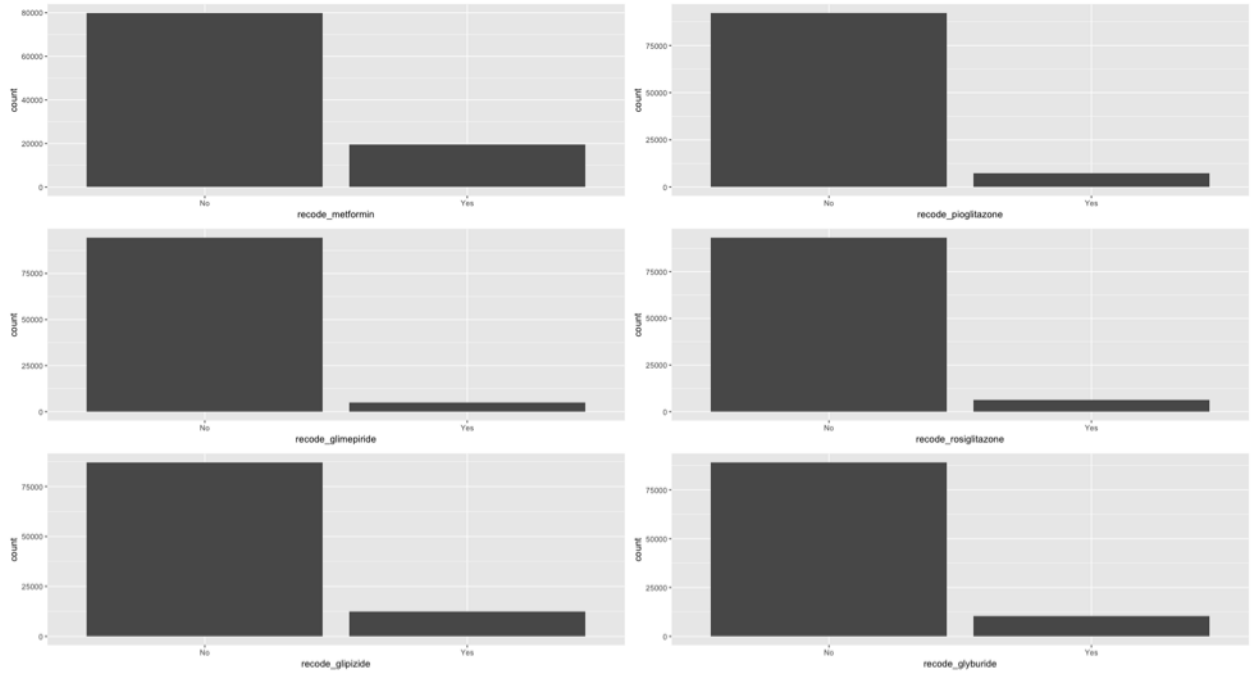

Figure 4: Log(Lambda) Values for LASSO w/ Deviance

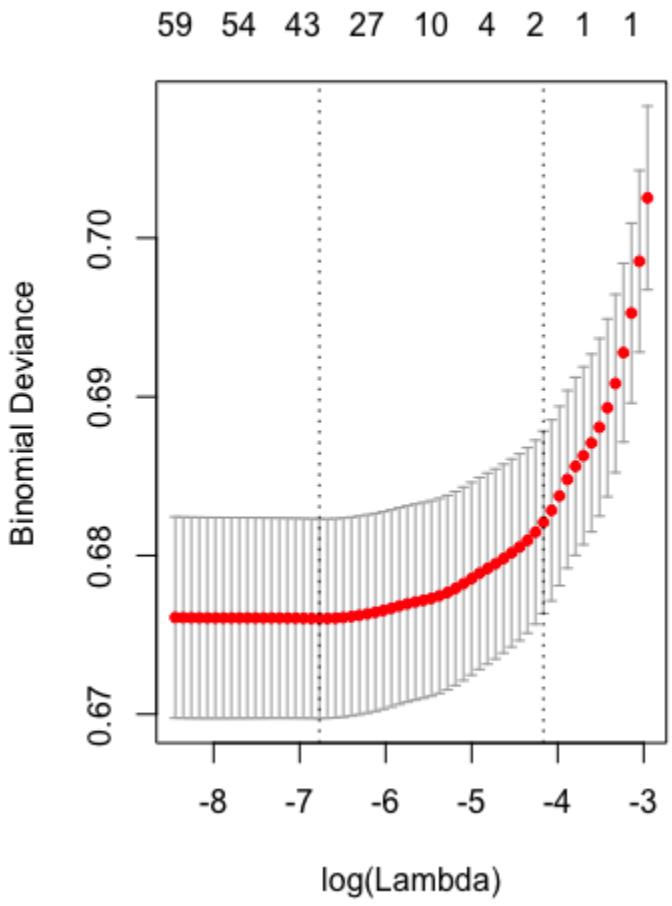

Figure 5: Log(Lambda) Values for LASSO w/ AUC

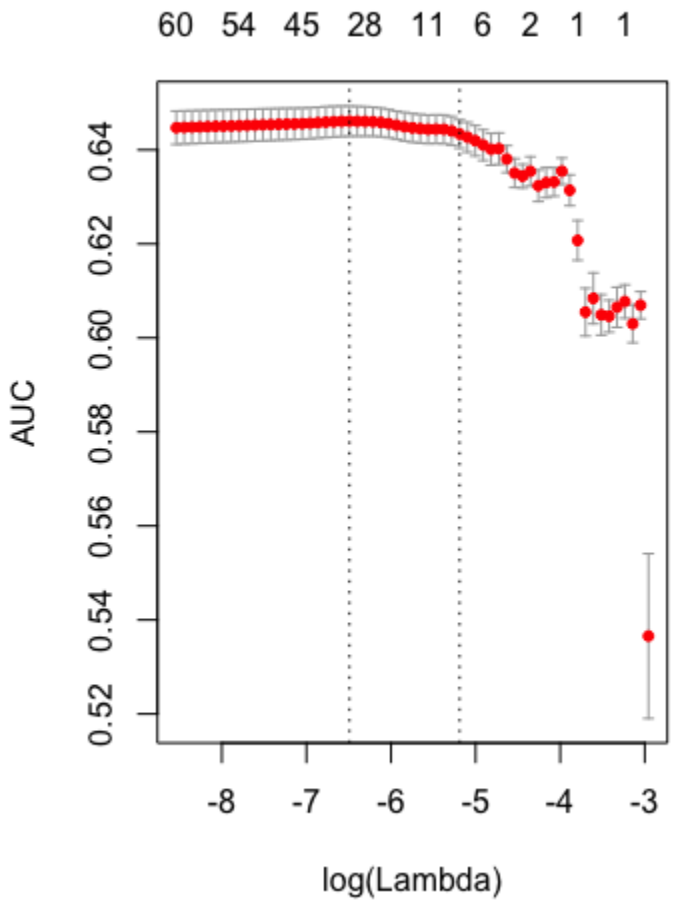

Table 5: Odds ratios predicting likelihood of readmission within 30 days

	exp(b)	
	AUC	Deviance
Number of Procedures	-	0.981*
Number of Medications	-	1.006**
Number of Emergency Visits	1.028*	1.028*
Number of Inpatient Visits	1.308***	1.305***
Number of Diagnoses	1.051***	1.047***
Insulin (Ref: No)		
Down	1.182***	1.165**
Steady	1.023	1.019
Up	1.062	1.042
DiabetesMed (Ref: No)	1.178***	1.164***
Disch_disp_modified (Ref: Other)		
Discharged to Home	0.627***	0.633***
Discharged to Home w/ Home Health Service	0.771***	0.767***
Discharged/Transferred to SNF	0.909	0.906*
Primary Diagnosis (Ref: Circulatory System)		
Digestive System	1.223	1.215
Endocrine, Nut, & Meta, Immunity	1.113	1.112
Genitourinary System	0.859	0.855
Injury and Poisoning	1.158	1.155
Musculoskeletal System & Connective Tissue	0.932	0.900
Other	0.964	0.962
Respiratory System	0.755***	0.735***
Sense Organs	0.752*	0.744*
Skin & Subcutaneous Tissue	0.758*	0.750*
Symptoms, Signs, & Ill-Defined Conditions	0.761***	0.760***
Constant	0.084***	0.082***
Observations	49,746	49,746
Log Likelihood	-16775.300	-16771.260
AIC	33,592.600	33,588.530
AUC	0.6459	0.6464

p<.001 *** , p<.01 ** , p<.05 *

Note. Standard errors omitted; results computed from testing data.