PREDICTIVE ANALYTICS FOR 30-DAY HOSPITAL READMISSIONS

LU XIONG a,b,* AND TINGTING SUN b

aDepartment of Mathematical Sciences
bComputational and Data Science Doctoral Program
Middle Tennessee State University
1301 E. Main St., Murfreesboro, TN 37132, USA

RANDALL GREEN
Fidelis Analytics Healthcare Consulting
7113 Commonwealth Circle, Nashville, TN 37221, USA

(Communicated by Ding-Xuan Zhou)

ABSTRACT. The 30-day hospital readmission rate is the percentage of patients who are readmitted within 30 days after the last hospital discharge. Hospitals with high readmission rates would have to pay penalties to the Centers for Medicare & Medicaid Services (CMS). Predicting the readmissions can help the hospital better allocate its resources to reduce the readmission rate. In this research, we use a data set from a hospital in North Carolina during the years from 2011 to 2016, including 71724 hospital admissions. We aim to provide a predictive model that can be helpful for related entities including hospitals, health insurance actuaries, and Medicare to reduce the cost and improve the clinical outcome of the healthcare system. We used R to process data and applied clustering, generalized linear model (GLM) and LASSO regressions to predict the 30-day readmissions. It turns out that the patient’s age is the most important factor impacting hospital readmission. This research can help hospitals and CMS reduce costly readmissions.

1. Introduction. 30-day hospital readmission means the patient who is readmitted to the same hospital, or another applicable acute care hospital for any reason was re-admitted within 30 days of discharge from the index (i.e., initial) admission [26]. About 20% of Medicare patients are readmitted within 30 days [27], 34% are readmitted within 90 days [13]. More than $41 billion [10] was spent on diabetes patients only, by hospitals in the United States in 2011, on their 30-day readmission, not even mention patients with other diseases. $17 billion could be saved per year for avoidable readmissions of Medicare patients alone [28]. Note that there is other 44% of the readmitted patients are not Medicare patients [29].

One readmission can double the cost of patient care. It costs Medicare $15,000 for a patient who was admitted once, but $33,000 for a patient who was readmitted once, according to the statistics in 2012 [27]. It is predicted that Medicare will be insolvent in 2024 or 2026 [30] if we don’t reform. Most of these readmissions

2020 Mathematics Subject Classification. Primary: 62P10, 62J05; Secondary: 91C20.
Key words and phrases. Generalized Linear Model (GLM), Predictive Analytics, Clustering,
The first author is supported by MTSU FRCAC grant 2019.
* Corresponding author: Lu Xiong: lu.xiong@mtsu.edu.
are avoidable [13]. Studies have shown that transitional care interventions can save nearly $4,000 per patient within 6 months [8]. Preventing 10% of readmissions could save Medicare $1 billion [20]. Research shows that hospital readmission rates vary by different factors, which allows us to improve the quality of care and save money by incentivizing providers to reduce excessive readmissions.

Hospital Readmission Reduction Program (HRRP) is a Medicare value-based purchasing program that lowers payments to Inpatient Prospective Payment System (IPPS) hospitals with excess readmissions, established in 2012 by the Affordable Care Act (ACA). ACA tries to reduce the number of excessive hospitalizations, reduce medical expenses, and improve patient safety and outcomes [7]. There are three initially target conditions in HRRP, acute myocardial infarction (AMI), heart failure (HF), and pneumonia (PN). Several target conditions were added in 2015, including chronic obstructive pulmonary disease (COPD), coronary artery bypass grafting (CABG) surgery, selective primary total hip replacement, and/or total knee replacement (THA / TKA) [31]. Total HRRP penalties in fiscal years (FY) 2013, 2014, and 2015 were $280 million, $227 million, and $428 million, respectively [22]. A study [22] shows that HRRP has led to an increase in the 30-day readmission rate (target indicator) of the policy target population and patients with non-target insurance types or clinical illnesses by 3% to 6%, and there are significant improvements in non-targeted measures such as 31–60 days readmissions and hospitalization costs.

The prediction of readmission is of significance in multiple aspects. From the patients’ view, this can improve the quality of recovery. From the view of the hospitals and Medicare system, 30-day readmission means low health outcomes, ineffective treatment, increasing the burden of the patients and the caregivers, causing waste for valuable and expensive healthcare resources. The research we propose here can help in reducing the huge spending of the system. From the view of hospitals, this can help hospitals improving the effectiveness of their healthcare service and reducing the penalties they may get from the Centers for Medicare & Medicaid Services (CMS) because of the high rate of readmission. 78% of the hospitals in the US received the Hospital Readmission Reduction Penalty because of exceeding the readmission rate limit set by CMS [2] in the financial year 2015. Also, the development of this predictive model can help further analyze the factors that cause readmission, which can help to increase the quality of healthcare the patients receive in the future. Once the high-risk readmission discharges are identified, more resources can be spent on them to reduce their readmission rate, for example, contacting them within 24~48 hours, depending on the patient’s risk of readmission to make sure they are following the discharge plan. Discharging the high-risk readmission patients to the most appropriate transitional care setting can also be a choice to reduce their readmission after the model predicted that the patient was at high risk of readmission before discharge.

CMS published 3 models in 2008 to predict readmission for three types of diseases. The performances of their models are relatively poor, which have c-statistics (a measurement of model accuracy) 0.63 for acute myocardial infarction (AMI), 0.61 for congestive heart failure (CHF), and 0.63 for pneumonia [15, 16]. Note that only when the c-statistics is greater than 0.8, the model can be considered as a strong model. In 2011, a study [14] reviewed 30 models in readmission prediction, it turns out most of them didn’t have high enough prediction accuracy, with c-statistics ranging from 0.55 to 0.83. In 2015, a study [3] reviewed the machine learning
models including Random Forest, LogitBoost, etc. on readmission projection, their area under the curve (AUC, a measurement of model classification accuracy) ranging from 0.75 to 0.8, which are fair but not strong accuracy. In 2016, a study [23] established a model with total accuracy for the ensemble ranges from 81% to 85%, which is better accuracy. In 2018, Health Catalyst published a report that says they developed a model with AUC=0.784 [32] that met the goal of their hospital client, which outperformed the LACE index method that is currently being used in many hospitals, but this method usually performs poorly in clinical practice under imbalanced data. A study [4] proposed an SVM-based method that achieves an accuracy of 81.02%, a sensitivity of 82.89%, a specificity of 79.23%, which performs better in clinical practice. Another study [9] reviewed four prediction models including logistic regression, gradient boosting, max out networks, and DUNs with AUC ranging from 0.664 to 0.705. In 2020, a study [25] reviewed the LACE scores system and six single machine learning classifiers as the initial benchmark, including LR (Logistic Regression), NB (Naive Bayes), DT (Decision Tree), SVM (Support Vector Machine), ANN (Artificial Neural Network), and RF (Random Forest) with AUC ranging from 0.496 to 0.63. It applied a joint ensemble-learning model which improves AUC to 0.879, considering data imbalances and various misclassifications of clinical interventions.

For chronic obstructive pulmonary disease (COPD), a study [17] applied seven models including LR, KNN (K-Nearest Neighbors), NB, NN (Neural Network), SVM, RM, BN (Bayesian Network), with accuracy ranging from 0.77 to 0.88. For heart failure (HF), a study [18] showed a deep learning approach based on convolutional neural networks (CNN), which achieves an F1 score of 0.756 (F1 score is a measurement of model classification accuracy). The model developed by Huynh QL et al. has the best predictive value with C-statistic = 0.85 for HF readmission and mortality [18]. “Early readmissions among older adults hospitalized for acute myocardial infarction (AMI) are costly and difficult to predict. Aging-related functional impairments may inform risk prediction but are unavailable in most studies,” as reported by NewsRx journalists [5]. A study [5] applied a multivariate model with AUC=0.753. A model was developed to predict 30 days readmissions for pneumonia with C-statistic ranging from 0.66 to 0.75 [11]. For coronary artery bypass grafting (CABG) surgery, a study [19] sought to develop a model with C-statistic = 0.63.

The innovations of our paper are as follows:

1. The current methods reviewed above focus on the classification problem, which is to predict whether a patient would be readmitted within 30-days or not. In this research, we focus on the regression problem, which is to predict how many days a discharged patient would be readmitted. Very little research has studied the days to readmission from discharge (gap days). This gap days variable is an important metric to measure the health outcome as well as reducing the readmission rate.

2. We apply the generalized linear model (GLM), which is different than the methods we listed in the literature reviews. As far as we know, no research has been done before by applying GLM to study the readmissions of all disease types. Though, one researcher has used GLM to study the readmission of chronic obstructive pulmonary diseases (COPD) [31]. Our research is more universal for all types of diseases. The GLM has the advantage of easy interpretation and high robustness. It doesn’t require the target variable to
follow the normal distribution. The target variable (gap days) in readmission
is not normal distributed, instead, it’s right skewed distributed, where lots of
readmissions happen within a short time, fewer readmitted after a long time.
It’s a long tail right-skewed distribution. Therefore, GLM works better for
this type of distributed data. Also, the GLM can find out the importance
each variable with a clear explanation. Plus, it’s easy to be implemented
even in a spreadsheet with an easy single formula to calculate the predicted
target variable using the predictor variables, and this is not easy to be done
for algorithms like random forest or CNN.

We propose here applying the GLM based on feature selection using Akaike in-
formation criterion (AIC) and Bayesian information criterion (BIC) methods to
predict the 30-day readmissions. We will discuss the advantages of choosing this
method compared with other methods such as LASSO, ordinary least squares re-
gression (OLS). We considered OLS, GLM, with feature selections using AIC, BIC,
and considered different combinations of distribution and transform function, then
finally comes to the final model. The final model is helpful to the related entities
including hospitals, health insurance actuaries, and Medicare to help to reduce the
cost, improving the clinical outcome of our healthcare system.

2. Data description, visualization, and pre-processing. The data used in
this study contains 71724 records of patient visits in a network of all the hospitals
in North Carolina from 2011 to early 2016. A description of the data is provided
in Appendix A. The number of missing entries is relatively small compared with
the total amount of data. Therefore, we think the data is reliable for the intended
purpose to predict the 30-day readmission. Because the data is from NC state
only, the findings in this research may not be generalized well to the readmissions
nationwide.

We define the target variable: days between the admissions (DBA) as the number
of days from the discharge date of the first admission to the start of the second ad-
mission. The smaller DBA means faster readmission, which indicates worse health-
care outcomes. Figure 1 shows the distribution of DBA, which is a right-skewed
distribution, with mostly low values and few high values.

As a common practice to deal with right-skewed distribution, we take the natural
log of the (DBA+1), so that it will be transformed to a normal-like distribution
(Figure 2), which makes a later analysis more convenient. The reason the log
transform is on DBA+1 instead of DBA is that some of the original DBA are 0,
which does not allow to take the natural log.

To explore the relationship between each predictor and the target variable, we
did a boxplot for each predictor split by its levels. One example is given in Figure
3. For this predictor (marital status), there is a higher DBA for marital status code
= P (Domestic partner), lower for code W (Widow). This makes sense because
research shows being widowhood undermines people’s health [6]. The larger dif-
ference among the levels implies the strong predictive power this predictor has in
terms of predicting the target variable. Other variables with different target values
within the levels are:

• **HOSPITALSERVICE.CODE**: low target value for NP, IVT, WND, WWC.
• **ADMITSOURCE**: high target value for **ADMITSOURCE**=9, 1, 6. Low for 2, 4, 5.
• **DISCHARGESTATUS**: low target value for **DISCHARGESTATUS**=9
• **PATIENTSEX.CODE**: high target value for females.
• **PATIENT.RACE.CODE**: high target value for **PATIENT.RACE.CODE=**T, low for D.

To further explore the relationships between the predictors and the target, we computed the statistics of the target variables in each level of the predictors. One example is given in Table 1.
Figure 3. Box plot of patient marital status VS log(DBA+1).

By looking at the mean and median of the log(DBA+1) within each level of the predictors, some additional differences are observed:

- **Age**: high for age 0-10, low for 100+. The target value decreases with age increases.
- **Patient.Days**: low for Patient.Days 20-30.

Based on the similarity the target behaves in two levels, we can combine the levels, to reduce the levels within a predictor so that the predictive power of the model can be increased. The marital status D, U, W, X are combined to one level renamed as Unmarried; status P (Domestic partner), S (Single) are combined to another level called Single; M stays as the other level called Married. The reason D, W, X can be combined is the target has a similar mean and median on these three levels. U has very little data, we decide to group them to Unmarried also, because of their low mean and median of the target value. It’s clearer after levels are reduced to 3: ranging from low, middle to high DBA. For single people, they are usually young people thus with better health than senior people, it is reasonable

| PATIENT.MARITAL.STATUS | mean   | median  | n    |
|------------------------|--------|---------|------|
| D                      | 4.06432| 4.20469 | 8122 |
| M                      | 4.31861| 4.48864 | 26568|
| P                      | 4.84009| 5.01728 | 21   |
| S                      | 4.42558| 4.6783  | 22164|
| U                      | 3.41953| 2.83321 | 120  |
| W                      | 4.05355| 4.11087 | 7196 |
| X                      | 4.05785| 4.07754 | 1492 |
for them to have low readmission rates. A person who has a partner can get more
care than others, which also means a low readmission rate.

The following other combinations are made in other predictors:

1. **Patient Days** less than or equal to 19 are combined as one level called 0-19. 
   **Patient Days** from 20 to 30 are grouped to the level called 20-30. The others 
   are grouped to 30+.

2. **Patient_Race_Code**: W and X are combined to a new group W-X because 
   they similar relationship to target. B stays as one level B. All other codes are 
   grouped as Other because the data counts are small.

3. **Hospital_Service_Code**: WND, WWC, PAT, SO, NP, IVT are combined as 
   Level1; DX, EOB, MED, OBS, OPB, OPV as Level2; ER as Level3; CTH, 
   NB, OPS as Level4. The mean and median of the target increases from Level1 
   to Level4.

4. **Discharge_Status**: Code 2, 5, 43, 62, 63, 65, 70, 72 are combined as D_level1; 
   Code 1, 7, 30 are combined as D_level2; others are combined to D_level3.

The predictors like “Doctor.Number” have 480 levels, which means there are 480 
different doctors in this dataset. It is not feasible to combine the groups manually. 
Therefore, we apply the clustering method to reduce its levels. The idea is the levels 
with a similar mean value of the target variable would be merged. The clustering 
algorithm used is Ward’s method [21]:

\[
TD_{C_1 \cup C_2} = \sum_{x \in C_1 \cup C_2} D(x, \mu_{C_1 \cup C_2})^2
\]

Where \(x\) are the means of the target in every 480 levels of the predictor Doctor.Number. 
\(\sum_{x \in C_1 \cup C_2} D(x, \mu_{C_1 \cup C_2})^2\) is the total distance of all the points in data to the center 
of the merged cluster of \(C_1\) and \(C_2\). The distance between two clusters is measured 
by \(TD_{C_1 \cup C_2}\) [1]. Figure 4 is the result of the clustering.

There are 16 predictors totally, while some of them are correlated. The principal 
component analysis (PCA) [24] can be used to reduce the dimensionality of a data 
set that contains multiple variables correlated to each other. We run PCA based on 
several variables that could have correlations: for instance, Surgeon and Doctor. 
The number could be related to, PATIENT.DRG and HOSPITALSERVICE.CODE and 
IC99.PROCEDURE.CODE and ICD9_DIAGNOST_CODE could be related.

Table 2 shows the largest loadings in the first principal component (PC1) of 3 
predictors. We can use these loadings to generate a new feature to replace the 3 predictors 
“HOSPITALSERVICE.CODE”, “PATIENT.DRG”, “IC99_DIAGNOST_CODE”. This 
new feature is the linear combination of the levels from table 5, and the coefficients 
are the loadings.

3. **Generalized linear model (GLM) with feature selection using AIC and BIC**. In regression, when the influence of the independent variable on the 
dependent variable changes according to the value of one or more other independent 
variables, the interactions exist [24]. When we say two predictors interact with 
each other, that means the change of value in one predictor alters how the other 
predictor affects the target. In GLM, the interaction between predictor \(x_1\) and \(x_2\) 
can be expressed as:

\[
\log(y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_{12}x_1x_2
\]
Figure 4. The cluster dendrogram of Doctor.Number. Each vertical line at Height 0 represents one doctor. The doctors inside the same red box are clustered as one.

| Variable Name                  | PC1         |
|--------------------------------|-------------|
| ICD.PROCEDURE.CODE1            | -0.38498161|
| HOSPITALSERVICE.CODEG          | -0.38284437|
| ICD9_DIAGNOST_CODE3            | -0.29296143|
| PATIENT.DDRG1                  | -0.19955252|
| ICD9_DIAGNOST_CODE2            | 0.38297941  |
| ICD.PROCEDURE.CODE5            | 0.38210699  |
| HOSPITALSERVICE.CODEB          | 0.36536618  |
| PATIENT.DRG4                   | 0.24330584  |

Table 2. The largest loadings of related predictors in the first principal component (PC1).

We investigate the interactions of IO_CODE vs PATIENT.SEX_CODE, and IO_CODE vs Age. IO_CODE stands for inpatient/outpatient code, where code I is inpatient, O is outpatient. From Figure 5, we can see the interaction exists between IO_CODE and PATIENT.SEX_CODE. For IO_CODE=I, the target values are significantly higher in females than in males, while for IO_CODE=O, the target values are almost the same in both genders. The reason could be that inpatient service has a better outcome for females than males, while outpatient service has the same outcome for both genders. Or it could be female inpatients who involve birth deliveries, who stayed in the hospital not because they are sick but to give birth etc. Figure 6 shows the interaction between the IO_CODE and Age. We can see the DBA decreases with the age increases for outpatients, while DBA stays the same or even slightly increasing for inpatients age greater than 40 years old. This indicates investing more resources...
on inpatients age from 40 to 90 could be effective in reducing the readmission rate (increasing the DBA) because their DBAs don’t behave in the pattern it’s supposed to be.

Figure 5. Box plot of PATIENT.SEX.CODE VS log(DBA+1) split by inpatient (I) and outpatient (O).

Figure 6. Box plot of Age vs log(DBA+1) split by inpatient (I) and outpatient (O).
We split the data into 75% training data and 25% testing data. The average values of the target variable in training and testing data are 207.233 and 207.438. This shows the training and testing data have very similar distributions by using the build-in stratification function in R. Before running the GLM on training data, we first run the ordinary least squares regression (OLS) model to get the baseline model as a benchmark. Its Akaike information criterion (AIC) is 692151 and the root means square error (RMSE) is 275.102. AIC is defined as $-2l + 2p$, where $l$ is the loglikelihood of the model on the training data, $p$ is the number of parameters (or features) in the model. Using AIC as the metric to evaluate overfitting, the model with a smaller AIC is better. $2p$ acts as the penalty term, to prevent the model from adding too many features.

Two combinations of distribution and link function of GLM are considered: Gamma and Inverse Gaussian distributions, both with log link function. Because the target variable is right-skewed, we choose the Gamma and Inverse Gaussian distributions, both of which are right-skewed. The reason we choose the log link is that it ensures the non-negative output of the prediction, and the target variable DBA is non-negative. The GLM models also include the interactions selected above: the interaction between IO\_CODE and PATIENT.SEX\_CODE, IO\_CODE, and Age.

The AIC for GLM Gamma distribution with log link is 607627 and RMSE is 271.102. The GLM Inverse Gaussian distribution does not converge when running it in R. The Gaussian distribution with log link can also produce non-negative output. Running it, we get AIC=692012, RMSE=274.642. Given two models have the same number of predictors, the model with lower AIC indicates the larger value of loglikelihood and thus better fits the training data. GLM with Gamma distribution and log link has lower AIC than the OLS model and Gaussian GLM, we select it as the model to be used for further analysis.

Features are selected using forward or backward stepwise selection method using AIC or BIC (Bayesian information criterion) as the information criteria. BIC is defined as $-2l + \log(n)p$, where $n$ is the number of rows in training data. From the formulas of AIC and BIC, we know BIC is stricter than AIC, thus BIC is more likely to build a model with fewer features. Since our goal is to select the most important features to be contained in the final model, we prefer BIC to AIC as the criteria. In the forward stepwise selection, we start with the empty model where no feature is selected, then gradually add features one by one. If adding one feature reduces the information criteria measured by AIC or BIC, then it will be added. The stepwise selection stops when the information criteria no longer decrease. It is the opposite for backward selection, where it starts from the full model with all the features included, then deleting features one after another until the BIC doesn’t decrease. It is more likely the forward selection produces fewer features than backward selection. We use forward selection to do feature selection since we prefer a model with fewer parameters.

Running the forward selection with BIC, the following features (or variables) are selected:

- DOCTOR\_NUMBER
- ServCode\_DRG\_ICD
- Age
- Surgeon
Note that the interactions are not selected as features by the BIC to be included in the later model. The interaction between \texttt{IO\_CODE} and \texttt{PATIENT.SEX\_CODE} is not selected could be due to the unimportance of patient gender as a predictor. The interaction between \texttt{IO\_CODE} and \texttt{Age} is not selected could be because both predictors have been selected, and their interaction doesn’t add much new information. But the exploration of the interactions at the beginning of this session is still worthwhile to be conducted because it provides us new perspectives to understand and interpret the data.

When running the GLM model using the features selected above, we get the following output with the p-value of each predictor listed in the last column. Notice that majority of the features have small p-values marked with star(s) except few features (or levels). This indicates that the most features selected are significant. 6 levels are not significant. For example, \texttt{DOCTOR.NUMBER=5} has a p-value=0.13795. To deal with this issue, we combine these levels to their base levels.

| Coefficients | Estimate | Std. Error | t value | p-value | Significant Code |
|--------------|----------|------------|---------|---------|------------------|
| (Intercept)  | 5.99381  | 0.1633     | 36.7    | < 2e-16 | ***              |
| DOCTOR.NUMBER2 | -0.60542 | 0.06083    | -8.67   | < 2e-16 | ***              |
| DOCTOR.NUMBER3 | -0.1012  | 0.04012    | -2.52   | 0.01166 | *                |
| DOCTOR.NUMBER4 | 0.31089  | 0.03528    | 8.81    | < 2e-16 | ***              |
| DOCTOR.NUMBER5 | 0.52643  | 0.35486    | 1.48    | 0.13795 |                  |
| ServCode\_DRG\_ICD | -0.20195 | 0.00903    | -33.48  | < 2e-16 | ***              |
| Age10-20    | -0.26285 | 0.05948    | -6.76   | < 2e-16 | ***              |
| Age100+     | -1.86377 | 0.39675    | -4.7    | 2.60e-06| ***              |
| Age20-30    | -0.61615 | 0.05559    | -11.08  | < 2e-16 | ***              |
| Age30-40    | -0.66055 | 0.05605    | -11.78  | < 2e-16 | ***              |
| Age40-50    | -0.63653 | 0.05632    | -11.13  | < 2e-16 | ***              |
| Age45-60    | -0.63357 | 0.05637    | -11.24  | < 2e-16 | ***              |
| Age50-70    | -0.63919 | 0.05726    | -11.01  | < 2e-16 | ***              |
| Age70-80    | -0.60506 | 0.05855    | -10.33  | < 2e-16 | ***              |
| Age80-90    | -0.62985 | 0.06173    | -10.2   | < 2e-16 | ***              |
| Age90-100   | -0.6078  | 0.08093    | -7.51   | 6.00E-14| ***              |
| Surgeon2    | -0.27224 | 0.02279    | -12.08  | < 2e-16 | ***              |
| Surgeon3    | -0.18455 | 0.06706    | -2.75   | 0.00592 | *                |
| Surgeon4    | -0.36174 | 0.04333    | -8.35   | < 2e-16 | ***              |
| Surgeon5    | 0.53341  | 0.34909    | 1.53    | 0.12652 |                  |
| PATIENT.MARITAL\_STATUSM | 0.16705  | 0.01716    | 9.73    | < 2e-16 | ***              |
| PATIENT.MARITAL\_STATUSPS | 0.05911  | 0.02037    | 2.9     | 0.03711 | **               |
| NurStat\_CCU\_ER\_PC | 0.02249  | 0.03593    | 0.63    | 0.53137 |                  |
| NurStat\_WS | 0.2277   | 0.06175    | 3.69    | 0.0023  | ***              |
| DISCHARGE\_STATUSUR | 0.39629  | 0.0448     | 8.85    | < 2e-16 | ***              |
| DISCHARGE\_STATUSUS | 0.43827  | 0.05653    | 7.75    | 9.20e-15| ***              |
| income\_levellow | -0.11351 | 0.02611    | -4.35   | 1.40E-05| ***              |
| income\_levelmedium | -0.09598 | 0.01506    | -6.37   | 1.80E-10| ***              |
| Patient\_Days20-30 | -0.28387 | 0.04367    | -6.5    | 8.10E-11| ***              |
| Patient\_Days30+ | -0.48626 | 0.32127    | -1.51   | 0.13035 |                  |
| PATIENT\_RACE\_CODEBO | -0.18993 | 0.14455    | -1.31   | 0.18887 |                  |
| PATIENT\_RACE\_CODEDH | -0.99651 | 0.20834    | -4.78   | 1.70E-06| ***              |
| PATIENT\_RACE\_CODEWX | -0.11298 | 0.14369    | -0.79   | 0.43171 |                  |
| IO\_CODEO | -0.18799 | 0.04251    | -4.42   | 9.80E-06| ***              |

Significant codes: 0.001 ‘***’ 0.01 ‘**’ 0.05 ‘*’ 0.1 ‘.’ 1

Table 3: The results of GLM on training data generated by R.
Next, we validate this GLM model compared with the OLS model. Running both models on the same preprocessed data, we get the AIC of 607727 for the GLM model and 692151 for the OLS model. Thus, GLM has significantly lower AIC despite having fewer features, suggesting the GLM is a better model than OLS.

We obtained the diagnosis plots of two models. From the plot of Residual vs Fitted in Figure 7, the points for GLM are distributed more symmetrically against its mean and centered near zero than the points of OLS in the right-side figure. Therefore, the assumption of constant variance and zero means of residual is more valid for GLM than OLS.

The Q-Q plot of GLM in Figure 8 shows the normal assumption for residuals holds for most values, except for values near extreme. It suggests a right-skewed distribution like gamma or lognormal would do better for residuals distribution. The Q-Q plot on the right side of Figure 8 shows the normal assumption of residual is not proper for the OLS model.

We run the above trained GLM model on all data (train and test data combined), the coefficients of the regression model are listed in Table 4 below.
| Feature               | Coefficients (β) | exp(β) - 1 | Interpretation                                                                 |
|-----------------------|------------------|------------|--------------------------------------------------------------------------------|
| DOCTOR.GROUP=2        | -0.58933         | 0.45       | 45% decrease in DBA compared to the base group DOCTOR.GROUP=1. This means the group 2 of doctors are less effective in terms of improving the DBA (or reducing 30-day readmission) than group 1. |
| DOCTOR.GROUP=3        | -0.11050         | 0.10       | 10% decrease in DBA compared to the group DOCTOR.GROUP=1.                         |
| DOCTOR.GROUP=4        | 0.28857          | 0.33       | 33% decrease in DBA compared to the group DOCTOR.GROUP=1. Therefore group 4 of doctors are more effective in improving the DBA than group 1. |
| ServCode_DRG_ICD      | -0.20100         | 0.18       | 18% decrease in DBA for every 1.0 increase in the feature ServCode_DRG_ICD, which is the new artificial feature made using the PCA from the predictors PATIENT.DRG, HOSPITAL.SERVICE.CODE, ICD.PROCEDURE.CODE, ICD9_DIAGNOST_CODE |
| Age=10-20             | -0.30656         | 0.26       | 26% decrease in DBA compared to Age 0-10. This makes sense because younger people are less likely to be readmitted. |

Since the log link was took on the target variable, to explain the effect of the coefficients, we should take the exponent of the coefficients and minus 1. The following table is the interpretation of some coefficients:
From the results and interpretation in Table 5 and Table 4, we can see the Age variables have obvious larger coefficients than other predictors. Therefore, Age is the most important predictor to predict the hospital readmission. Knowing this fact can help the hospitals to better know the readmission risks associated with Age. They can provide more resources to reduce the readmissions of senior patients. For instance, by following up more frequently with the older patients to remind them to take pills in time after they are discharged. Another example is instead of discharging the older patients to home, the hospitals can discharge them to inpatient rehabilitation facilities (IRF) where the patients can receive rehabilitation care that
reduce the readmissions. Or simply charging more from the health insurance companies for senior patients’ admissions because of the higher readmission risks, though the moral risks of doing so would need to be considered also.

We also tried the LASSO regression, an algorithm that can select the features by itself, and used cross-validation to reduce the overfitting. An RMSE of 274.9847 is obtained when running the LASSO. It retained more predictors than the stepwise AIC methods. The RMSE for the GLM model developed is 275.6046. The LASSO produced a more complicated model but doesn’t significantly reduce the RSME, thus we still recommend the GLM method using the features selected from the stepwise method.

4. Summary. To summarize the paper, we investigated the relationships of various predictors to the target variable: days between the admissions (DBA). Data visualization was applied to explore the potential relationship between the predictors and the target. We preprocessed the data by deleting the variables we believe there is no predictive power, combined the levels within the predictors to improve their predicting power. Then we compared the OLS model, GLM with different distribution and link functions. The interaction between the predictors is also included in the model. We discussed the cons and pros of regularization methods on this problem. The features input for the regressions is selected using forward selection with BIC. The GLM outperforms the OLS model in terms of RMSE and AIC. The regression diagnostics validated the GLM we built. According to this GLM, we found Age is the most important factor to predict the DBA. Other predictors such as the doctor, surgeon, patient marital status, and inpatient days also have a significant effect on DBA.

Hospitals can use our study to predict which patient discharge can lead to 30-day readmission so that they can better allocate the resources on these patients to reduce the rate of 30-day readmission. Notice the interactions identified between IO_CODE and PATIENT.SEX.CODE, IO_CODE, and Age are not selected by the BIC as the features to be included in the final GLM model. If there is more data, it would be interesting to further investigate these interactions. We would also be interested in implementing a loss-gain analysis about each wrong or correct prediction if the information of the loss-gain associate with each type of prediction is available.

REFERENCES

[1] J. Benuzillo, W. Caine, R. S. Evans, C. Reborts, D. Lappe and J. Doty, Predicting readmission risk shortly after admission for CABG surgery, Journal of Cardiac Surgery, 33 (2018), 163–170.

[2] C. Boccuti and G. Casillas, Aiming for fewer hospital u-turns: The medicare hospital readmission reduction program, Henry J. Kaiser Family Foundation, (2017). https://collections.nlm.nih.gov/catalog/nlm:nlmuid-101707559-pdf.

[3] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm and N. Elhadad, Intelligible models for healthCare: Predicting pneumonia risk and hospital 30-day readmission, KDD, 10 (2015), 1721–1730.
[4] S. Cui, D. Wang, Y. Wang, P. Yu and Y. Jin, An improved support vector machine-based diabetic readmission prediction, Computer Methods and Programs in Biomedicine, 166 (2018), 123–135.

[5] J. A. Dodson, A. M. Hajduk, T. E. Murphy, M. Geda, H. M. Krumholz, S. Tsang, M. G. Nanna, M. E. Tinetti, D. Goldstein, D. E. Forman, K. P. Alexander, T. M. Gill and S. I. Chaudhry, Thirty-day readmission risk model for older adults hospitalized with acute myocardial infarction, Circulation: Cardiovascular Quality and Outcomes, 12 (2019).

[6] M. Fionn and P. Legendre, Ward’s hierarchical agglomerative clustering method: Which algorithms implement ward’s criterion?, J. Classification, 31 (2014), 274–295.

[7] Y. Gai and D. Pachamanova, Impact of the Medicare hospital readmissions reduction program on vulnerable populations, BMC Health Services Research, 19 (2019), 1–15.

[8] R. Gardner, Q. Li, R. R. Baier, K. Butterfield, E. A. Coleman and S. Gravenstein, Is implementation of the care transitions intervention associated with cost avoidance after hospital discharge?, Journal of General Internal Medicine, 29 (2014), 878–884.

[9] S. B. Golas, T. Shibahara, S. Agboola, H. Otaki, J. Sato, T. Nakae, T. Hisamitsu, G. Kojima, J. Felsted, S. Kakarmath, J. Kvedar and K. Jethwani, A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: A retrospective analysis of electronic medical records data, BMC Medical Informatics and Decision Making, 18 (2018), 44.

[10] A. L. Hines, M. L. Barrett, H. J. Jiang and C. A. Steiner, Conditions with the largest number of adult hospital readmissions by payer, Healthcare Cost and Utilization Project (HCUP) Statistical Briefs, Statistical brief 172, 2011.

[11] Q. L. Huynh, K. Negishi, L. Blizzard, K. Sanderson, A. J. Venn and T. H. Marwick, Risk factors for 30-day readmissions after acute myocardial infarction, International Cardiovascular Forum Journal, 4 (2015), 30–36.

[12] Q. L. Huynh, K. Negishi, L. Blizzard, K. Sanderson, A. J. Venn and T. H. Marwick, Predictive score for 30-day readmission or death in heart failure, JAMA Cardiol, 1 (2016), 362–364.

[13] S. F. Jencks, M. V. Williams and E. A. Coleman, Rehospitalizations among patients in the Medicare fee-for-service program, New England J. Medicine, 360 (2009), 1418–1428.

[14] D. Kansagara, H. Englander, A. Salanitro, D. Kagen, C. Theobald, M. Freeman and S. Kripalani, Risk prediction models for hospital readmission: A systematic review, JAMA, 306 (2011), 1688–1698.

[15] H. M. Krumholz, A. R. Merrill, E. M. Schone, G. C. Schreiner, J. Chen, E. H. Bradley, Y. Wang, Y. Wang, Z. Lin, B. M. Straube, M. T. Rapp, S. T. Normand and E. E. Drye, Patterns of hospital performance in acute myocardial infarction and heart failure 30-day mortality and readmission, Circulation: Cardiovascular Quality and Outcomes, 2 (2009), 407–413.

[16] H. M. Krumholz, A. R. Merrill, E. M. Schone, G. C. Schreiner, J. Chen, E. H. Bradley, Y. Wang, Y. Wang, Z. Lin, B. M. Straube, M. T. Rapp, S. T. Normand and E. E. Drye, Hospital 30-day pneumonia readmission measure: Methodology, Centers for Medicare & Medicaid Services, https://qualitynet.cms.gov/inpatient/measures/readmission/methodology.

[17] S. Lee, S. Wang, P. A. Bain, C. Baker, T. Kundinger, C. Sommers and J. Li, Reducing COPD readmissions: A causal bayesian network model, IEEE Robotics and Automation Letters, 3 (2018), 4046–4053.
[18] X. Liu, Y. Chen, J. Bae, H. Li, J. Johnston and T. Sanger, Predicting heart failure readmission from clinical notes using deep learning, *2019 IEEE International Conference on Bioinformatics and Biomedicine*, (2019), 2642–2648.

[19] J. F. Mather, G. J. Fortunato, J. L. Ash, M. J. Davis and A. Kumar, Prediction of pneumonia 30-day readmissions: A single-center attempt to increase model performance, *Respiratory Care*, **59** (2014), 199–208.

[20] C. K. McIlvennan, Z. J. Eapen and L. A. Allen, Hospital readmissions reduction program, *Circulation*, **131** (2015), 1796–1803.

[21] J. E. Simmering, L. A. Polgreen, A. P. Comellas, J. E. Cavanaugh and P. M. Polgreen, Identifying patients with COPD at high risk of readmission, *Chronic Obstructive Pulmonary Diseases*, **3** (2016), 729.

[22] M. Soltani, R. Batt and H. Bavafa, Quality improvement spillovers: Evidence from the hospital readmissions reduction program, *SSRN*, [https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3132770](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3132770).

[23] L. Turgeman and J. H. May, A mixed-ensemble model for hospital readmission, *Artificial Intelligence in Medicine*, **72** (2016), 72–82.

[24] K. Williams, The transition to widowhood and the social regulation of health: Consequences for health and health risk behavior, *The Journals of Gerontology: Series B*, **59** (2004), 343–349.

[25] K. Yu and X. Xie, Predicting hospital readmission: A joint ensemble-learning model, *IEEE Journal of Biomedical and Health Informatics*, **24** (2020), 447–456.

[26] Hospital Readmissions Reduction Program (HRRP), Reported of the U. S. Centers for Medicare & Medicaid Services, 2020. Available from: [https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html](https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html).

[27] Cost of Hospital Readmissions, Report of MBA Medical, 2015. Available from: [http://mbamedical.com/cost-hospital-readmissions/](http://mbamedical.com/cost-hospital-readmissions/).

[28] Performance of the Massachusetts Health Care System Series: A Focus on Provider Quality, Agency for Healthcare Research and Quality (US), 2015. Available from: [https://archives.lib.state.ma.us/bitstream/handle/2452/265367/ocn913253238-report.pdf](https://archives.lib.state.ma.us/bitstream/handle/2452/265367/ocn913253238-report.pdf).

[29] Analysis: Hospital Readmissions of all Ages, Insurance Types Identifies High Risk Groups, Report of Beth Israel Deaconess Medical Center, 2017. Available from: [https://medicalxpress.com/news/2017-07-analysis-hospital-readmissions-ages-high.html](https://medicalxpress.com/news/2017-07-analysis-hospital-readmissions-ages-high.html).

[30] *Medicare Could Be Insolvent In 2024: How To Prevent It*, Report of Forbes, 2021. Available from: [https://www.forbes.com/sites/nextavenue/2021/03/05/medicare-could-be-insolvent-in-2024-how-to-prevent-it/?sh=99cf1ab26f00](https://www.forbes.com/sites/nextavenue/2021/03/05/medicare-could-be-insolvent-in-2024-how-to-prevent-it/?sh=99cf1ab26f00).

[31] Readmissions Reduction Program, Report of Centers for Medicare and Medicaid Services, 2018. Available from: [http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html](http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html).

[32] Accuracy of Readmission Risk Assessment Improved by Machine Learning, Health Catalyst, 2018. Available from: [https://www.healthcatalyst.com/wp-content/uploads/2021/05/Accuracy-of-Readmission-Risk-Assessment-Improved-by-Machine-Learning-1.pdf](https://www.healthcatalyst.com/wp-content/uploads/2021/05/Accuracy-of-Readmission-Risk-Assessment-Improved-by-Machine-Learning-1.pdf).
Appendix A. Data dictionary.

| Variable Name         | Definition                                      | Date type and values |
|-----------------------|-------------------------------------------------|----------------------|
| PATIENT.DRG Patient   | Patient diagnosis-related group                 | Integer 0-999        |
| NurStat               | Nurses Station                                  | Letters code representing the type of nurses station, with majority values missing. |
| DOCTOR NUMBER         | ID of the doctor                                | Integer              |
| Surgeon               | ID of the surgeon                               | Integer              |
| I0_CODE               | Inpatient or outpatient                         | 1: inpatient         |
| HOSPITAL SERVICE CODE | A code representing the type of healthcare service | Letters code. No missing value. |
| ADMIT SOURCE          | The code indicating the source of the referral for the admission or visit. | 1: Physician Referral |
|                       |                                                  | 2: Clinic Referral   |
|                       |                                                  | 3: HMO Referral      |
|                       |                                                  | 4: Transfer from a Hospital                          |
|                       |                                                  | 6: Transfer from Another Health Care Facility        |
|                       |                                                  | 8: Court/Law Enforcement                              |
|                       |                                                  | 9: Information Not Available                          |
| DISCHARGE STATUS      | Patient discharge status                        | Integer code         |
| PATIENT SEX CODE      | A code indicating the sex of the patient.       | F: Female          |
|                       |                                                  | M: Male            |
|                       |                                                  | U: Unknown         |
| PATIENT MARITAL STATUS| Marital status                                  | D: divorced        |
|                       |                                                  | S: single          |
|                       |                                                  | M: married         |
|                       |                                                  | W: widowed         |
|                       |                                                  | U: unknown         |
|                       |                                                  | P: partnered       |
|                       |                                                  | X: legally Separated |
| PATIENT RACE CODE     | Code indicating the racial or ethnic background of a person. | A: Asian or Pacific Islander |
|                       |                                                  | B: Black           |
|                       |                                                  | D: Subcontinent Asian American                         |
|                       |                                                  | H: Hispanic       |
|                       |                                                  | I: American Indian or Alaskan Native                   |
|                       |                                                  | N: Black(Non-Hispanic)                                 |
|                       |                                                  | O: White(Non-Hispanic)                                 |
|                       |                                                  | W: widowed        |
|                       |                                                  | X: legally separated                                    |
| ICD PROCEDURE CODE    | ICD-10 Procedure Coding                         | Integer code        |
| ICD9 DIAGNOSIS CODE   | ICD-9-CM Diagnosis Codes                        | Integer code        |
| PATIENT ZIP           | Patient zip code                                | Integer code        |
| DBA                   | days between the admissions                     | Non-negative integer |

Table 6: Data Dictionary

Appendix B. Levels combined for predictors.

| Variable Name | Levels before combined | Levels after combined |
|---------------|------------------------|-----------------------|
| Nur Stat      | isNA                   | isNA                  |
|               | MS                     | WS                    |
|               | CCU                    | WS                    |
|               | ER                     | WS                    |
|               | PC                     | WS                    |
|               | 0-19                   | 0-19 or 30+           |
|               | 30+                    | 20-30                 |
| HOSPITAL SERVICE CODE | OPS  | Y                  |
|               | CTH                    | R                    |
|               | NB                     | R                    |
|               | MA                     | R                    |
|               | MBD                    | R                    |
|               | OBS                    | R                    |
|               | OPB                    | R                    |
|               | OPV                    | R                    |
|               | EOB                    | R                    |
|               | WND                    | R                    |
|               | IVT                    | B                    |
|               | SO                     | B                    |
|               | WWC                    | B                    |
|               | NP                     | B                    |
|               | ER                     | B                    |
Table 7: Levels combined for predictors

| ADMIT SOURCE |   |   |
|--------------|---|---|
| 9            | 9 |
| 6            |   |
| 8            | 8 |
| 4            | 4 |
| 2            | 25|

| PATIENT SEX CODE |   |
|------------------|---|
| M                | M |
| U                |   |
| F                | F |

| PATIENT MARITAL STATUS |   |
|------------------------|---|
| M                      | M |
| P                      | PS |
| U                      | DUWXNA |
| W                      |   |
| X                      |   |
| NA                     |   |

| PATIENT RACE CODE |   |
|-------------------|---|
| W                 | WX |
| X                 |   |
| A                 |   |
| I                 |   |
| M                 | AIMNT |
| N                 |   |
| T                 |   |
| B                 |   |
| O                 |   |
| D                 | DH |

| DISCHARGE STATUS |   |
|------------------|---|
| 1                |   |
| 7                | R |
| 30               |   |
| 2                |   |
| 5                | Y |
| 43               |   |
| 62               |   |
| 63               |   |
| 65               |   |
| 70               |   |
| 72               |   |
| 3                |   |
| 4                |   |
| 6                |   |
| 9                |   |
| 21               | G |
| 50               |   |
| 51               |   |
| 64               |   |
| 81               |   |
| 82               |   |
| 83               |   |

Table 7: Levels combined for predictors

Appendix C. The mean and median of log(DBA+1) within three levels of Patient.Days.

| Patient.Days | Mean of log(DBA+1) | Median of log(DBA+1) |
|--------------|--------------------|----------------------|
| 0-19         | 4.34939            | 4.54329              |
| 30+          | 3.86552            | 3.98744              |
| 20-30        | 2.59868            | 2.07944              |

Table 8: The mean and median of log(DBA+1) within three levels of Patient.Days.

Received April 2021; revised August 2021; early access December 2021.

E-mail address: lu.xiong@mtsu.edu
E-mail address: ts7f@mtmail.mtsu.edu
E-mail address: rgreen@evigils.com