A Mixed-Effects Logistic Regression Model for Predicting Hospitalization in Dialysis Patients

Samuel Zalewski[1,†]

[1]*M.S. University of Barcelona, Polytechnic University of Catalunya*

Abstract

This study develops a mixed-effects logistic regression model to predict hospitalization among patients undergoing hemodialysis for end-stage renal disease (ESRD), a significant health concern due to its high cost and the burden on healthcare resources. Using a retrospective dataset of 1311 patients with 9547 observations, hospitalization was modeled as a binary outcome, with robust model and component selection processes used to select a best-performing model. This mixed-effects logistic regression model that predicted hospitalization with an AUC of 0.865, confirming its efficacy in classifying high-risk patients. Further, the predictive significance of treatment facilities (p= 0.019) in patient outcomes emphasizes the need for targeted research to explore how facility-related factors influence hospitalization rates.

Corresponding author: Samuel Zalewski *E-mail address:* samzalewski@yahoo.com
Published: Sep 3, 2024

1. Objective and Dataset

In the treatment of end-stage renal disease (ESRD), frequent hospitalization is often a challenge that poses threats for both patients and healthcare providers. In addition to simply disrupting the quality of life of patients, hospitalization is expensive, requires the occupancy of valuable hospital beds, and increases the risk of infection.

In order to investigate causes of hospitalization, a retrospective dataset was organized containing 9547 observations of 1311 patients undergoing dialysis treatment. This dataset included 23 columns of information, a few of them being general patient characteristics such as age or gender, and others being more specific measurements related to their treatment, facility information, or latest bloodwork. The response variable in each observation, the hospitalization in a given month, is presented as an integer counting the total hospitalizations in that time frame. The majority of patients did not experience hospitalization, and one hospitalization was not uncommon, but any more than that was quite rare, with only 18 observations describing the maximum of 3 hospitalizations in a month.

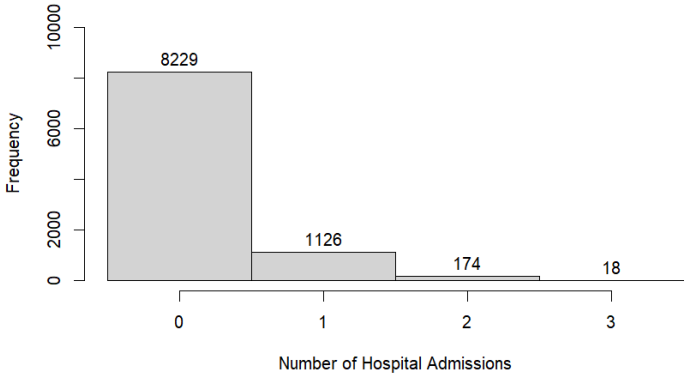

Figure 1. Histogram of Hospital Admissions During a One-Month Period

Because nearly all (∼ 98%) of the hospital admission counts are either 0 or 1, training a model to discriminate between 1 and values larger than 1 would be quite difficult. Further, the marginal benefit of being able to predict that a patient will be hospitalized twice in a month instead of once is conceivably less important than being able to whether they will need to be hospitalized at all. For this reason, hospitalization counts can be simplified to two categories: "Hospitalization not Occurred" for values of 0, and "Hospitalization Occurred" for values of 1 or more. This binary response variable makes this data well-suited for logistic regression, as the logit link function guarantees output values between 0 and 1. In the case of our dataset, this regression model would take the following form:

$$logit(p) = \log\left(\frac{p}{1-p}\right) = X\beta + Zu + \epsilon \tag{1}$$

Where X is the design matrix for known fixed effects multiplied by its effects vector β, Z is the design matrix for random effects multiplied by its vector u, and ϵ is the residual error of the model. As a mixed-effects model, the incorporation of these random effects allows the model intercepts to adjust for repeated measures of patients, resulting in more accurate estimations of model components.

2. Feature Selection and Model Validation

Unfortunately for this dataset, there were several columns with significant amount of missing data. The majority of these NAs came from the final three columns, which described the total amount of Epogen, Heparin, and Venofer, administered in that month, respectively. Because these three variables do not have particularly strong correlations with any other more complete columns, simply deleting these columns could result in loss of uniquely valuable information. Therefore, it was decided to delete all rows with NAs, instead of deleting problematic columns, resulting in a total of 3646 of the original 9547. Although this is a serious reduction in the data, the non-linear relationship between sample size and statistical power are such that any loss of power is worth the additional information maintained by keeping every predictor column. A post-hoc sensitivity analysis was performed regarding this decision to delete rows with NAs, and is discussed further in section 3, but for now the analysis will be continued with the mentioned 3646 observations.

The only downside of now preserving all columns is that the possibility of including too many features and creating an overfitted model must be addressed. To select model components, a forward stepwise procedure was performed to minimize Akaike information criterion (AIC). This procedure begins with an empty model, and then adds additional features one at a time, choosing the one that results in the greatest reduction in AIC. Eventually, the stepwise procedure arrived at a model that optimized AIC by including 11 of the 22 possible predictors. Before diving into the model specifications, a quick model validation should be done to make sure the model is able to perform well on new data.

The 3646 rows of usable data was divided 80/20 into a train and test set, with which 80% of the data was used to train a model using the glmer() function from the lme4 package in R 4.3.2.

Next, the model is to be tested on the remaining 20% of the data

to see how well it can predict hospitalization outcomes. Because logistic regression always returns a decimal between 0 and 1, a common practice is to interpret a value of 0.5 or over as a 1, and anything below 0.5 as a 0. In this context, a 1 would mean that hospitalization is predicted, as opposed to a 0 which would be predicting no hospitalization. Although 0.5 as a threshold is naturally intuitive, it doesn't always yield the best performance. To select a threshold to use as a classification boundary, Youden's Index was used. Simply put, Youden's Index selects the threshold which maximizes the vertical distance between the ROC curve and the diagonal classifier curve (both shown in Figure 2).

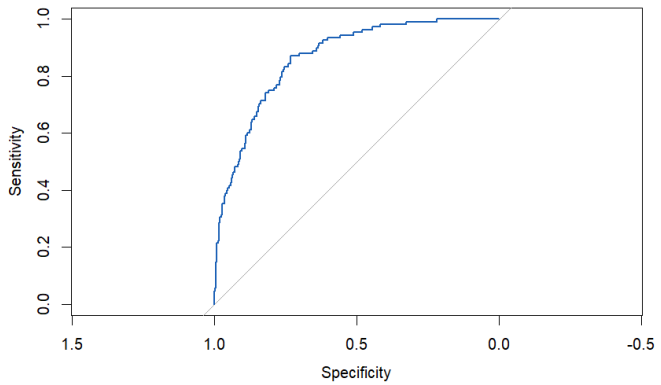

Figure 2. ROC Curve of Hospital Admittance Predictions on the test set. Specificity is the proportion of true negatives, and sensitivity is the proportion of true positives. The curve shows the tradeoff between the two values throughout all possible values of the threshold. From "pROC" for R.

In practice, choosing an optimal threshold depends on the costs associated with each type of error. According to the Youden's Index, the best threshold is 0.102, which results in a sensitivity of 0.87 and a specificity of 0.73. Given that it's more important to correctly predict the outcome of a patient in need of hospitalization than one who doesn't, the higher sensitivity obtained by this threshold is desirable.

This threshold can always be adjusted as seen fit depending on the extent to which true positives are prioritized over true negatives. Figure 3 shows graphically the results of the model's confusion matrix on the test set, demonstrating a strong ability to predict positive hospitalizations, and to a lesser extent, patients not at risk of hospitalization.

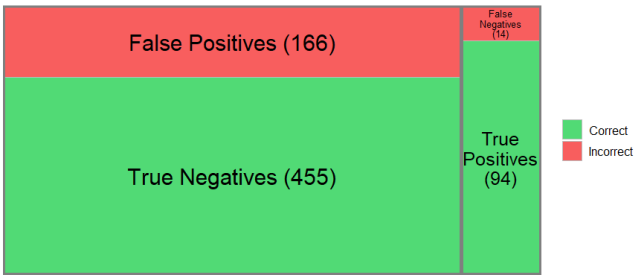

Figure 3. Tree plot of hospital admittance predictions. Positive denotes hospitalization. From "ggplot2"

Finally, it's a good idea to compare this model's performance against a few others before moving forward. In Table 1, four models are compared: a mixed-effects logistic regression model with no feature selection using all 22 columns, a logistic RIDGE regression model with fixed effects only features selected by a Lagrange multiplier, a fixed-effects model with 11 features selected via AIC, and finally, our mixed-effects logistic regression model with 11 the same 11 features. Although some of the models are close, our previously mentioned 11-feature mixed-effects model yields the highest the area under the ROC curve, or AUC (shown in Figure 2).

Table 1. Comparison of Model Performance Metrics. Shown in bold is the mixed-effects logistic regression model from equation (1) with features selected via AIC

Model	Thres.	Sens.	Spec.	AUC
GLMM with 20 factors	0.095	0.861	0.709	0.862
Logistic RIDGE regression	0.129	0.815	0.775	0.853
GLM (11 features, AIC, fixed only)	0.104	0.935	0.663	0.855
GLMM with 11 features (AIC)	**0.102**	**0.870**	**0.733**	**0.865**

3. Model Analysis

With the model validation process complete, we can now look at the individual components of the model and their role in explaining hospitalization rates. Many of the predictors with the highest statistical significance, as it turns out, are not very surprising. HB, representing hemoglobin level in g/Dl, is well known to cause complications in individuals undergoing dialysis treatment from previous studies [2]. The most significant predictor of hospitalization turned out to be TOTAL_MTXS, the number of missed treatments that month. Not far below, the third most significant coefficient was the total number of attended treatments, which, unlike missed treatments, reduced odds of hospitalization. Three components that also had some statistical significance were Alibumin levels, bloodstream infection count, and VINTAGE, or the total days since starting dialysis.

At the bottom of Table 2, there are four coefficients that didn't have much statistical significance, but were still incorporated into the model. These were: a flag indicating excess body weight, age, total Epogen administered, and gender. While some of these variables have correlations with hospitalization, it's not quite certain that there is as strong of a causal relationship between them as with other predictors. For example, Epogen is administered to increase hemoglobin , which has been shown to decrease odds of mortality[1]. However, the model suggests that the higher epogen administration increases odds of hospitalization. Although it likely isn't the epogen itself causing hospitalization, the need for epogen administration itself is a result of one or more other risk factors in a patient.

Table 2. Effects Estimates as Returned by the Summary of glmer's Mixed-Effect Logistic Regression Model

Effect	Estimate	Std. Error	Pr(>|z|)
(Intercept)	4.377×10^{0}	1.000×10^{0}	1.20×10^{-5}
TOTAL_MTXS	2.516×10^{-1}	3.417×10^{-2}	1.82×10^{-13}
HB	-3.839×10^{-1}	6.418×10^{-2}	2.20×10^{-9}
TOTAL_TXS	-1.602×10^{-1}	3.499×10^{-2}	4.70×10^{-6}
FACILITY_ID4154	-9.007×10^{-1}	2.901×10^{-1}	1.91×10^{-3}
FACILITY_ID4426	-1.317×10^{0}	7.357×10^{-1}	0.07347
ALBUMIN	-4.691×10^{-1}	1.698×10^{-1}	5.74×10^{-3}
INFECTIONS	3.022×10^{0}	1.384×10^{0}	2.903×10^{-2}
VINTAGE	1.233×10^{-4}	5.619×10^{-5}	2.820×10^{-2}
PW_FLAG	2.107×10^{-1}	1.520×10^{-1}	0.16576
AGE	9.768×10^{-3}	5.478×10^{-3}	0.07456
EPOPTX	4.360×10^{-5}	2.407×10^{-5}	0.07005
GENDERM	-1.859×10^{-1}	1.515×10^{-1}	0.21986

Despite the inability to derive a strong causal relationship between some of these model components, the inclusion of the last four components in Table 2 proved to be beneficial for model performance. When removing one or more of these components, the performance on the test set, in terms of AUC, always worsens. For this reason,

despite p-values that are theoretically insignificant, all effects from Table 2 will continue to be considered for this model.

Last and most unexpectedly, Facility 4154 was included in the model with p=.00191. According to the model summary, treatment at this facility is associated with a log odds ratio of -0.81, a large decrease in hospitalization. Another facility, facility 4426, also had a large negative log odds ratio, but with much less statistical significance because of its small sample size.

Figure 4 shows an effect plot for Facility ID. For each facility, a prediction and corresponding 95% confidence interval are generated for the chance of a patient experience hospitalization. These estimates are made by changing only the facility, and assume that hemoglobin level, treatment data, and all other model components are the same.

While the 95% confidence intervals of most facilities capture more or less a similar range, that of facility 4154 is nearly entirely below facilties 4057 and 4111. The prediction of 4226 is even lower than 4154, but because of the previously mentioned sample size the confidence interval is much larger. If a more formal hypothesis test were to be conducted on the differences, these confidence intervals could be adjusted for multiple comparisons, increasing the interval along with the number of compared facilities. In the context of this analysis, the low p-values of the facility components, combined with the little confidence overlap in Figure 4, is sufficient evidence to justify a further investigation in the facilities and the impact they might have on hospitalization rates.

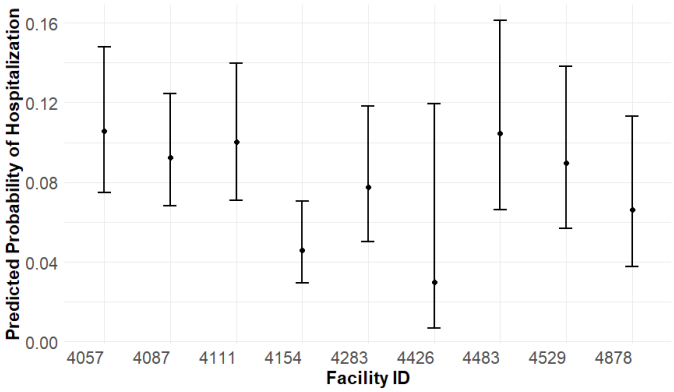

Figure 4. Effect plot of Facility ID versus Hospitalization with 95% CI. Calcualted using "ggpredict" from the "ggeffects" package, and ploted with "ggplot2".

3.1. Sensitivity Analysis

In section 2, it was mentioned that the original dataframe of 9547 observations was reduced down to 3646 due to the presence of NAs. However, after completing the model selection and validation process, it was noticed that many of these NAs belonged to rows of columns that were ultimately not included in any of the models that involved feature selection. Therefore, it is possible to remove these columns and run the model pipeline once again, but this time excluding these unused columns, allowing us to preserve more of the original data. As previously stated, three columns were mainly responsible for the NAs: Epogen, Heparin, and Venofer. While Epogen was a component of the top performing model, Heparin and Venofer were not. By removing these two columns from the dataframe, the total amount of NA-free observations increased form 3646 to 6876, a considerable improvement.

After training the four model types from section 2 on 80% of the 6876 observations, their performance was compared on the remaining test set in the same manner as the models in section 2.

Although there is a difference in the models' AUC with the larger dataset, their ranking amongst each other in Table 3 other mirrors that of Table 1. Mixed effects models prove to outperform the other two, with the model with features selected by AIC performing the best. The consistency of these rankings suggests a robustness of our model selection process, and that the decision to remove NA rows at the beginning of the analysis did not noticeably impact the conclusion that the 11-feature GLMM with features selected via AIC forward selection is the strongest of the compared models.

Table 3. Comparison of Model Performance Metrics as in Table 1, but trained and tested on 6876 rows instead of the previous 3646.

Model	Thres.	Sens.	Spec.	AUC
GLMM with 20 factors	0.145	0.696	0.849	0.838
Logistic RIDGE regression	0.195	0.608	0.886	0.812
GLM (11 features, AIC, fixed only)	0.116	0.779	0.730	0.812
GLMM with 11 features (AIC)	**0.144**	**0.706**	**0.847**	**0.840**

4. Conclusions and Next Steps

Based on the results from the analysis, there are possible two avenues that can be followed to solidify the conclusions of this exploratory research. First, the predictive capability of mixed-effects logistic regression model to predict patient hospitalization has been demonstrated, with an AIC forward selection proving to be much more effective at predicting hospitalization risk than any individual predictor. In practice, a model like this could be used to flag patients that may have significant that has gone unnoticed. From there, a healthcare expert can monitor their specific details and take prophylactic action, knowing that they are at possible risk of hospitalization.

Second, the model summary showed a strong indication that the treatment facility impacts the hospitalization risk of patients. From this, a further investigation should be carried out to explore possible root causes of this discrepancy. Although a demographic comparison of patients at each facility showed no clear patterns that would explain the hospitalization differences, it's possible that there are additional variables not included in the study that are responsible. As a recommendation, more data should be collected from facilities 4154 and 4426 to be compared with other facilities. This data should include not just details of the patients, but of the medical staff at the facility, as information such as their experience level could be a possible cause of outcome discrepancy. Finally, environmental data from the facilities can be explored, as factors such as the cleanliness of a facility can certainly affect treatment success.

This study has displayed the complex relationship of patient characteristics, treatment specifics, and facility-related factors in influencing hospitalization rates among hemodialysis patients. Future research should not only replicate these findings in larger, more comprehensive studies, but should also encapsulate advanced facility details to further catch the nuanced causes of hospitalization risk.

References

[1] J. Fink, S. Blahut, M. Reddy, and P. Light, "Use of erythropoietin before the initiation of dialysis and its impact on mortality," *American Journal of Kidney Diseases*, vol. 37, no. 2, pp. 348–355, 2001. DOI: 10.1053/ajkd.2001.21305.

[2] N. Ofsthun, J. Labrecque, E. Lacson, M. Keen, and J. Lazarus, "The effects of higher hemoglobin levels on mortality and hospitalization in hemodialysis patients," *Kidney International*, vol. 63, no. 5, pp. 1908–1914, 2003. DOI: 10.1046/j.1523-1755.2003.00937.x.