STA 440 Case 2

Annie Lott, Wuming Zhang, William Yang

October 27, 2017

Goals and Data Description

Getting treatment quickly after a stroke is crucial to a positive long-term prognosis. For emergency room patients who may have had a stroke, we are exploring the variables that influence the time it takes for these patients to receive a neurological assessment. The time from getting to the emergency room to receiving a neurological assessment, such as a CT scan, factors into the total time it takes for the patient to get treated for a stroke. This has a direct impact on the stroke patient's subsequent neurological health and survival.

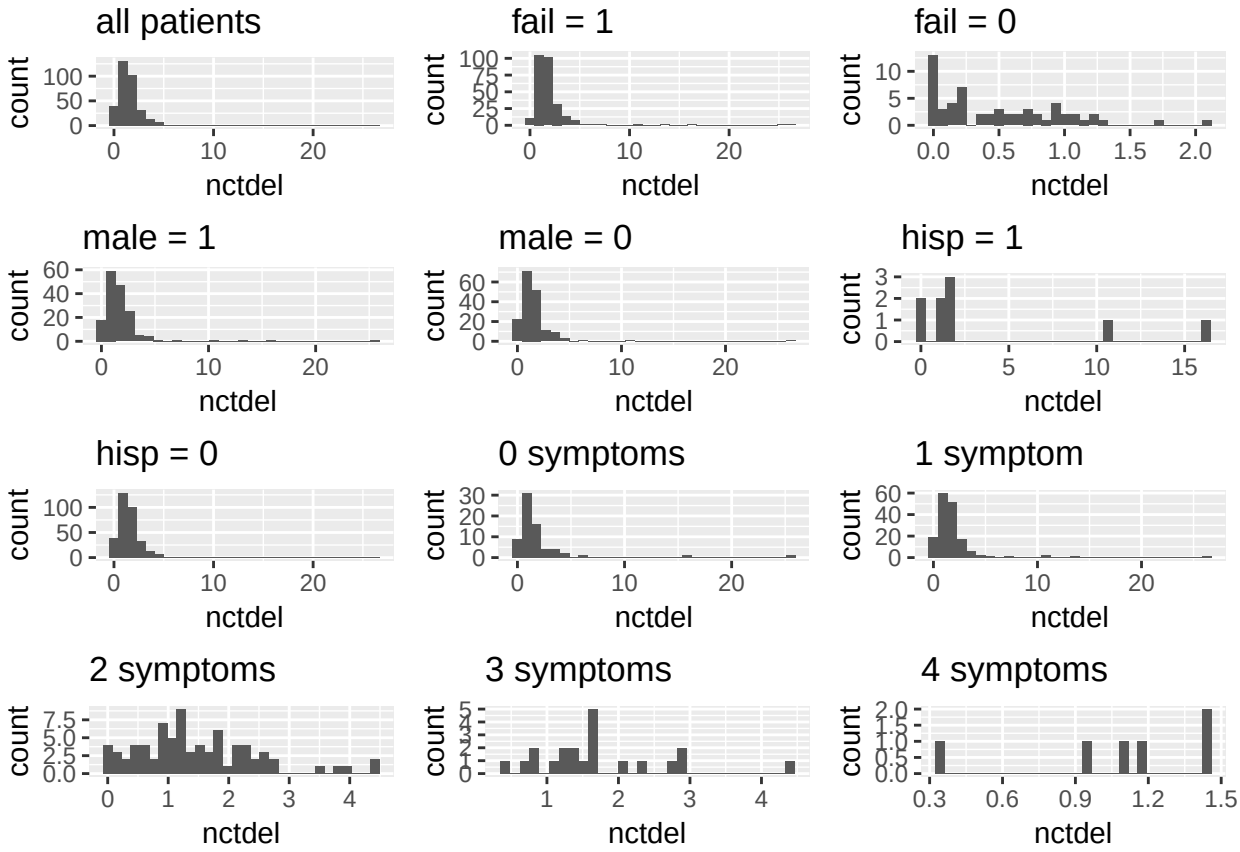
Based on data from 335 emergency room patients with mild to moderate motor impairment, possibly indicative of a stroke, we are analyzing whether sex, race, ethnicity, and the number of symptoms displayed affects the time until the patients receive a neurological assessment, or whether or not they receive a neurological assessment at all. Sex is given as a binary variable of whether the patient is male (1) or female (0), race is a binary variable of whether the patient is black (1) or not (0), ethnicity is a binary variable of whether the patient is hispanic (1) or not (0), and the number of symtoms ranges from zero to four, with binary variables for if the patient exhibited one, two, three, or four symptoms (1 for each number of symptoms if they did exhibit this number, 0 for each number if they didn't). The four possible symptoms include a headache, loss of motor skills or weakness, trouble talking or understanding, and vision problems. Thus, given these variables, our goal is to build a model that predicts the amount of time until neurological assessment for potential stroke patients based on clinical presentation (how many symptoms the patient seems to have), gender and race and ethnicity, and to perform inference on the impact of these variables based on the model.

Data Cleaning

For ease of analysis, we cleaned the data by introducing a categorical variable for the number of symptoms for each patient- zero, one, two, three, or four- rather than using binary variables for whether a particular number of symptoms were displayed or not. We treat this variable as a categorical variable rather than a continuous variable, because we predict that the impact of having 0 symptoms versus 1 symptom in influencing wait time will be different from the effect of having 3 symptoms versus 4 symptoms in determining wait time. There were no missing data to recode for our analysis.

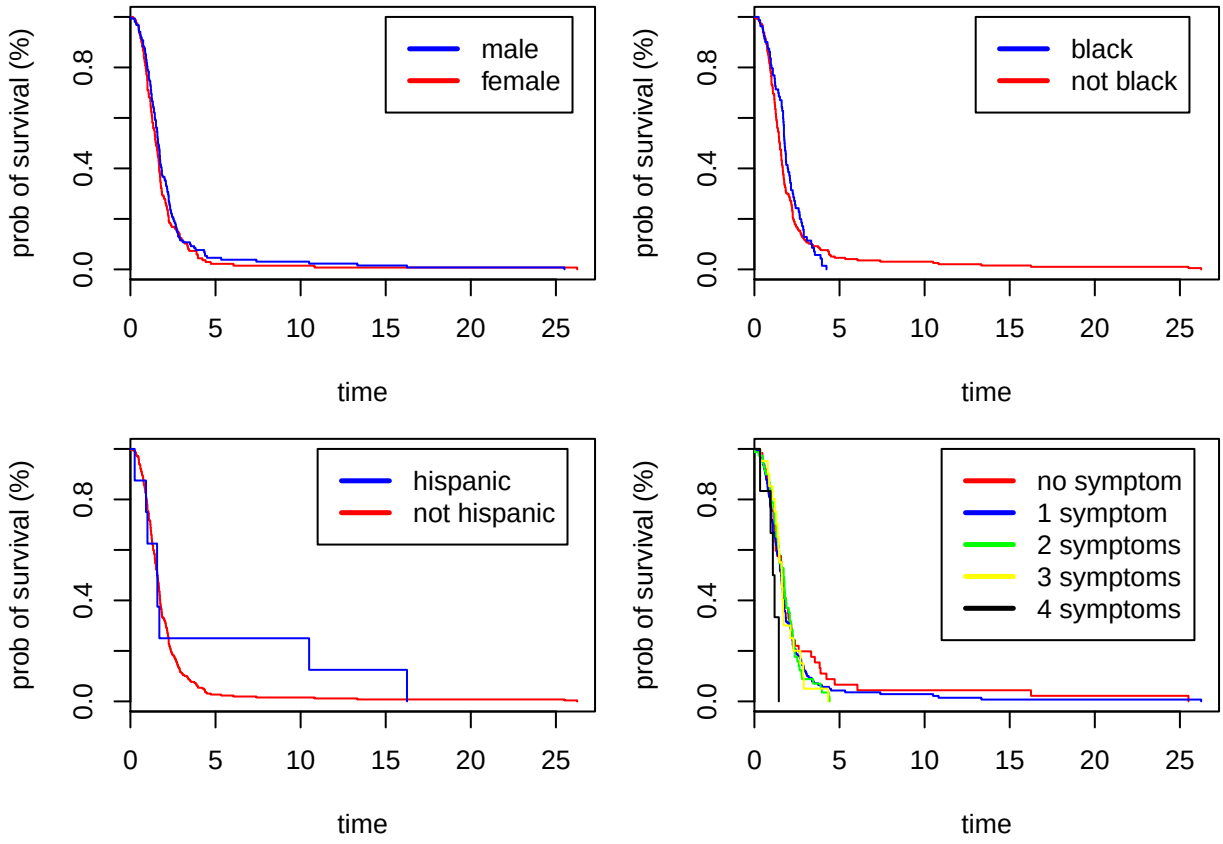
Exploratory Data Analysis

Histograms of each Indicator



To get a better sense of the distributions in the data, we first plot histograms of the nctdel times for each level of all the variables. We notice some general trends here - nctdel time seems to go down with more symptoms, the overall distribution of times for all patients is right-skewed with a mean around 1.5. However, we note that there are only 9 data points for hispanic patients, and only 6 for patients with all 4 symptoms. Inference around these variables will therefore have high variance.

Kaplan-Meier Analysis



Note that in our case, survival means that the patient still keeps waiting. The survival curves for males versus females look almost the same, suggesting that there is no significant difference in their chance of waiting at varied time. The survival curves for black patients versus non-black patients look a little different, with the probability of still waiting dropping off slightly more quickly for non-black patients as waiting time increases, but the difference is not substantial. There is a big difference between the survival curves for hispanic patients versus non-hispanic patients, but this difference may arise because there is so little data on hispanic patients. There seems to be large differences in the curves for the number of symptoms displayed, with most people having four symptoms being seen before about 2 time units, while people with no symptoms had a much less dramatic drop-off in the probability of still waiting, meaning that they were generally seen later.

Modeling Approach

Generalized Linear Modeling with Kernel Regression

Generalized linear modeling is a form of inference that abstracts linear regression so that the error distribution of the response variable does not have to be normal. In a generalized linear model, the linear predictor for the explanatory variables is related to the expected value of the response variable through the link function, which in this case is the log odds. We use the log odds as the link function because we are implementing logistic regression through the generalized linear model paradigm. The only relevant assumption made for our generalized linear model is the the wait times for patients are independent of each other, which we are comfortable making. We do not need to assume that the residuals are normally distributed or that they have equal variance.

For this study, we used generalized linear modeling with kernel regression. Kernel regression is a non-parametric method that allows for flexible, non-linear regression by iteratively fitting the model within regions of the domain called kernels, where the set of kernels spans the entire domain. Kernel regression better models the differences in the data along the domain, especially when more data points are clustered in some areas of the domain and not in others. This is the case for our data, where most wait times are below 3 but can go up to around 26.

To implement generalized linear modeling, we needed to transform the data into a discrete format suitable for applying this technique, while also taking into account censoring. To do so, we divided up the wait times into bins of equal width, with one bin per time unit and 26 bins overall. We used fourteen kernels for the regression, and weighted the kernels for the lower bins (corresponding to the shorter wait times) more than the higher bins (corresponding to the longer wait times), because, again, more patients had shorter wait times until clinical assessment.

To test whether non-binary variables, such as number of symptoms, have a significant impact on wait time, we compared full models with the non-binary variables included to reduced models with the the non-binary variables excluded. If the full model was shown to be significant through ANOVA, then we concluded that the non-binary variables themselves were significant. We used this approach for the variable of clinical presentation, or the number of symptoms displayed. We did not do a full versus reduced model comparison for race, because we differentiated race (being black or not black) from ethnicity (being Hispanic or not Hispanic), thus keeping race and ethnicity binary. We first compared a full model with all explanatory variables included with a reduced model with the symptoms variable excluded. To evaluate the full versus reduced model, we used ANOVA based on the Chi-Squared likelihood ratio test, which is typically applied for comparing generalized linear models (Columbia Statistics). In all models, full and reduced, we included the kernel values.

```
##
## Call:
## glm(formula = y ~ 0 + ., family = "binomial", data = d2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4833  -0.8204  -0.6693   1.0284   2.5741
##
## Coefficients: (4 not defined because of singularities)
##         Estimate Std. Error z value Pr(>|z|)
## X1    -1.757e+83  2.650e+83  -0.663   0.5074
## X2            NA         NA      NA       NA
## X3            NA         NA      NA       NA
## X4            NA         NA      NA       NA
## X5            NA         NA      NA       NA
## X6     3.510e+13  5.296e+13   0.663   0.5074
## X7    -1.299e+10  1.960e+10  -0.663   0.5074
## X8     7.507e+03  1.088e+04   0.690   0.4901
## X9    -5.159e+01  2.373e+01  -2.174   0.0297 *
## X10   -4.549e+01  2.479e+01  -1.835   0.0665 .
## X11   -7.095e+01  3.347e+01  -2.120   0.0340 *
## X12   -1.080e+02  4.948e+01  -2.183   0.0290 *
## X13   -3.883e+02  1.801e+02  -2.156   0.0310 *
## X14   -4.564e+02  2.188e+02  -2.086   0.0370 *
## male  -2.519e-01  1.738e-01  -1.449   0.1473
## black -1.527e-02  1.986e-01  -0.077   0.9387
## hisp  -2.017e-01  4.565e-01  -0.442   0.6586
## sn1    3.716e+01  1.801e+01   2.063   0.0391 *
## sn2    3.719e+01  1.801e+01   2.065   0.0389 *
```

```
## sn3    3.716e+01  1.802e+01   2.063   0.0391 *
## all4   3.833e+01  1.803e+01   2.126   0.0335 *
## sn0    3.683e+01  1.801e+01   2.045   0.0409 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 944.07  on 681  degrees of freedom
## Residual deviance: 782.63  on 663  degrees of freedom
## AIC: 818.63
##
## Number of Fisher Scoring iterations: 25
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ 0 + (X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11 +
##     X12 + X13 + X14 + male + black + hisp + sn1 + sn2 + sn3 +
##     all4 + sn0)
## Model 2: y ~ 0 + X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11 +
##     X12 + X13 + X14 + male + black + hisp
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       663     782.63
## 2       668     794.92 -5  -12.293  0.03099 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

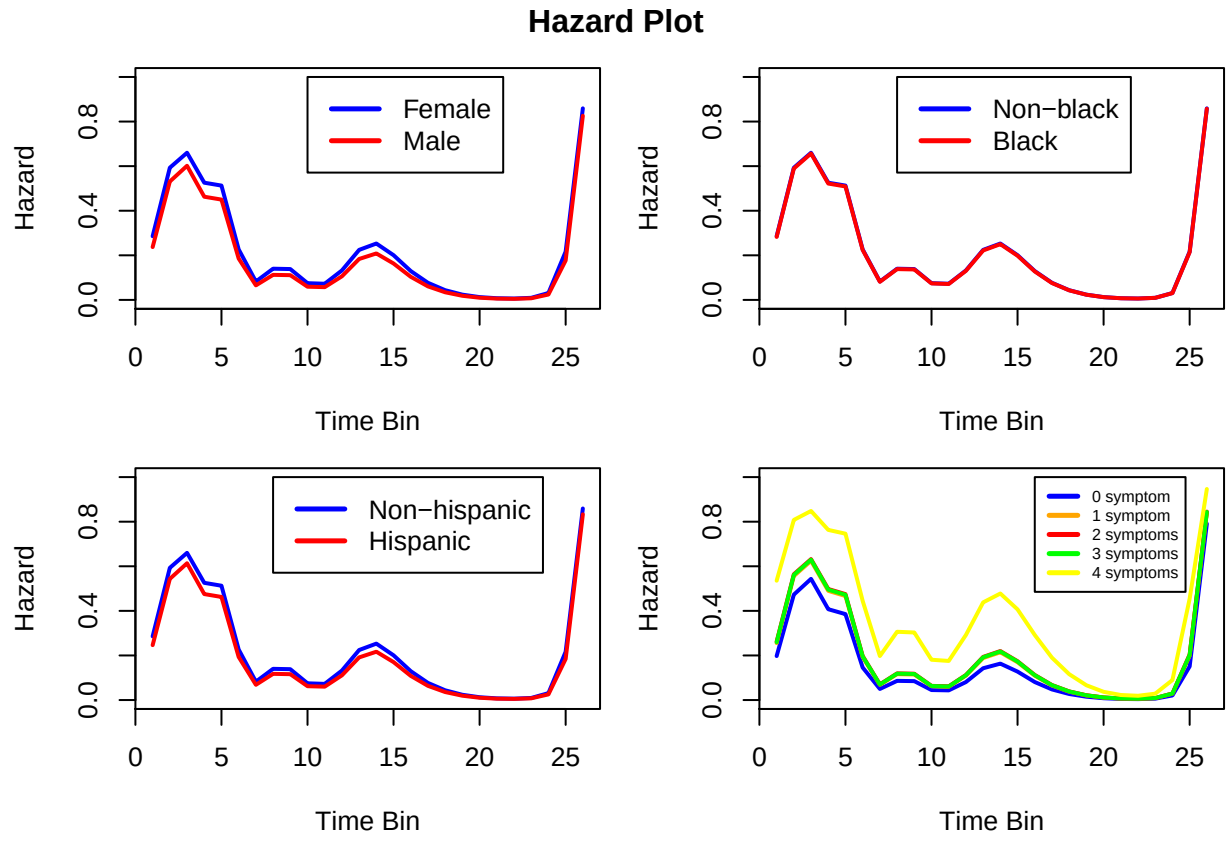
Based on the ANOVA results for the Chi-squared likelihood ratio test, the full model is statistically significant at an alpha of 0.05, meaning that the variable of number of symptoms displayed is significant in explaining wait times. Looking at the summary of the full model, (shown above the output for the ANOVA) we see that all the other explanatory variables, such as gender, whether or not the patient was hispanic, and whether or not the patient was black, had coefficients that were not statistically significant. Thus, these variables do not impact wait time. We considered this result and decided to compare the full model, including all the the binary explanatory variables and the clinical presentation variable, to a different reduced model, which contained just the clinical presentation variable but not the other explanatory variables, hypothesizing that this new reduced model could be regarded as the better model.

```
## Analysis of Deviance Table
##
## Model 1: y ~ 0 + (X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11 +
##     X12 + X13 + X14 + male + black + hisp + sn1 + sn2 + sn3 +
##     all4 + sn0)
## Model 2: y ~ 0 + X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11 +
##     X12 + X13 + X14 + sn0 + sn1 + sn2 + sn3 + all4
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       663     782.63
## 2       666     784.85 -3  -2.2176   0.5285
```

The results of this Chi-squared likelihood ratio ANOVA test show that the full model, including all the binary and non-binary explanatory variables, is not significant at an alpha of 0.05 when compared to the reduced model with only the clinical presentation variable included. We can surmise from these results that the reduced model is better than the full model. Therefore, including only the variables showing number of symptoms displayed as well as the kernel values produces the best model overall.

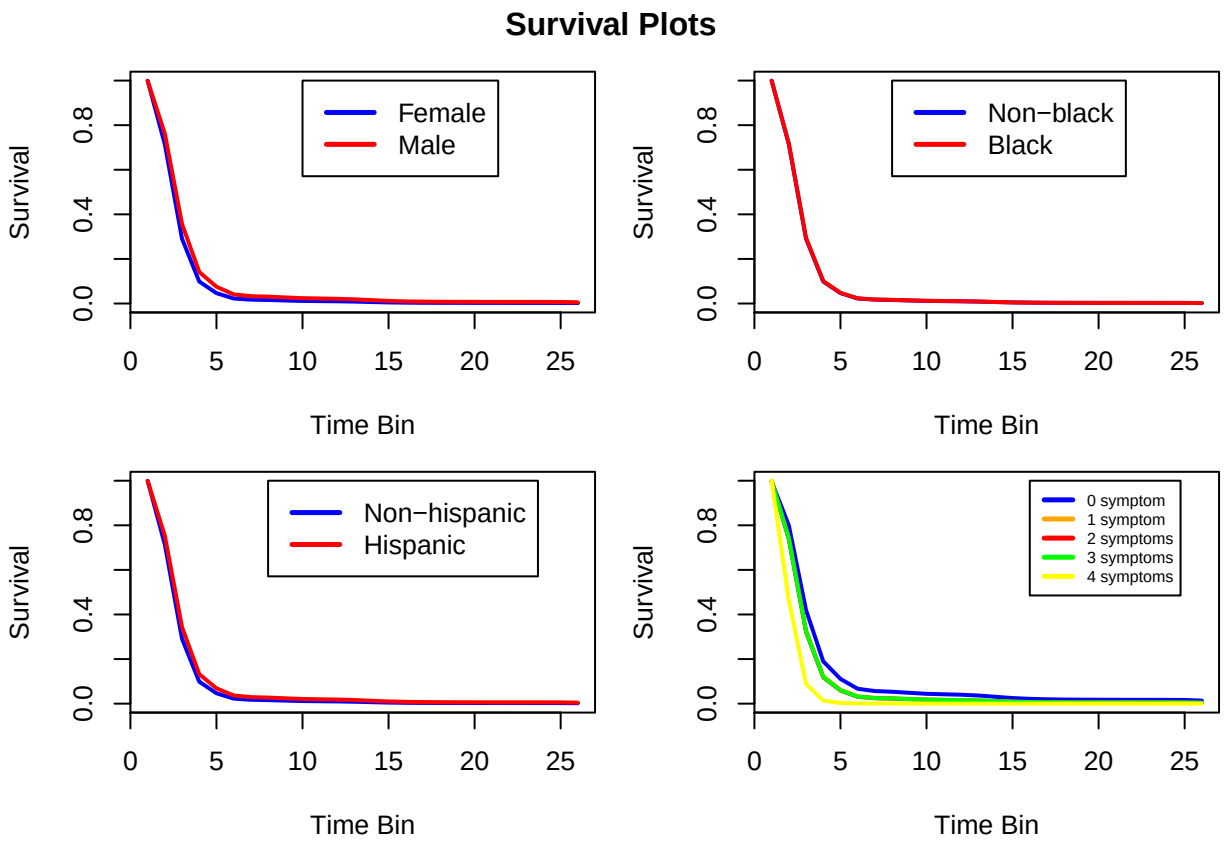
We next graph the hazard plots for each of the variables, while fixing the other variables that aren't the subject of the plot to arbitrary values. We expect the hazard plots generated from fixing different values of variables

that aren't the focus to be similar, so we only graph hazard plots for one configuration of the fixed variables.



In our models we see significance in the number of symptoms a patient has. This is evident in hazard plots above when comparing the levels of different explanatory variables. In each of the plots displayed, we plot a hazard curve for the different values of each predictor variable. We see that for Male, Black, and Hispanic, the different levels of the predictors result in very similar hazard curves that overlap heavily. This suggests that the different values of these predictors don't significantly affect the chance that a patient gets evaluated at any point in time. However, the hazard curve for a patient with 4 symptoms present is significantly different from the hazard curves for patients with 0-3 symptoms. This suggests that a patient displaying 4 symptoms has a higher chance at each point in time to be evaluated, and indicates that clinical presentation overall is a significant variable. Therefore, our hazard curves confirm what our models indicated about the significance of clinical presentation.

We can similarly confirm the significance of variables using survival plots, which are built through a probability transformation of the hazard values from the plots above. Therefore we allow the variable being graphed in the surival plot to vary while fixing the other variables in the same configuration as for the hazard plots above. Again, we don't expect the survival curves to change much with different configurations of the fixed variables.



The survival plots show little separation between male and female, black and not black and Hispanic and non-Hispanic. These variables are therefore shown visually, once more, to be insignificant in influencing wait time. There is separation between the survival curves for symptoms displayed, indicating again that the variable of clinical presentation is significant.

Conclusions

We evaluated a set of three generalized linear models with kernel regression, where one model was considered the best, and the other two we used to show that selected explanatory variables were not statistically significant predictors of the wait time for stroke patients in the ER. Based on the analysis of these models, we found that the variables of gender, being black or not being black, and being Hispanic or non-Hispanic had no impact on the wait times of stroke patients. However, the variable of clinical presentation, or how many symptoms were displayed, did influence wait time in our model set.

Although our models indicated that most of this study's explanatory variables were not significantly associated with wait time, this outcome may be a result of having too little data. For example, there were only nine Hispanic patients in this data set. If there had been more Hispanic patients, perhaps our models would have shown a statistically significant difference between the wait times of Hispanic and non-Hispanic patients. The problem of having too few observations affects the reliability of our results, but the only solution is to gather more data.

Even though our models show that most explanatory variables have no statistically significant impact on wait times, this result has positive implications even though it may initially seem uninteresting. Based on our analysis, the predictors of race, ethnicity, and gender have no impact on wait times, indicating that the selection process for clinical assessment in the ER is non-discriminatory. The only factor that may be relevant in influencing wait time is the number of symptoms displayed by the patient. We would expect and hope

that a patient with more stroke symptoms would have to wait less time for a clinical assessment, because it is even more urgent that this patient see a doctor. We have found evidence that suggests that a patient with stroke symptoms need not worry about discrimination in the emergency room based on race, ethinicity, or gender, and the only variable that may impact wait time is the number of symptoms exhibited.

References

"Introduction to Generalized Linear Models." *STAT 504*, Eberly College of Science at Penn State, onlinecourses.science.psu.edu/stat504/node/216.

"Generalized Linear Models." *Columbia Statistics*, Columbia University.

Contributions

William contributed generally to the code and the ideas underpinning the code, while also contributing to writing the model approach section. Wuming wrote the code for analyzing the models and creating the hazard plots, and helped interpret the results. Annie wrote parts of the model approach section, generated the survival plots, and wrote the conclusion.