Case 2 Part 1

Sonia Xu, Grant Goettel, Ian Hua

September 29, 2017

Preliminary Analysis

This report explores the dataset provided from a study of time critical to neurological assessment for patients, so our group can form an initial understanding of demographic bias in regards to projected wait time. It covers how the data was cleaned, exploratory data analysis findings, and potential approaches to answer this challenge.

Cleaning the Data

Since the symptoms cannot be uniquely identified, we decided to to combine the count of all symptoms into one covariate called $count_sn$. This new variable is the count of the number of symptoms each observation may have, so it ranges from 0 to 4.

nctdel	fail	male	black	hisp	count_sn
1.2000000	1	0	0	0	2
0.8666667	1	0	0	0	2
1.6666667	1	0	0	0	1
1.8500000	1	1	1	0	1
1.2666667	1	1	0	0	2
0.9833333	1	0	0	0	1

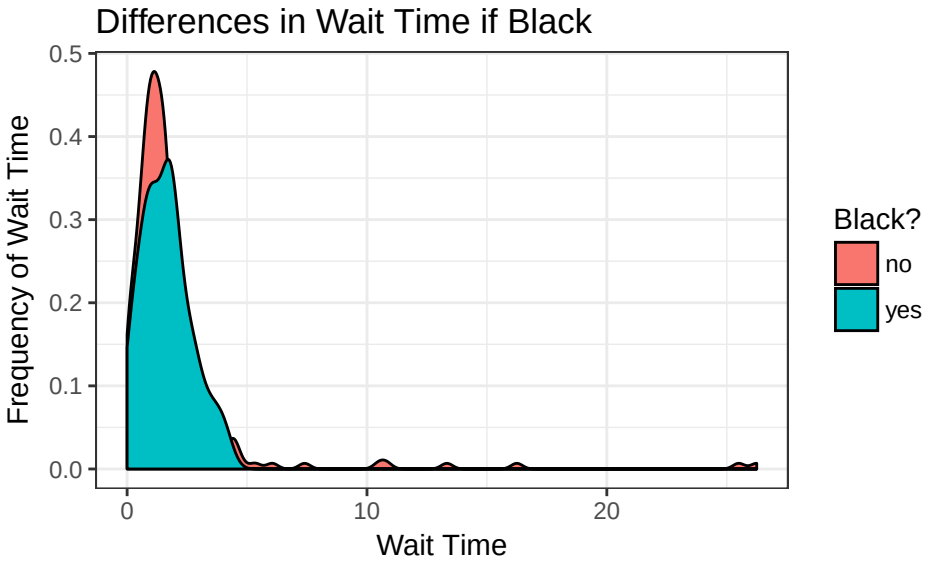
Exploratory Data Analysis

After cleaning the dataset, there are five covariates that can predict the response, wait time. Since three of these covariates are associated with patient demographics, we hypothesize that there may exist social bias in the wait time of a patient, and explore this assumption via exploratory data analysis.

Racial Bias

Black vs. Non-Black

There could potentially be bias here, but more analysis should be done. Non-blacks appear to have shorter wait times, but some non-blacks have extremely long wait times.



Hispanic vs. Non-Hispanic

There could potentially be bias here, but the lack of sample size for Hispanics appears to be an issue.

Gender Bias

There could potentially be bias here, but more analysis should be done. Females appear to have shorter wait times.

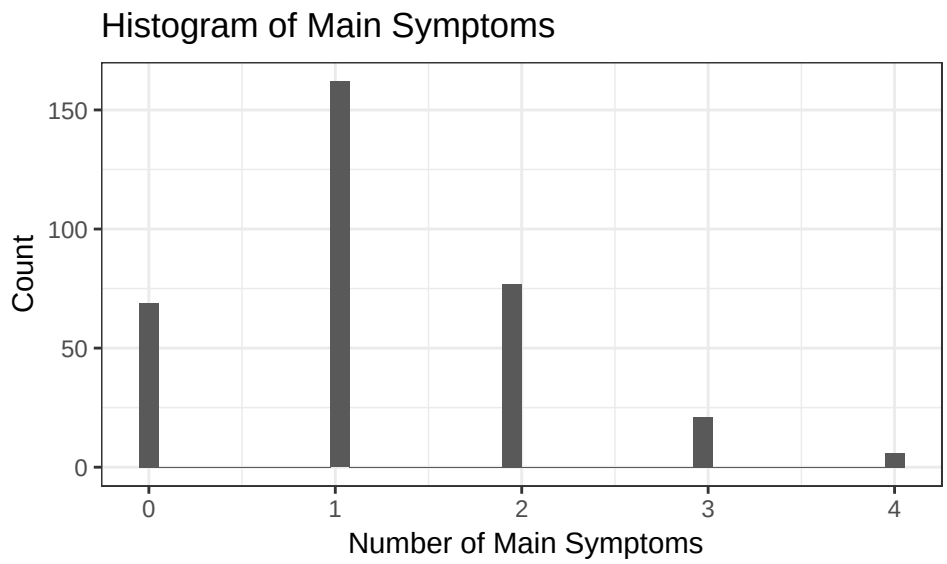
Demographic Differences for Those Who Fail to Receive Scans?

From comparing the correlation between who was rejected and who wasn't, there appears to be no significant difference.

nctdel	fail	male	black	hisp	count_sn
0.2342803	1	0.0062187	-0.0136992	0.0272369	0.094639

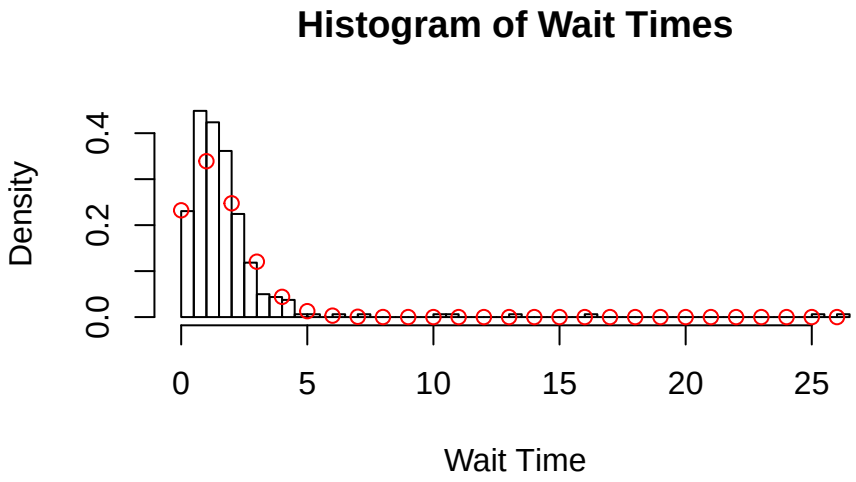
Distribution of Main Symptoms

Most observations showed only one symptom, so there is enough data to analyze the demographic bias between patients who solely exhibited one symptom. However, we need to remain cognisant of the small sample size for observations that exhibited three or four symptoms before we reach any final conclusions (sample size of 21, 6 respectively).



Possible Approaches to the Data

In an effort to understand factors predictive of wait time, we could model the wait time for each patient as a Poisson distribution. However, after attempting to fit a Poisson distribution with $\lambda = 1.46$, we noticed that it did not capture the data robustly enough. The observed data seemed to have a higher peak than the Poisson model we tried to fit.



Below are a couple potential approaches to alleviate this challenge:

1. We plan to separate the model into bins, and to model each bin with a piecewise constant hazard model. The separated model might fit the data better, so from there we could get a more accurate picture of whether there are biases in the determination of wait time.
2. We could non-parametrically estimate it via a Kaplan-Meier estimate. This would be used in combination with other techniques, as a non-parametric test would not be as powerful as a parametric one, but it would be able to point us in the right direction in terms of trends to look for.

3. Similarly, we could apply a Cox proportional hazards model to this data. By assuming that the effects of the predictor variables upon survival are constant over time and additive in one scale, a Cox regression could provide better estimates of survival probabilities and cumulative hazard.

4. We could also cluster the data to check to see whether or not people with similar characters have similar wait times from the waiting room to the ER.