Case Study V Final Report

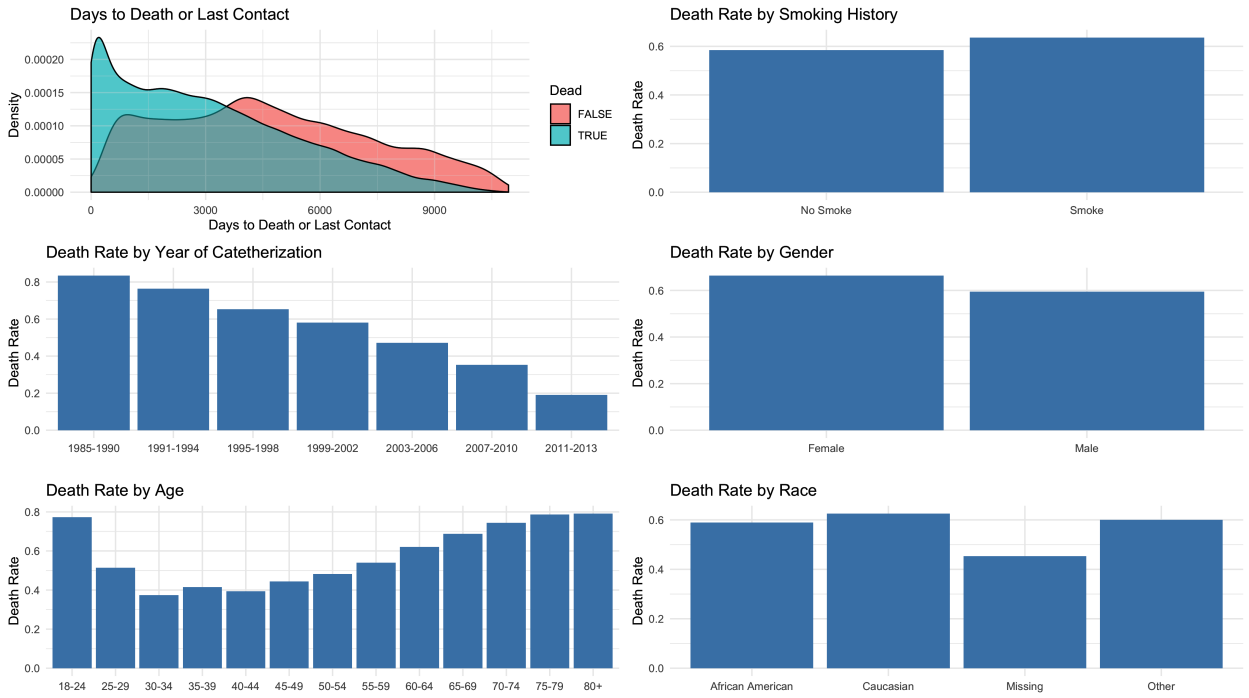
Jake Epstein, Daniel Spottiswood, Sahil Patel, Michael Tan, Man-Lin Hsiao

11/10/2019

Introduction

The goal of this case study is to evaluate risk factors for death for Duke Cardiology Patients. Data comes from Duke Databank for Cardiovascular Diseases and includes observations on about ~85,000 patients. As this is medical data, much of the data is missing. In order to account for biases from missing data, we will utilize multiple imputation. We will create and interpret a Cox Model on complete case data in order to evaluate risk factors for death. We will then compare that model to a model built with imputed data in order to evaluate the impact of missing data on the model.

Exploratory Data Analysis



First, we examine the distribution of days to death or last known contact. This chart confirms what we expect. We see that most of the early deaths are accounted for but that longer term survivors often lose contact before a confirmed death.

To explore how demographics of patients relate to death rate, we plotted death rate by smoking history, by gender, by race, by year of catetherization, and by age. We see that smokers have higher death rate than non-smokers, females have higher death rate than males, and Caucasians have the highest death rate followed by Other races, then African Americans. However, these differences are relatively small. On the other hand, we can see a very significant pattern in death rate by year of catetherization as well as age. The earlier the catetherization was performed, the higher death rate - this can likely be attributed to progress being made in medicine as catetherization becomes increasingly safer and more effectively treating patients as years progress. The youngest and oldest age groups have highest death rates. Death rates for 18-24 year olds are high, then decreases progressively until the 30-34 age group, then increases progressively after that.

Missing Data

We see several variables with missing data rates over 50 percent, and many more in the ~20-40% range.

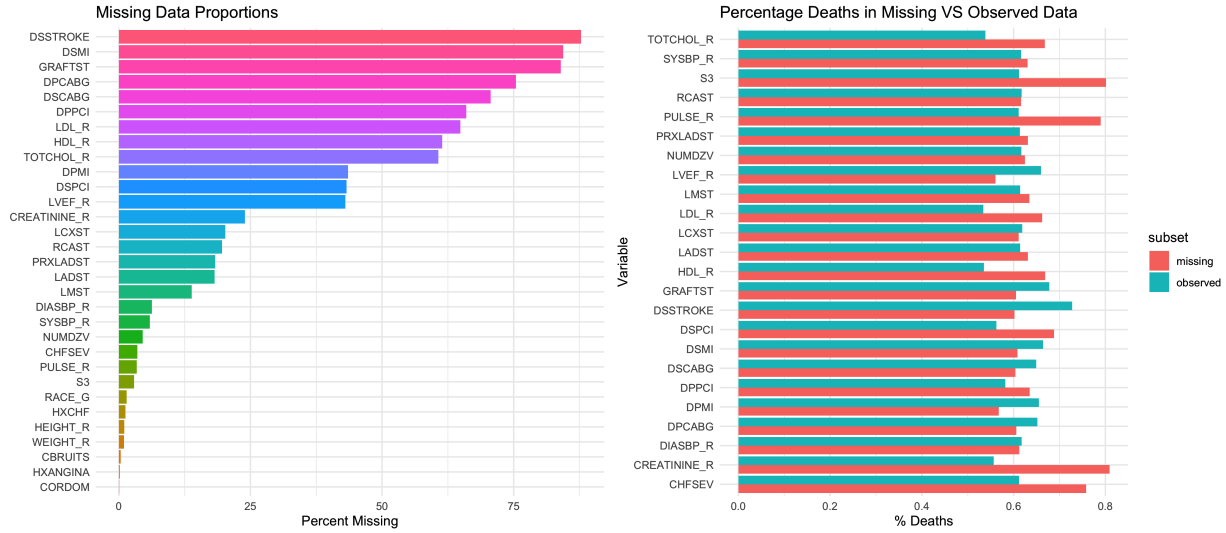
Most instances of large missing data are for events that likely never occurred, such as all the "days to first subsequent" variables like DSSTROKE (having a stroke in the future) and DSMI (having a non-fatal myocardianl infarction in the future), or "days to closest previous" variables like DPCABG (having a previous coronary artery bypass surgery).

On the other hand, some variables with a significant amount of missing data do not have easily interpretable explanations. For example laboratory measurements such as LDL_R, HDL_R, and TOTCHOL_R have significant proportions of missing data, but there is no straightforward explanation to why these measurements may not have been recorded for certain patients.

We see several variables with missing data rates over 50 percent, and many more in the ~20-40% range.

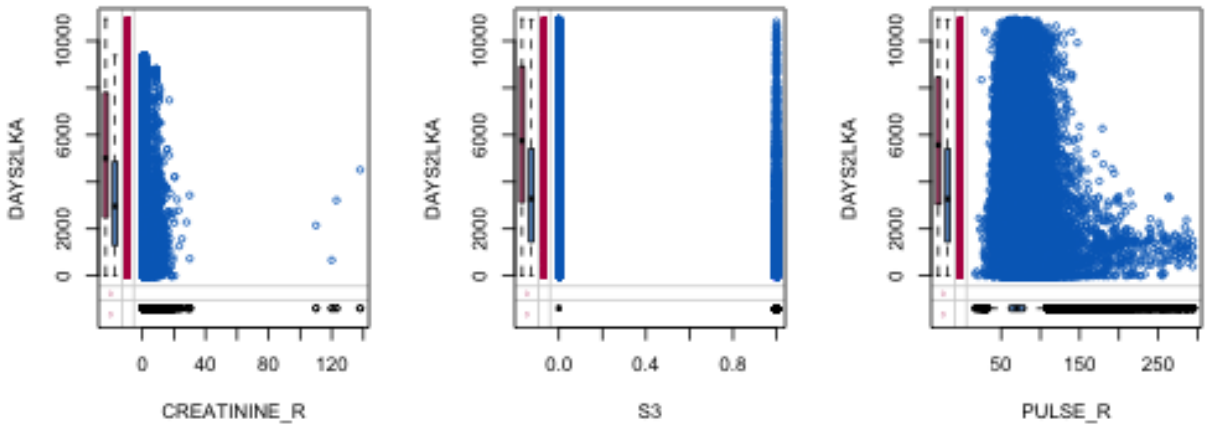
Most instances of large missing data are for events that likely never occurred, such as all the "days to first subsequent" variables like DSSTROKE (having a stroke in the future) and DSMI (having a non-fatal myocardianl infarction in the future), or "days to closest previous" variables like DPCABG (having a previous coronary artery bypass surgery).

On the other hand, some variables with a significant amount of missing data do not have easily interpretable explanations. For example laboratory measurements such as LDL_R, HDL_R, and TOTCHOL_R have significant proportions of missing data, but there is no straightforward explanation to why these measurements may not have been recorded for certain patients.



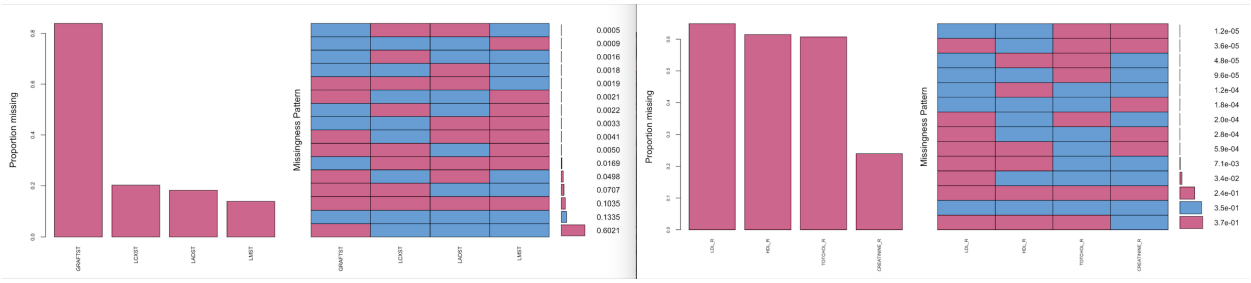
We see several variables with missing data rates over 50 percent, and many more in the ~20-40% range. Most instances of large missing data are for events that likely never occurred, such as all the "days to first subsequent" variables like DSSTROKE (having a stroke in the future) and DSMI (having a non-fatal myocardianl infarction in the future), or "days to closest previous" variables like DPCABG (having a previous coronary artery bypass surgery).On the other hand, some variables with a significant amount of missing data do not have easily interpretable explanations. For example laboratory measurements such as LDL_R, HDL_R, and TOTCHOL_R have significant proportions of missing data, but there is no immediate explanation to why these measurements may not have been recorded for certain patients.
In order to investigate how missing data affects the outcome of death, we plotted the percentage of deaths in every variable for the subset where data on the variable is missing, and the subset where data on the variable is observed. Through the graph, we can notice that five specific variables: CHFSEV, CREATININE_R, PULSE_R, S3, and TOTCHOL_R have particularly large differences in percentage deaths for their observed versus missing subsets. For all five variables, there was a higher percentage of deaths for the subset of data

that is missing. This suggests that for these variables, there is a possibility that data may be MAR (missing at random), since missingness is related to death, an observed variable.



We continued to investigate the five aforementioned variables and their effects on DAYS2LKA through margin plots. Although there were no noteworthy differences in distribution for CHSEV and TOTCHOL_R, we did find that missing data seems to effect DAYS2LKA for the variables CREATININE_R, S3, and PULSE_R. For all three of these variables, subsets of missing data have boxplots with higher means, indicating that patients with missing data on these variables died later than patients with observed data. Again, this suggests possible MAR (missing at random), since missingness is related to days to last known alive, an observed variable.

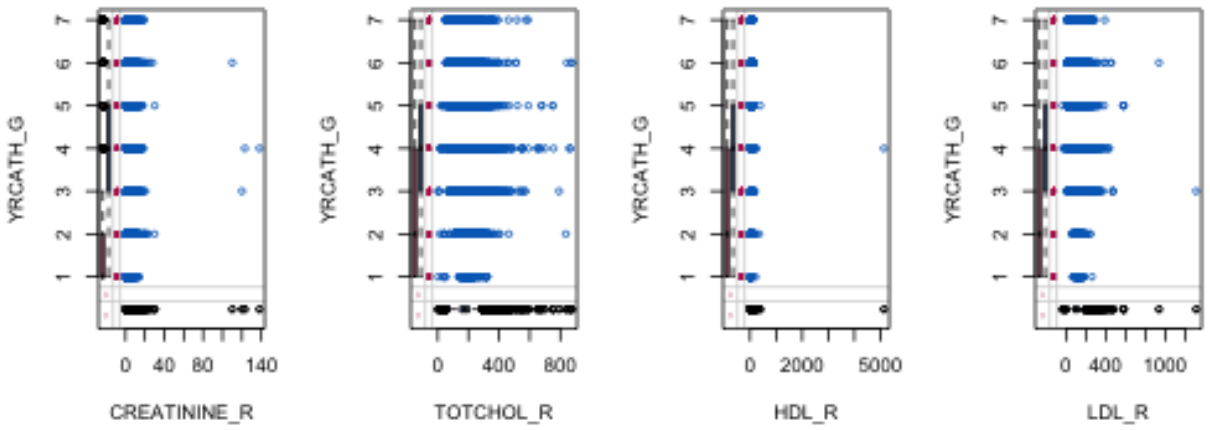
We will proceed with visualizations that may help determine any similar patterns for missing data in different variables.



Through examining the variables with the highest proportion of missing data with missingness pattern visualizations, the following can be seen:

1) When looking at GRAFTST, LCXST, LADST, and LMST, all four of these have missing values together 10.4% of the time. These are all catetherization results, so it could be the case the post-catetherization measurements were not done well or completely. It is likely that the missing data results from cases where the doctor just did not bother to measure all the metrics after catetherization, which would mean that the data is missing completely at random (MCAR).

2) When looking at LDL_R, HDL_R, TOTCHOL_R, and CREATININE_R, the variables excluding CREATININE_R have missing values together 37% of the time, and all four are missing 24% of the time. We continue to investigate this in the following graphs.



Through these margin plots, we can see there is a significant non-overlap in the distribution of YRCATH_G (year of catetherization) data depending on whether the laboratory measurements were missing versus observed. The majority of missing data on all these laboratory measurements comes from early years, while all the observed data comes from later years. A likely explanation is that these laboratory procedures were less common or accessible in the earlier years which would mean that this data is missing at random (MAR).

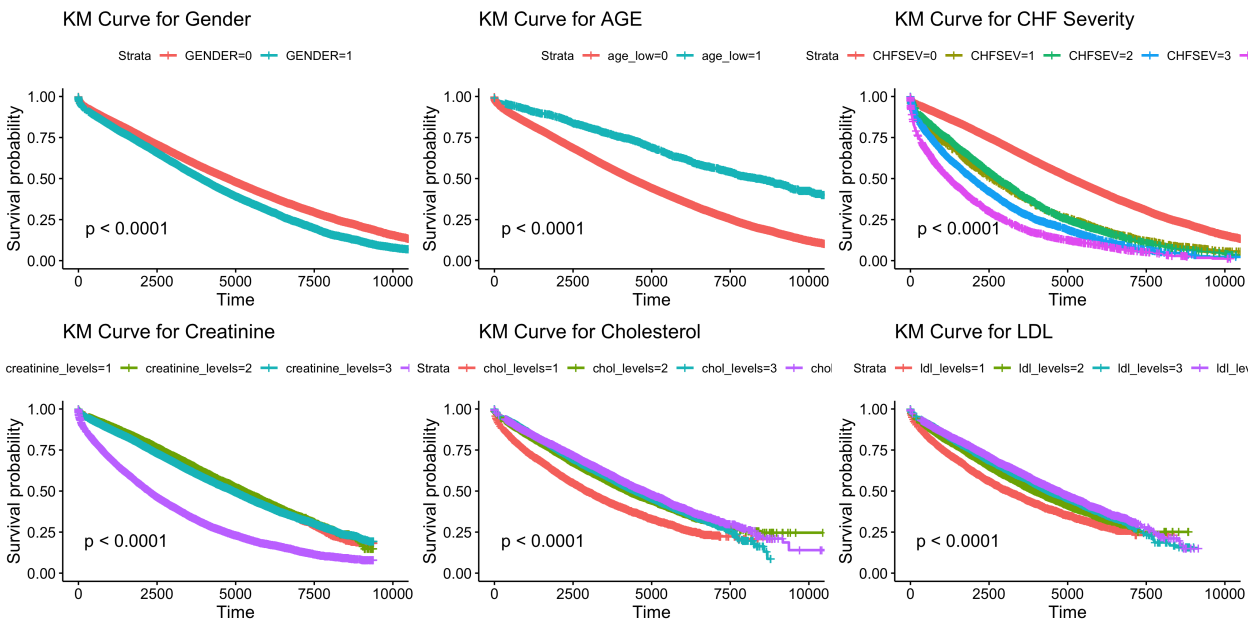
In summary, through these exploratory data analysis and missing data visualizations, we can determine that significant portions of missing data are due to either events that never occurred and therefore cannot be recorded, possible negligence in recording certain measurements (MCAR), and reasons that depend on other observed variables, such as less accessible laboratory procedures in earlier years leading to missing data (MAR). We also note that year of catetherization and age of patient are important variables that have especially significant effects towards death.

Build Model

In order to build a model to determine factors associated with survival, we first explored relationships in isolation using Kaplan-Meier curves. We then selected medically relevant variables and variables we found to be significant through our Kaplan-Meier analysis and used those to do Cox proportional hazards regression modeling. These variables included demographic data, data on the severity of heart conditions, and selected lab data. Finally, we used MICE to impute missing data, and created a cox model on our imputed data set to evaluate the impact of missing data.

Kaplan-Meier Curves

The plots below depict the Kaplan-Meier curves (the risk of death over time) of variables we found to demonstrate clear effects that we could draw inference from. The plots also include the p-value dervied from conducting log-rank tests on the Kaplan-Meier curves. The Kaplan-Meier curves are depictions of the variable's survival function over time. The probability of survival begins at 1 (time = 0), and approaches 0 as time goes to infinity. The log-rank test determines whether the differences between a variable's Kaplan-Meier curves are statistically significant. All of the variables examined had p-values <0.0001, implying that they were statistically significant.



The Kaplan-Meier curves for the variable gender shows that females have a statistically significant higher risk of dying than males. In addition, the curves show that this difference in risk of death increases as time increases (the slope of the red, male, curve increases faster than the blue, female curve).

In order to determine whether there was an effect for age, we categorised an individual's age to be low (<40) or high (>=40). The Kaplan-Meier curves for this variable show a distinguished difference in the risk of death for the two groups. Namely, individuals who are 40 or older have a statisitcally significant increased risk of death. Furthermore, as time passes, this difference increases furhter; in other words, as time passes individuals who are 40 or older have an increasing risk of dying when compared to those under the age of 40. Contextually, this is a reasonable outcome as individuals who are older are more likely to die than their juniors.

The Kaplan-Meier curves for congestive heart failure (CHF) severity indicate that individuals who never experienced CHF have the least risk of dying than those have had CHF. Moreover, the risk of dying increases as CHF severity increases. Intuitively, this is a feasible outcome as individuals with the most severe CHF are characterised by an "inability to perform any physical activity without discomfort" and potential discomfort while resting; logically, individuals who experience this level of pain are at higher risk of dying than those with less severe symptoms.

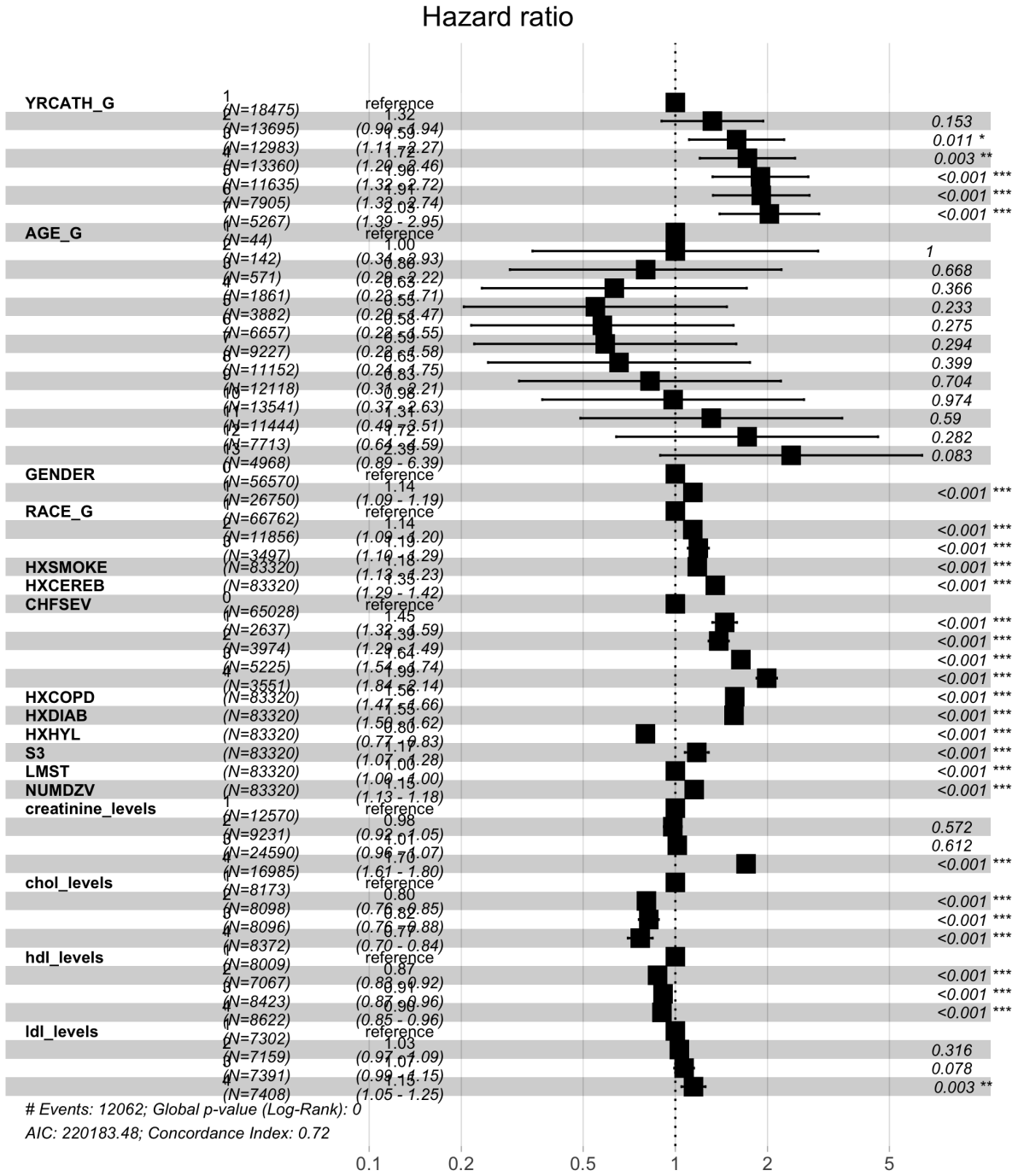
creatinine

cholesterol

ldl

We will explore these relationships further using Cox proportional hazards regression modelling. Hopefully, the Cox regression methodology will provide more insight into how these differences in risk of dying manifest at a more granular level.

Complete Case Cox Model



The forest plot above illustrates the hazard ratios for different values of the covariates explored as estimated through Cox proportional hazards regression modelling. A hazard ratio >1 implies that individuals with the associated value for the covariate have an increased chance of dying. Intuitively, hazard ratios <1 imply that the individual indicate a decreased risk of dying.

In order to run the Cox model, we established that our survival object would measure time with DAYS2LKA as we are working with censored, not interval, data, and we set the event to be 0 if the individual died

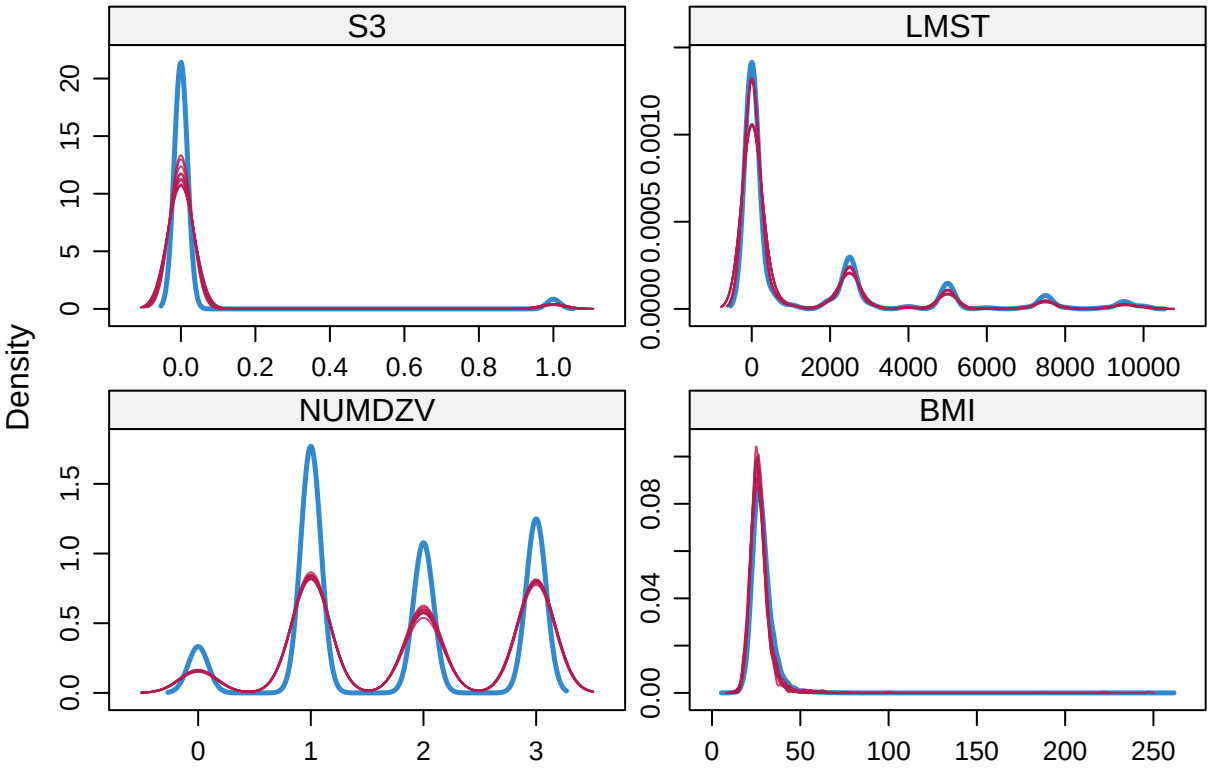
and 1 if the individual stayed alive. We modelled the survival object using the following covariates: year of cardiac catheter (categorised), age in years (categorised), gender, race, history of smoking (binary), history of cerbrovascular disease (binary), congestive heart failure severity (binary), history of chronic obsructive pulmonary disease (binary), history of diabetes (binary), history of hyperlipidemia (binary), third hear sound (binary), maximum stenosis of left main artery, number of significantly diseased valves.

Our forest plot indicates that as an individual's age increases from 18-39 their risk of dying decreases; however, this trend reverses for individuals ages 39 and over. Furthermore, the same trend occurs for white and non-white individuals, where non-white individuals are at higher risk of dying. Moreover, this trend appears in the variable CHFSEV as well, where individuals with a severity of 0 are the least at risk. Interestingly, YRCATH_G has decreasing risk as the group number moves from 1-3, then increases for groups 4-7.

[TODO: give context to why the relationship do or don't make sense]

[TODO: justification for why we want to do MICE, why we chose the m we did, etc]

Multiple Imputation



Comparison to Complete Case Model