1.0 Introduction

The dataset used in this report comes from the Duke University Cardiovascular Disease Databank and was obtained from http://biostat.mc.vanderbilt.edu/DataSets. It consists of data collected from 3504 patients who were sent to the Duke University Medical Center due to chest pain. The variables in the dataset are:

- `sex` (`sex = 1` for female; `sex = 0` for male)
- `age` (in years)
- `cad.dur` (duration of symptoms of coronary artery disease)
- `cholesterol`
- `sigz` (`sigdz = 1` if patient has significant coronary disease; `sigdz = 0` otherwise)

```
d <- read.csv("acath.csv")
d <- d[,-6]
head(d)
```

```
   sex age cad.dur choleste sigdz
1    0  73     132      268     1
2    0  68      85      120     1
3    0  54      45       NA     1
4    1  58      86      245     0
5    1  56       7      269     0
6    0  64       0       NA     1
```

Although this dataset has 3504 observations, there are 1246 patients whose cholesterol measurements are missing (approximately 36%). The purpose of the analysis in this report is to predict the odds of significant coronary disease given a patient's sex, age, duration of disease symptoms and cholesterol level, and to determine if these variables are significantly associated with the disease.

Since 36% of cholesterol values are missing, techniques to deal with missing data are applied: listwise deletion, inverse probability weighting, single imputation (mean, conditional mean, and regression imputation), as well as multiple imputation. The main results of each analysis are discussed, and the missing data techniques are compared.

2.0 Problem of Interest

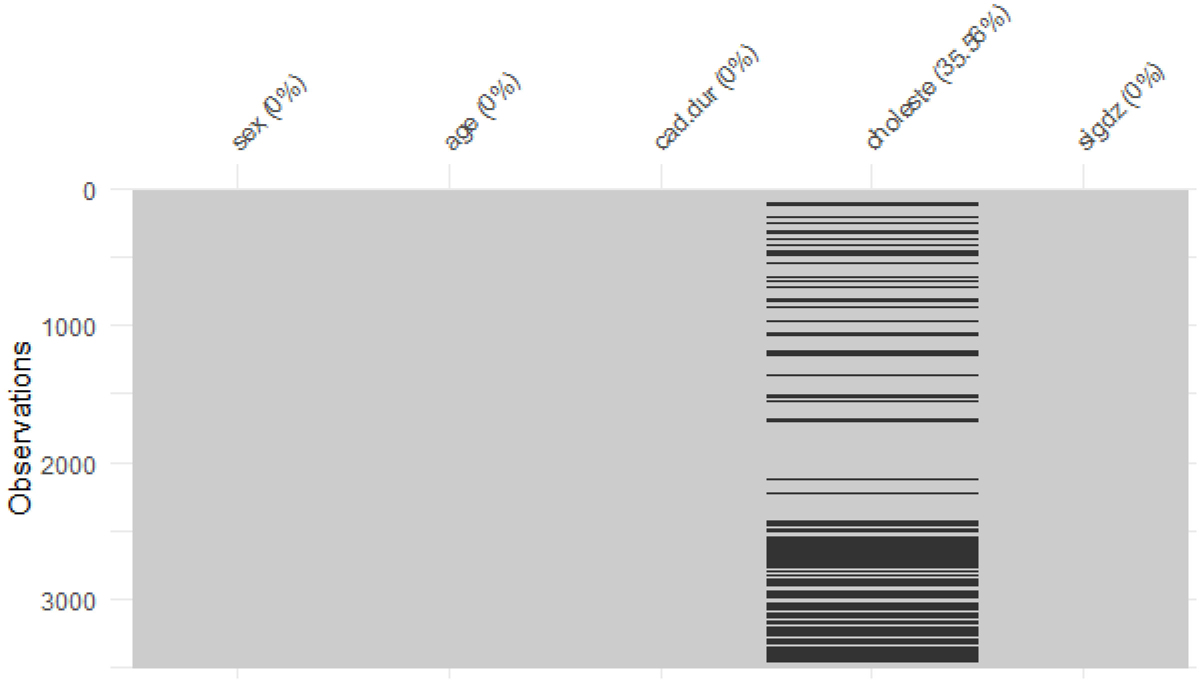
This analysis aims to answer the research question: For the patients who were referred to Duke Medical Center for chest pain, what are the odds of developing significant coronary disease given sex, age, duration of symptoms, and cholesterol level?

To deal with missing values in the cholesterol variable and to make better statistical inferences, various missing data techniques are applied.

3.0 Missing Mechanism

To get a better idea of the missingness in the data, the following plot shows the percentage of missing values in each variable and the observations at which the values are missing. As the plot indicates, the variable that contains missing values is cholesterol, with approximately 36% missing.

```
library(naniar)
vis_miss(d)
```



Also, a summary of the data provides an outline of the contents and shows that 1246 out of 3504 cholesterol values are missing.

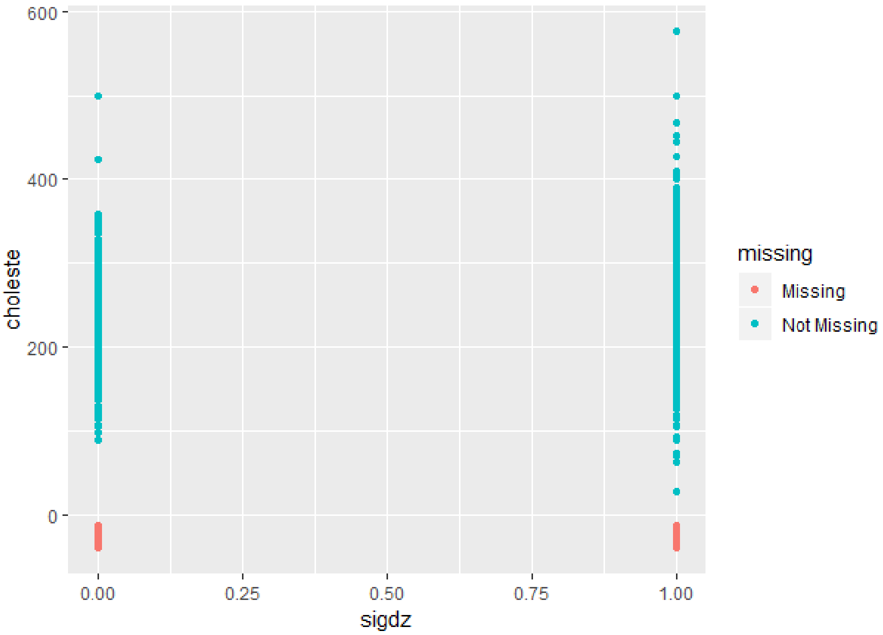
```
nrow(d) # 3504
```

```
summary(d)
```

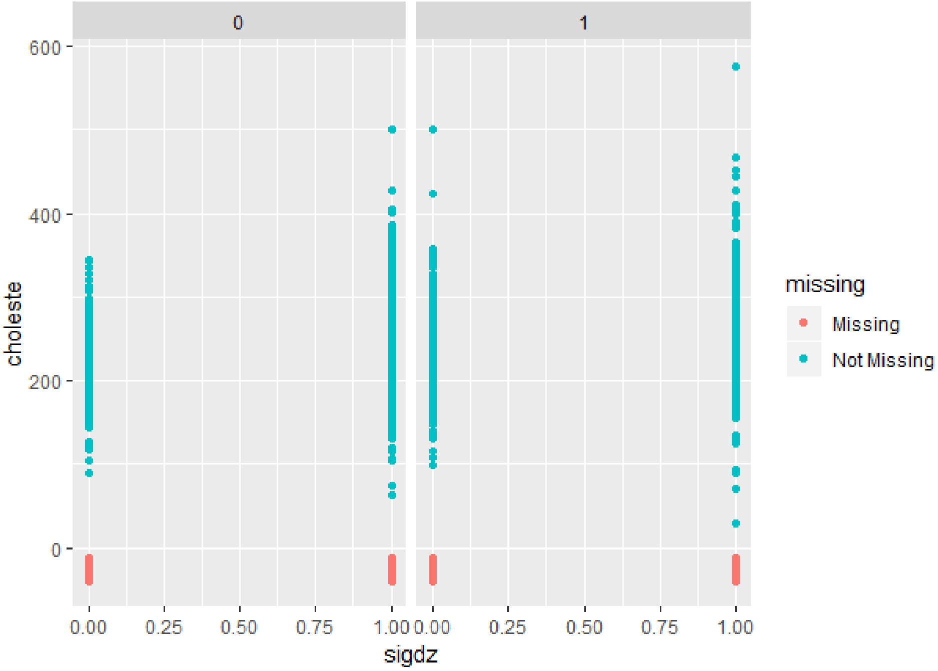
```
      sex              age            cad.dur        choleste          sigdz       
 Min.   :0.0000   Min.   :17.00   Min.   :  0   Min.   : 29.0   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:46.00   1st Qu.:  4   1st Qu.:196.0   1st Qu.:0.0000  
 Median :0.0000   Median :52.00   Median : 18   Median :224.5   Median :1.0000  
 Mean   :0.3136   Mean   :52.28   Mean   : 43   Mean   :229.9   Mean   :0.6661  
 3rd Qu.:1.0000   3rd Qu.:59.00   3rd Qu.: 60   3rd Qu.:259.0   3rd Qu.:1.0000  
 Max.   :1.0000   Max.   :82.00   Max.   :416   Max.   :576.0   Max.   :1.0000  
                                                NA's   :1246                    
```

Furthermore, we can look at the missingness between variables, such as cholesterol and coronary disease overall, then by gender groups.

```
library(ggplot2)
p <- ggplot(d,
         aes(x = sigdz,
             y = choleste)) +
     geom_miss_point()
p
```



```
p + facet_wrap(~sex)
```



While these plots are helpful in visualizing the missing data and possible associations between variables, they do not provide information on the missing data mechanism or type of missingness.

Using the observed data, we can check if the cholesterol data is missing completely at random (MCAR) by creating an indicator variable for missingness (R = 1 if missing; R = 0 if observed), then by fitting a logistic regression model of R as a function of sex, age, cad.dur and sigdz to see if any of these variables are significant.

```
d$R <- ifelse(is.na(d$choleste), 1, 0)

mech <- glm(R ~ sex + age + cad.dur + sigdz,
            family = "binomial", data = d)

summary(mech)

Coefficients:
              Estimate  Std. Error  z value  Pr(>|z|)
(Intercept) -2.9277771  0.2077119   -14.095   <2e-16 ***
sex         -0.0507309  0.0846696   -0.599   0.5491
age          0.0475965  0.0041443   11.485   <2e-16 ***
cad.dur     -0.0012319  0.0006351   -1.940   0.0524 .
sigdz       -0.1702029  0.0862924   -1.972   0.0486 *
```

From the output, we can see that the missingness in the cholesterol variable is related to the variables age and sigdz, which means that the missingness is not MCAR. Although missing at random (MAR) and missing not at random (MNAR) cannot be checked using the observed data, I assume that the missing cholesterol values are MAR because the missing mechanism may depend on other variables.

4.0 Methods and Main Results

Now that the missing mechanism of the cholesterol variable has been identified, the following sections discuss the application of approaches for analyzing the research question using various missing data techniques.

4.1 Listwise Deletion

Before we apply imputation or weighting techniques to deal with missing data, listwise deletion analyzes the data by dropping observations with missing values. Since the cholesterol missingness is not MCAR, this estimation will be biased, and standard errors will be inflated due to the smaller sample size. Nevertheless, we model the coronary disease outcome as a function of sex, age, cad.dur, and cholesterol.

```
d2 <- d[,-6]

model_ld <- glm(sigdz ~ sex + age + cad.dur + choleste,
        family = "binomial", data = d2)

summary(model_ld)

Coefficients:
           Estimate   Std. Error  z value   Pr(>|z|)
(Intercept) -4.286432   0.386800    -11.082   <2e-16 ***
sex         -2.101717   0.113559    -18.508   <2e-16 ***
age          0.073032   0.006146     11.883   <2e-16 ***
cad.dur     -0.001695   0.001014    -1.672    0.0945 .
choleste     0.009123   0.001079     8.453    <2e-16 ***

(1246 observations deleted due to missingness)
```

After dropping all the patients whose cholesterol measurements were missing, this analysis indicates that sex, age, and cholesterol are significantly associated with the odds of coronary disease, but duration of symptoms is not. The odds of coronary disease for females versus males is significantly less, and a one-year increase in age leads to higher odds of disease. Similarly, a one unit increase in cholesterol leads to an increase in the odds of coronary disease.

As previously noted, the estimates in this basic analysis are likely biased since the data is not missing completely at random and over a third of the data was dropped. The following sections employ more thorough techniques to account for the missing data.

4.2 Inverse Probability Weighting

The approach used in this section is inverse probability weighting; although this technique still uses the observed data only and drops observations with missing cholesterol values, it weights the data using the fitted values from the missing mechanism modeled in section 3.0. These weights compensate for the patients who are under-represented in the observed data.

If the data is assumed to be missing at random and the missing mechanism is correctly modeled, this analysis should provide unbiased estimates.

```
mech <- glm(R ~ sex + age + cad.dur + sigdz,
           family = "binomial", data = d)

qhat <- mech$fitted

wt <- 1/(1 - qhat)

model_ipw <- glm(sigdz ~ sex + age + cad.dur + choleste,
                weights = wt, family = "quasibinomial", data = d)

summary(model_ipw)
```

```
  Coefficients:
                Estimate Std. Error t value Pr(>|t|)
  (Intercept) -4.1549693  0.3929540 -10.574  < 2e-16 ***
  sex         -2.1203921  0.1138561 -18.623  < 2e-16 ***
  age          0.0714148  0.0061051  11.698  < 2e-16 ***
  cad.dur     -0.0014886  0.0009998  -1.489    0.137
  choleste     0.0086405  0.0010801   7.999 1.98e-15 ***

  (1246 observations deleted due to missingness)
```

The results of this model are very similar to the listwise deletion model results. The estimates for the coefficients are the same in both magnitude and sign as the listwise deletion model, and the significance of the variables has not changed (sex, age and cholesterol are significantly associated with coronary disease, but duration of symptoms is not).

The remaining sections in this report use imputation to assign values to missing cholesterol measurements and preserve the full sample size. Section 4.3 applies single imputation and Section 4.4 applies multiple imputation.

4.3 Single Imputation

By using single imputation, we keep the full sample size, but substitute in values for missing cholesterol measurements based on estimates from the rest of the observations.

4.3.1 Mean Imputation

Mean imputation is a simple approach in which the missing cholesterol values are replaced with the average of the observed cholesterol values. In the observed data, the average cholesterol is 229.9283.

```
d3 <- d2

mean_choleste <- mean(as.data.frame(filter(d3,
                !is.na(d3$choleste)))$choleste) # 229.9283

for (i in 1:nrow(d3)){

      if (is.na(d3$choleste[i]))
         {d3$choleste[i] = mean_choleste}

      else {d3$choleste[i] = d3$choleste[i]}

}

model_mean_imp <- glm(sigdz ~ sex + age + cad.dur +
                   choleste, family = "binomial", data = d3)

summary(model_mean_imp)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.6307612  0.3388281 -13.667   <2e-16 ***
sex         -2.0353795  0.0910867 -22.346   <2e-16 ***
age          0.0786127  0.0047453  16.567   <2e-16 ***
cad.dur     -0.0003107  0.0007909  -0.393    0.694
choleste     0.0088262  0.0010582   8.341   <2e-16 ***
```

Again, these results are very similar to the results from both the listwise deletion and inverse probability weighting models. The coefficient estimates have not changed very much, and the significance of the variables has not changed.

4.3.2 Conditional Mean Imputation

In conditional mean imputation, we replace missing cholesterol values with the conditional mean from different groups; in this case, we calculate the cholesterol means among males and females and then replace missing cholesterol values based on the patient’s sex. The mean cholesterol level for females is 236.7692 and the mean for males is 226.9242.

```
d4 <- d2

mean_f <- mean(as.data.frame(filter(d4, !is.na(d4$choleste)
           & d4$sex == 1))$choleste) # 236.7692

mean_m <- mean(as.data.frame(filter(d4, !is.na(d4$choleste)
          & d4$sex == 0))$choleste) # 226.9242

for (i in 1:nrow(d4)){

      if (d4$sex[i] == 1 & is.na(d4$choleste[i]))
           {d4$choleste[i] = mean_f}

      if (d4$sex[i] == 0 & is.na(d4$choleste[i]))
           {d4$choleste[i] = mean_m}

      else {d4$choleste[i] = d4$choleste[i]}

}

model_cond_imp <- glm(sigdz ~ sex + age +
                      cad.dur + choleste,
                      family = "binomial", data = d4)

summary(model_cond_imp)

Coefficients:
           Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.6402205  0.3386462 -13.702   <2e-16 ***
sex         -2.0682581  0.0918928 -22.507   <2e-16 ***
age          0.0784370  0.0047443  16.533   <2e-16 ***
cad.dur     -0.0003036  0.0007909  -0.384    0.701
choleste     0.0089437  0.0010665   8.386   <2e-16 ***
```

Just like the listwise deletion, inverse probability weighting and simple mean imputation models, the variables sex, age and cholesterol are significant while duration of symptoms is not. The sign of the variables is also the same as in the other models.

4.3.3 Regression Imputation – Random Forest

Compared to mean and conditional mean imputation, regression imputation is more useful since it predicts the missing value using information from the rest of the variables. Using random forest in R, we can get the imputed values for the missing cholesterol measurements, and then fit the outcome model using the new imputed cholesterol variable.

```
library(randomForest)

d5 <- d2

fit_rf <- randomForest(choleste ~ sex + age +
                       cad.dur + sigdz,
                       data = d5, na.action = na.omit)

d5$chol.impute <- predict(fit_rf, newdata = d5)

d5$chol.impute[!is.na(d5$choleste)] <-
                           d5$choleste[!is.na(d5$choleste)]


model_rf <- glm(sigdz ~ sex + age + cad.dur + chol.impute,
           family = "binomial", data = d5)

summary(model_rf)

Coefficients:
           Estimate     Std. Error z value  Pr(>|z|)
(Intercept) -5.3286952  0.3480167 -15.312   <2e-16 ***
sex         -2.1157369  0.0932548 -22.688   <2e-16 ***
age          0.0788628  0.0047841  16.484   <2e-16 ***
cad.dur     -0.0006420  0.0007996  -0.803    0.422
chol.impute  0.0120208  0.0011162  10.769   <2e-16 ***
```

The coefficient estimates from this model are slightly different than the previous models (the signs are the same, but the magnitudes are marginally different), but the variables still have the same significance as in the other models. By using random forest, the coefficient of the imputed cholesterol variable is approximately 1.3 times higher than in the other models.

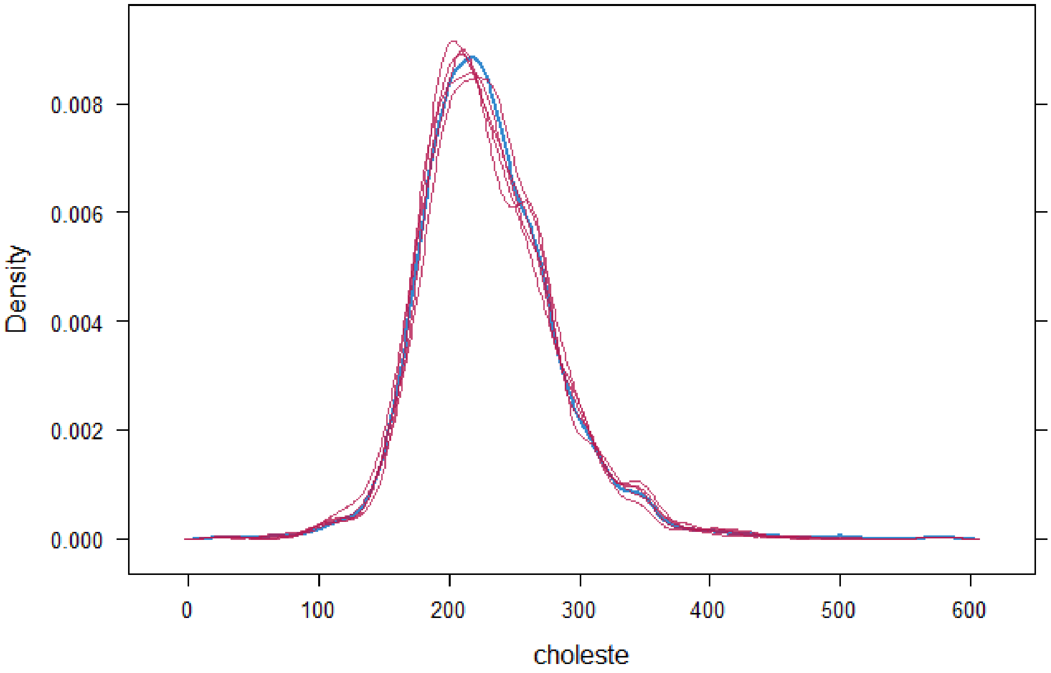
4.4 Multiple Imputation

In this section, multiple imputation is used to help account for the uncertainty in the missing cholesterol values. If the data is missing at random like I have assumed, this approach will work well. To do multiple imputation, the R package `mice` is used, and the imputation process is repeated 5 times.

```
library(mice)

d6 <- d2

mult_imp <- mice(d6, m = 5, seed = 12345, print = F)
```

Before the outcome model with the imputed values is fit, van Buuren and Groothuis-Oudshoorn (2011) note that it is important to check the imputations by comparing the imputed values with the observed values to see whether the imputed values are reasonable. This can be done by using the `densityplot()` function.

```
densityplot(mult_imp, scales = list(x = list(relation = "free")))
```



This plot shows that the imputed values (red lines) are reasonably close to the observed values (blue line), which provides evidence that the imputed values are reasonable.

Now, the outcome model using the imputed values can be fit.

```
fit_mult_imp <- with(mult_imp, glm(sigdz ~ sex + age +
                          cad.dur + choleste,
                          family = "binomial"))

round(summary(pool(fit_mult_imp)), 3)
```

```
           estimate   std.error  statistic       df   p.value
(Intercept)   -4.692     0.356   -13.178     76.551   0.000
sex           -2.100     0.093   -22.578   3358.809   0.000
age            0.080     0.005    16.482   1949.061   0.000
cad.dur       -0.001     0.001    -0.652   1943.088   0.514
choleste       0.009     0.001     7.918   24.005     0.000
```

The results from this model, again, are very similar to all other models. The coefficient estimate for cholesterol from this model is close to the estimates from all models except the random forest model (the coefficient of the imputed cholesterol variable is approximately 1.3 times higher in the random forest model than in the other models). Like the previous models, sex, age, and cholesterol are significant, but duration of symptoms is not.

5.0 Conclusion

In this report, the Duke Cardiac Catheterization Coronary Artery Disease Diagnostic Dataset was used to predict the odds of developing significant coronary disease among the patients given sex, age, duration of symptoms, and cholesterol level. Since 1246 out of 3504 cholesterol values were missing, various techniques for dealing with missing data were applied, including mean and conditional mean imputation, regression imputation, and multiple imputation.

After applying each missing data technique, a simple logistic regression model that fit the disease outcome as a function of sex, age, duration of symptoms and cholesterol was run. The models for each technique produced very similar results for the coefficient estimates and the significance of the variables. Each model found that age, sex and cholesterol were significantly associated with coronary disease, but duration of symptoms was not. For females, the odds of coronary disease is significantly less compared to males, and a one-year increase in age is associated with higher odds of disease. Also, a one unit increase in cholesterol is associated with an increase in the odds of coronary disease.

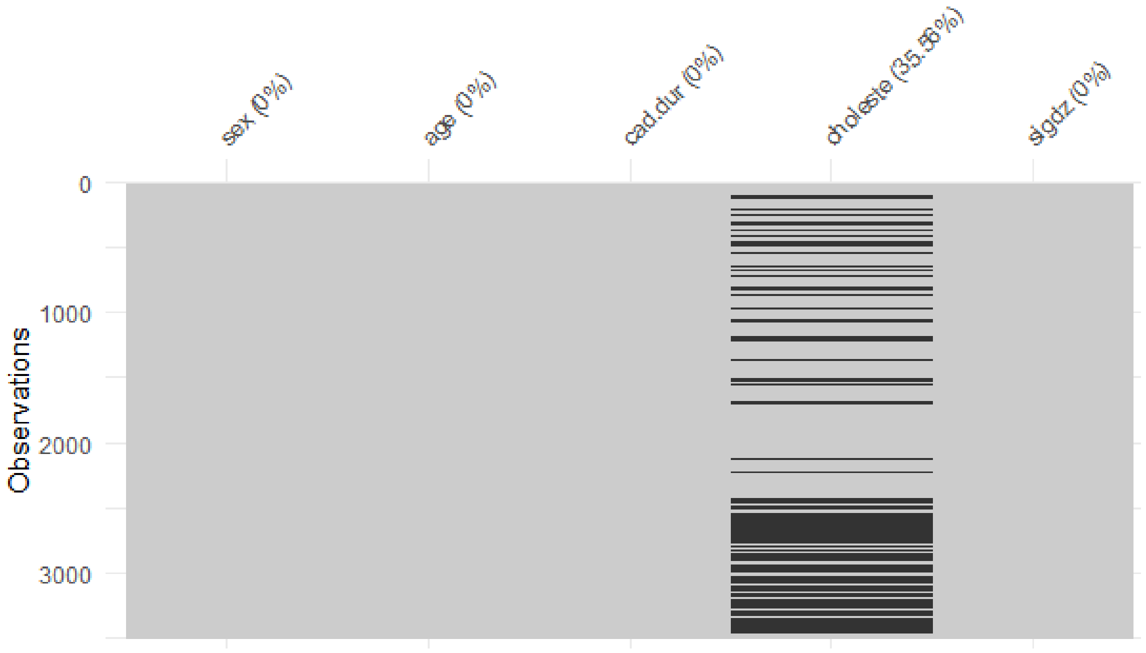
Appendices

Appendix A: The Dataset

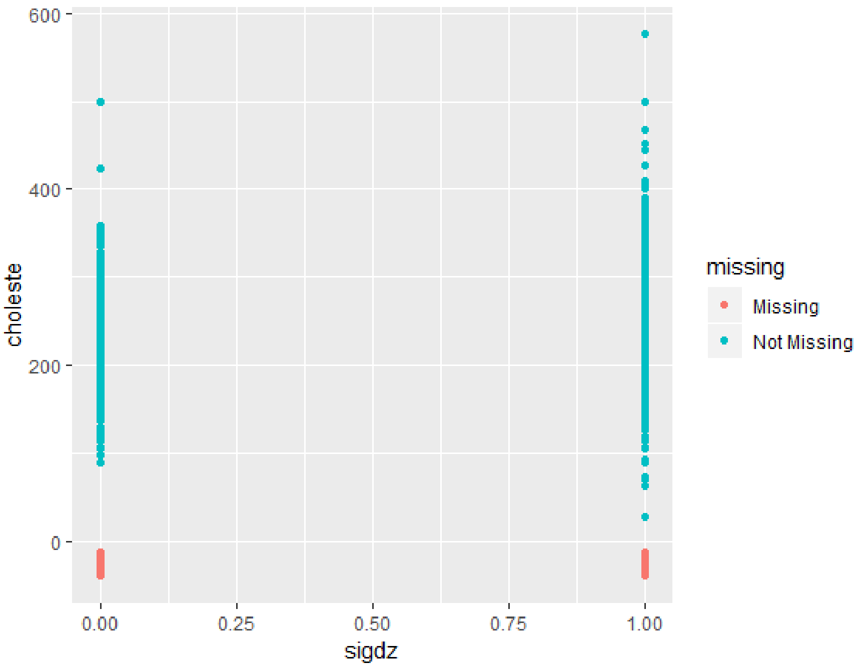
```
d <- read.csv("acath.csv")
d <- d[,-6]

head(d)

  sex age cad.dur choleste sigdz
1   0  73     132      268     1
2   0  68      85      120     1
3   0  54      45       NA     1
4   1  58      86      245     0
5   1  56       7      269     0
6   0  64       0       NA     1

tail(d)

     sex age cad.dur choleste sigdz
3499   0  46       1      196     1
3500   0  58      14      295     1
3501   1  71      27       NA     1
3502   0  67      11       NA     1
3503   1  66     247       NA     1
3504   0  67     123       NA     1

summary(d)

      sex              age           cad.dur       choleste          sigdz       
 Min.   :0.0000   Min.   :17.00   Min.   :  0   Min.   : 29.0   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:46.00   1st Qu.:  4   1st Qu.:196.0   1st Qu.:0.0000  
 Median :0.0000   Median :52.00   Median : 18   Median :224.5   Median :1.0000  
 Mean   :0.3136   Mean   :52.28   Mean   : 43   Mean   :229.9   Mean   :0.6661  
 3rd Qu.:1.0000   3rd Qu.:59.00   3rd Qu.: 60   3rd Qu.:259.0   3rd Qu.:1.0000  
 Max.   :1.0000   Max.   :82.00   Max.   :416   Max.   :576.0   Max.   :1.0000  
                                                NA's   :1246                    
```

Appendix B: Missing Mechanism

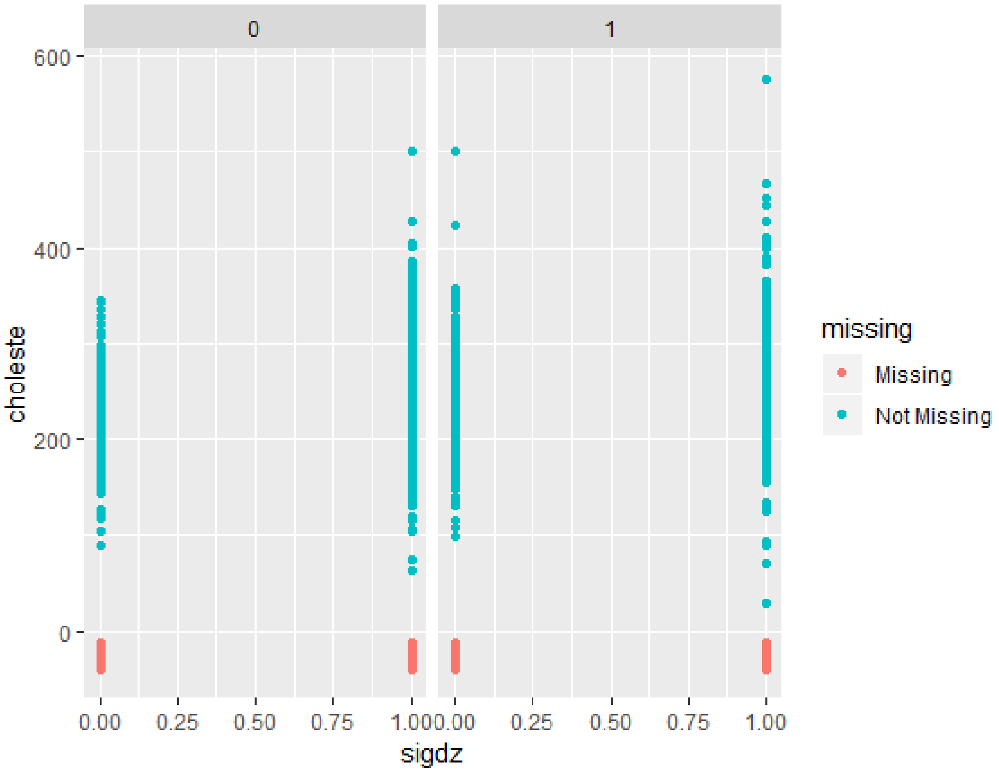
```
library(naniar)

vis_miss(d)
```



```
missing <- miss_var_summary(d)

View(missing)

  variable n_miss pct_miss

1 choleste   1246     35.6
2 sex           0      0
3 age           0      0
4 cad.dur       0      0
5 sigdz         0      0

library(ggplot2)

p <- ggplot(d,
         aes(x = sigdz,
             y = choleste)) +
     geom_miss_point()

p
```



```
p + facet_wrap(~sex)
```



```
d$R <- ifelse(is.na(d$choleste), 1, 0)

mech <- glm(R ~ sex + age + cad.dur + sigdz, family = "binomial", data = d)
summary(mech)

Call:
glm(formula = R ~ sex + age + cad.dur + sigdz, family = "binomial", data = d)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.5101  -0.9512  -0.7881   1.2751   2.0927  

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -2.9277771  0.2077119 -14.095   <2e-16 ***
sex         -0.0507309  0.0846696  -0.599   0.5491    
age          0.0475965  0.0041443  11.485   <2e-16 ***
cad.dur     -0.0012319  0.0006351  -1.940   0.0524 .  
sigdz       -0.1702029  0.0862924  -1.972   0.0486 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4561.1  on 3503  degrees of freedom
Residual deviance: 4414.6  on 3499  degrees of freedom
AIC: 4424.6

Number of Fisher Scoring iterations: 4
```

Appendix C: Listwise Deletion

```
d2 <- d[,-6]

model_ld <- glm(sigdz ~ sex + age + cad.dur + choleste,
              family = "binomial", data = d2)

summary(model_ld)
```

```
Call:
glm(formula = sigdz ~ sex + age + cad.dur + choleste, family = "binomial",
    data = d2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5401  -0.8704   0.5282   0.7663   2.4125

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.286432   0.386800 -11.082   <2e-16 ***
sex         -2.101717   0.113559 -18.508   <2e-16 ***
age          0.073032   0.006146  11.883   <2e-16 ***
cad.dur     -0.001695   0.001014  -1.672   0.0945 .
choleste     0.009123   0.001079   8.453   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2895.3  on 2257  degrees of freedom
Residual deviance: 2344.6  on 2253  degrees of freedom
  (1246 observations deleted due to missingness)
AIC: 2354.6

Number of Fisher Scoring iterations: 4
```

Appendix D: Inverse Probability Weighting

```
qhat <- mech$fitted

wt <- 1/(1 - qhat)

model_ipw <- glm(sigdz ~ sex + age + cad.dur + choleste,
                 weights = wt, family = "quasibinomial", data = d)

summary(model_ipw)
Call:
glm(formula = sigdz ~ sex + age + cad.dur + choleste, family =
"quasibinomial",
    data = d, weights = wt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.1940  -1.0485   0.6894   0.9379   2.5835

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.1549693  0.3929540 -10.574  < 2e-16 ***
sex         -2.1203921  0.1138561 -18.623  < 2e-16 ***
age          0.0714148  0.0061051  11.698  < 2e-16 ***
cad.dur     -0.0014886  0.0009998  -1.489    0.137
choleste     0.0086405  0.0010801   7.999 1.98e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasibinomial family taken to be 1.568169)

    Null deviance: 4467.7  on 2257  degrees of freedom
Residual deviance: 3624.0  on 2253  degrees of freedom
  (1246 observations deleted due to missingness)
AIC: NA

Number of Fisher Scoring iterations: 4
```

Appendix E: Mean Imputation

```
d3 <- d2

mean_choleste <- mean(as.data.frame(filter(d3,!is.na(d3$choleste)))$choleste)

for (i in 1:nrow(d3)){

  if (is.na(d3$choleste[i])){d3$choleste[i] = mean_choleste}

  else {d3$choleste[i] = d3$choleste[i]}

}

model_mean_imp <- glm(sigdz ~ sex + age + cad.dur + choleste,
                       family = "binomial", data = d3)

summary(model_mean_imp)

Call:
glm(formula = sigdz ~ sex + age + cad.dur + choleste, family = "binomial",
    data = d3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5416  -0.8666   0.5146   0.7861   2.4857

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.6307612  0.3388281 -13.667   <2e-16 ***
sex         -2.0353795  0.0910867 -22.346   <2e-16 ***
age          0.0786127  0.0047453  16.567   <2e-16 ***
cad.dur     -0.0003107  0.0007909  -0.393    0.694
choleste     0.0088262  0.0010582   8.341   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4463.5  on 3503  degrees of freedom
Residual deviance: 3621.6  on 3499  degrees of freedom
AIC: 3631.6

Number of Fisher Scoring iterations: 4
```

Appendix F: Conditional Mean Imputation

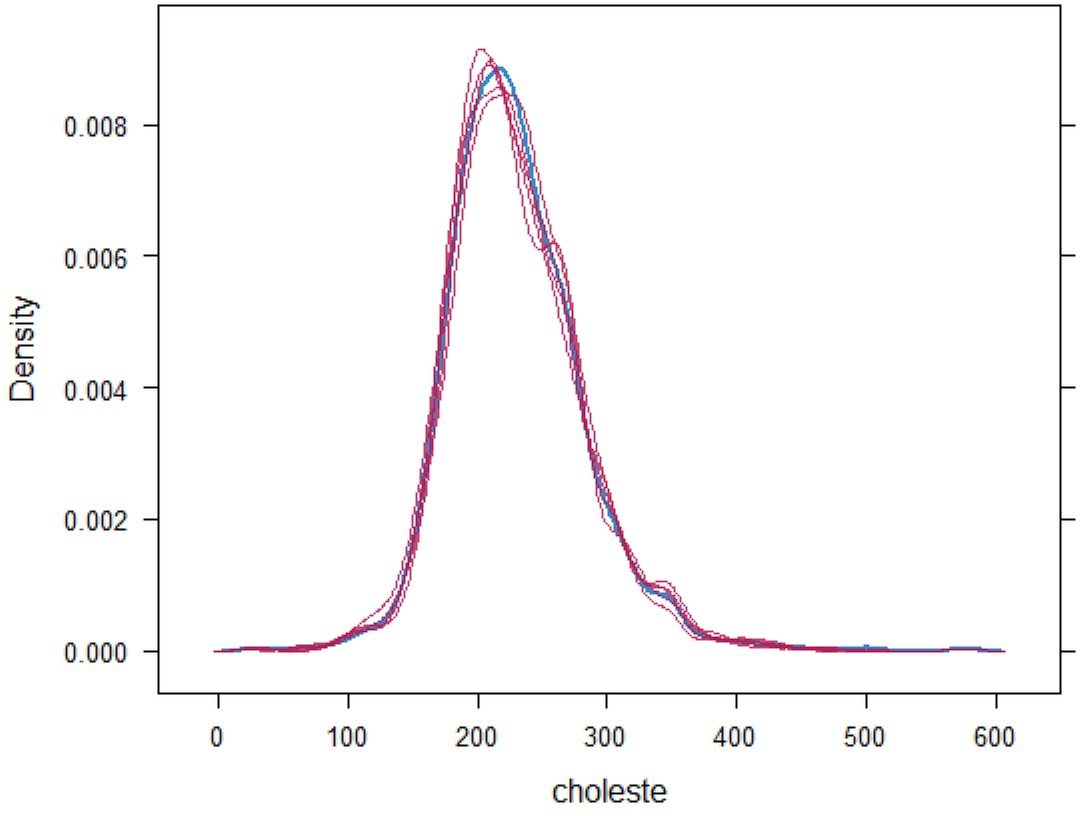
```
d4 <- d2

mean_f <- mean(as.data.frame(filter(d4, !is.na(d4$choleste)
      & d4$sex == 1))$choleste)

mean_m <- mean(as.data.frame(filter(d4, !is.na(d4$choleste)
      & d4$sex == 0))$choleste)

for (i in 1:nrow(d4)){

  if (d4$sex[i] == 1 & is.na(d4$choleste[i])) {d4$choleste[i] = mean_f}

  if (d4$sex[i] == 0 & is.na(d4$choleste[i])) {d4$choleste[i] = mean_m}

  else {d4$choleste[i] = d4$choleste[i]}

}

model_cond_imp <- glm(sigdz ~ sex + age + cad.dur + choleste,
                      family = "binomial", data = d4)

summary(model_cond_imp)

Call:
glm(formula = sigdz ~ sex + age + cad.dur + choleste, family = "binomial",
    data = d4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5432  -0.8782   0.5180   0.7831   2.4951

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.6402205  0.3386462 -13.702   <2e-16 ***
sex         -2.0682581  0.0918928 -22.507   <2e-16 ***
age          0.0784370  0.0047443  16.533   <2e-16 ***
cad.dur     -0.0003036  0.0007909  -0.384    0.701
choleste     0.0089437  0.0010665   8.386   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4463.5  on 3503  degrees of freedom
Residual deviance: 3620.5  on 3499  degrees of freedom
AIC: 3630.5

Number of Fisher Scoring iterations: 4
```

Appendix G: Regression Imputation - Random Forest

```
library(randomForest)

d5 <- d2

fit_rf <- randomForest(choleste ~ sex + age + cad.dur + sigdz,
                       data = d5, na.action = na.omit)

d5$chol.impute <- predict(fit_rf, newdata = d5)

d5$chol.impute[!is.na(d5$choleste)] <- d5$choleste[!is.na(d5$choleste)]


model_rf <- glm(sigdz ~ sex + age + cad.dur + chol.impute,
                family = "binomial", data = d5)

summary(model_rf)

Call:
glm(formula = sigdz ~ sex + age + cad.dur + chol.impute, family = "binomial",
    data = d5)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6292  -0.8351   0.4996   0.7686   2.5514

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.3286952  0.3480167 -15.312   <2e-16 ***
sex         -2.1157369  0.0932548 -22.688   <2e-16 ***
age          0.0788628  0.0047841  16.484   <2e-16 ***
cad.dur     -0.0006420  0.0007996  -0.803    0.422
chol.impute  0.0120208  0.0011162  10.769   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4463.5  on 3503  degrees of freedom
Residual deviance: 3564.9  on 3499  degrees of freedom
AIC: 3574.9

Number of Fisher Scoring iterations: 4
```

Appendix H: Multiple Imputation

```
library(mice)

d6 <- d2

mult_imp <- mice(d6, m = 5, seed = 12345, print = F)

fit_mult_imp <- with(mult_imp, glm(sigdz ~ sex + age + cad.dur + choleste,
                                   family = "binomial"))

round(summary(pool(fit_mult_imp)), 3)

            estimate std.error statistic       df p.value
(Intercept)   -4.692     0.356   -13.178   76.551   0.000
sex           -2.100     0.093   -22.578 3358.809   0.000
age            0.080     0.005    16.482 1949.061   0.000
cad.dur       -0.001     0.001    -0.652 1943.088   0.514
choleste       0.009     0.001     7.918   24.005   0.000

densityplot(mult_imp, scales = list(x = list(relation = "free")))
```



References

Buuren, S. V., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate Imputation by Chained Equations inR. *Journal of Statistical Software*,*45*(3). doi:10.18637/jss.v045.i03

Vanderbilt University Biostatistics. (2002). Vanderbilt University Datasets. Retrieved from http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets