Exploration and Modeling of Warfarin Dosage

Tiffany Pang

12/14/2017

The objective of the project is to classify and predict therapeutic warfarin dosage. The warfarin dosing estimator [http://warfarindosing.org/Source/Home.aspx (http://warfarindosing.org/Source/Home.aspx)] is used as a reference as to which variables to include in the model.

```
library(dplyr)
library(stringr)
setwd("/Users/tiffany/Documents/USF_MSHI/HS 630/HS630_Assignments/Warfarin")

iwpc_data <- read.delim("iwpc_warfarin.txt", header = TRUE, sep = "\t", dec = ".")
```

Data Cleaning

The variables are renamed for easier referencing.

```
# Rename the variables
names(iwpc_data)[names(iwpc_data) == "PharmGKB.Subject.ID"] <- "subject_id"
names(iwpc_data)[names(iwpc_data) == "PharmGKB.Sample.ID"] <- "sample_id"
names(iwpc_data)[names(iwpc_data) == "Project.Site"] <- "project_site"
names(iwpc_data)[names(iwpc_data) == "Gender"] <- "gender"
names(iwpc_data)[names(iwpc_data) == "Race..Reported."] <- "race_reported"
names(iwpc_data)[names(iwpc_data) == "Race..OMB."] <- "race_omb"
names(iwpc_data)[names(iwpc_data) == "Ethnicity..Reported."] <- "ethnicity_reported"
names(iwpc_data)[names(iwpc_data) == "Ethnicity..OMB."] <- "ethnicity_omb"
names(iwpc_data)[names(iwpc_data) == "Age"] <- "age"
names(iwpc_data)[names(iwpc_data) == "Height..cm."] <- "height"
names(iwpc_data)[names(iwpc_data) == "Weight..kg."] <- "weight"
names(iwpc_data)[names(iwpc_data) == "Indication.for.Warfarin.Treatment"] <- "indicatio
n"
names(iwpc_data)[names(iwpc_data) == "Comorbidities"] <- "comorbidities"
names(iwpc_data)[names(iwpc_data) == "Medications"] <- "medications"
names(iwpc_data)[names(iwpc_data) == "Target.INR"] <- "target_inr"
names(iwpc_data)[names(iwpc_data) == "Estimated.Target.INR.Range.Based.on.Indication"] <
- "target_inr_estimated"
names(iwpc_data)[names(iwpc_data) == "Subject.Reached.Stable.Dose.of.Warfarin"] <- "reac
hed_stable_dose"
names(iwpc_data)[names(iwpc_data) == "Therapeutic.Dose.of.Warfarin"] <- "therapeutic_war
farin_dose"
names(iwpc_data)[names(iwpc_data) == "INR.on.Reported.Therapeutic.Dose.of.Warfarin"] <-
"inr_on_warfarin"
names(iwpc_data)[names(iwpc_data) == "Current.Smoker"] <- "smoker"
names(iwpc_data)[names(iwpc_data) == "VKORC1..1639.consensus"] <- "VKORC1"
names(iwpc_data)[names(iwpc_data) == "CYP2C9.consensus"] <- "CYP2C9"
```

Next Excel date error for target_inr_estimated and age variables are fixed. The age variable holds age as the number of decades for that patient.

```
# Fix Excel date error for target_inr_estimated
levels(iwpc_data$target_inr_estimated)[levels(iwpc_data$target_inr_estimated)=="3-Feb"]
 <- "2-3"
iwpc_data$target_inr_estimated[iwpc_data$target_inr_estimated == levels(iwpc_data$target
_inr_estimated)[7]] <- levels(iwpc_data$target_inr_estimated)[6]
iwpc_data$target_inr_estimated <- droplevels(iwpc_data$target_inr_estimated)
iwpc_data$target_inr_estimated <- factor(iwpc_data$target_inr_estimated, levels=c("1.7-
2.8", "1.7-3.3", "2-3", "2-3.5", "2.5-3.5", "3.0-4.0"))
summary(iwpc_data$target_inr_estimated)
```

```
## 1.7-2.8 1.7-3.3     2-3   2-3.5 2.5-3.5 3.0-4.0    NA's
##     263     250    2656     436     209       4    1882
```

```
# Fix Age format
levels(iwpc_data$age)[levels(iwpc_data$age)=="19-Oct"] <- "10 - 19"
summary(iwpc_data$age)
```

```
## 10 - 19 20 - 29 30 - 39 40 - 49 50 - 59 60 - 69 70 - 79 80 - 89     90+
##      14     130     230     540    1085    1384    1570     670      35
##    NA's
##      42
```

For medications the patient is taking, the following two Boolean variables (amiodarone_bool and enzyme_inducer_bool) are created to indicate a) whether the patient is taking Amiodarone, and b) whether the patient is taking an Enzyme Inducer (rifampin, carbamazepine, phenytoin or rifampicin).

```
# Create Boolean column for Amiodarone
iwpc_data$amiodarone_bool <- ifelse(!is.na(iwpc_data$medications) & str_detect(iwpc_data
$medications, "amiodarone"),
                                    yes = 1,
                                    no = 0)
iwpc_data$amiodarone_bool[str_detect(iwpc_data$medications, "not amiodarone")==T] = 0
iwpc_data$amiodarone_bool[str_detect(iwpc_data$medications, "no amiodarone")==T] = 0
count(iwpc_data, iwpc_data$amiodarone_bool)
```

```
## # A tibble: 2 x 2
##   `iwpc_data$amiodarone_bool`     n
##                         <dbl> <int>
## 1                           0  5510
## 2                           1.00   190
```

```
# Create Boolean column for rifampin, carbamazepine, phenytoin or rifampicin
iwpc_data$enzyme_inducer_bool <- ifelse(!is.na(iwpc_data$medications) & str_detect(iwpc_
data$medications, "rifampin|carbamazepine|phenytoin|rifampicin"),
                                    yes = 1,
                                    no = 0)
iwpc_data$enzyme_inducer_bool[str_detect(iwpc_data$medications, "not rifampin|not carbam
azepine|not phenytoin|not rifampicin")==T] = 0
iwpc_data$enzyme_inducer_bool[str_detect(iwpc_data$medications, "no rifampin|no carbamaz
epine|no phenytoin|no rifampicin")==T] = 0
count(iwpc_data, iwpc_data$enzyme_inducer_bool)
```

```
## # A tibble: 2 x 2
##   `iwpc_data$enzyme_inducer_bool`     n
##                             <dbl> <int>
## 1                             0     5676
## 2                             1.00    24
```

Data Exploration

Only the key variables are chosen for further data analysis.

```
iwpc_df <- select(iwpc_data, gender, race_omb, ethnicity_omb, age, height, weight, targe
t_inr, target_inr_estimated, therapeutic_warfarin_dose, smoker, CYP2C9, VKORC1, amiodaro
ne_bool, enzyme_inducer_bool)
```

The dataframe is further divided into two subsets: a subset for patients with high warfarin dosage (0.2 or more standard deviations above the mean) and a subset for patients with low warfarin dosage (0.2 or more standard deviations below the mean).

```
summary(iwpc_df$therapeutic_warfarin_dose)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    2.10   19.53   28.00   30.98   38.50  315.00     172
```

```
sd_cutoff <- 0.2 * sd(iwpc_df$therapeutic_warfarin_dose, na.rm = TRUE)
mean_val <- mean(iwpc_df$therapeutic_warfarin_dose, na.rm = TRUE)
high <- mean_val + sd_cutoff
low <- mean_val - sd_cutoff

# create dosage_level column
iwpc_df$dosage_level <- ifelse(iwpc_df$therapeutic_warfarin_dose >= high, "high", NA)
iwpc_df$dosage_level <- ifelse(iwpc_df$therapeutic_warfarin_dose <= low, "low", iwpc_df
$dosage_level)
iwpc_df$dosage_level <- ifelse(iwpc_df$therapeutic_warfarin_dose < high & iwpc_df$therap
eutic_warfarin_dose > low, "med", iwpc_df$dosage_level)
count(iwpc_df, iwpc_df$dosage_level)
```

```
## # A tibble: 4 x 2
##   `iwpc_df$dosage_level`     n
##   <chr>                  <int>
## 1 high                    2044
## 2 low                     2654
## 3 med                      830
## 4 <NA>                     172
```

```
# Subset dataset based on high warfarin dosage (0.2 or more standard deviations above th
e mean) and low warfarin dosage (0.2 or more standard deviations below the mean)
high_wafarin_dosage <- filter(iwpc_df, therapeutic_warfarin_dose >= high)
low_wafarin_dosage <- filter(iwpc_df, therapeutic_warfarin_dose <= low)

summary(high_wafarin_dosage)
```

```
##     gender                            race_omb
##  female: 749   Asian                       : 219
##  male  :1295   Black or African American: 301
##                Unknown                     : 195
##                White                       :1329
##
##
##
##                 ethnicity_omb       age          height
##  Hispanic or Latino    :  19   60 - 69:503   Min.   :140.5
##  not Hispanic or Latino:1606   50 - 59:482   1st Qu.:165.1
##  Unknown               : 419   70 - 79:406   Median :172.7
##                                40 - 49:274   Mean   :172.3
##                                80 - 89:154   3rd Qu.:180.3
##                                (Other):223   Max.   :202.0
##                                NA's   :  2   NA's   :393
##      weight         target_inr    target_inr_estimated
##  Min.   : 38.00   Min.   :1.750   1.7-2.8:  62
##  1st Qu.: 71.17   1st Qu.:2.500   1.7-3.3:  23
##  Median : 84.00   Median :2.500   2-3    :1071
##  Mean   : 87.56   Mean   :2.552   2-3.5  : 191
##  3rd Qu.: 99.80   3rd Qu.:2.500   2.5-3.5: 101
##  Max.   :237.70   Max.   :3.500   3.0-4.0:   1
##  NA's   :80       NA's   :1389    NA's   : 595
##  therapeutic_warfarin_dose     smoker           CYP2C9      VKORC1
##  Min.   : 34.5             Min.   :0.0000   *1/*1  :1624   A/A :115
##  1st Qu.: 37.5             1st Qu.:0.0000   *1/*2  : 259   A/G :521
##  Median : 42.5             Median :0.0000   *1/*3  :  87   G/G :795
##  Mean   : 47.5             Mean   :0.1721   *2/*2  :  13   NA's:613
##  3rd Qu.: 52.5             3rd Qu.:0.0000   *2/*3  :   5
##  Max.   :315.0             Max.   :1.0000   (Other):  10
##                            NA's   :591      NA's   :  46
##  amiodarone_bool   enzyme_inducer_bool dosage_level
##  Min.   :0.00000   Min.   :0.00000     Length:2044
##  1st Qu.:0.00000   1st Qu.:0.00000     Class :character
##  Median :0.00000   Median :0.00000     Mode  :character
##  Mean   :0.02642   Mean   :0.00636
##  3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000
##
```

```
summary(low_wafarin_dosage)
```

```
##     gender                            race_omb
##  female:1176   Asian                       :1173
##  male  :1474   Black or African American:  95
##  NA's  :   4   Unknown                     : 212
##                White                       :1174
##
##
##
##                 ethnicity_omb       age          height
##  Hispanic or Latino    :  16   70 - 79:887   Min.   :125.0
##  not Hispanic or Latino:2205   60 - 69:657   1st Qu.:157.5
##  Unknown               : 433   50 - 59:402   Median :164.6
##                                80 - 89:402   Mean   :164.7
##                                40 - 49:166   3rd Qu.:172.0
##                                (Other):105   Max.   :195.6
##                                NA's   : 35   NA's   :539
##      weight         target_inr    target_inr_estimated
##  Min.   : 30.00   Min.   :1.300   1.7-2.8: 121
##  1st Qu.: 57.92   1st Qu.:2.500   1.7-3.3: 215
##  Median : 68.00   Median :2.500   2-3    :1048
##  Mean   : 70.27   Mean   :2.521   2-3.5  : 181
##  3rd Qu.: 80.00   3rd Qu.:2.500   2.5-3.5:  72
##  Max.   :177.30   Max.   :3.500   3.0-4.0:   1
##  NA's   :162      NA's   :2269    NA's   :1016
##  therapeutic_warfarin_dose     smoker           CYP2C9      VKORC1
##  Min.   : 2.10             Min.   :0.0000   *1/*1  :1818   A/A :1172
##  1st Qu.:14.00             1st Qu.:0.0000   *1/*2  : 338   A/G : 616
##  Median :18.93             Median :0.0000   *1/*3  : 328   G/G : 216
##  Mean   :18.50             Mean   :0.0969   *2/*3  :  56   NA's: 650
##  3rd Qu.:22.88             3rd Qu.:0.0000   *2/*2  :  34
##  Max.   :27.51             Max.   :1.0000   (Other):  24
##                            NA's   :1570     NA's   :  56
##  amiodarone_bool   enzyme_inducer_bool dosage_level
##  Min.   :0.00000   Min.   :0.000000    Length:2654
##  1st Qu.:0.00000   1st Qu.:0.000000    Class :character
##  Median :0.00000   Median :0.000000    Mode  :character
##  Mean   :0.04371   Mean   :0.002638
##  3rd Qu.:0.00000   3rd Qu.:0.000000
##  Max.   :1.00000   Max.   :1.000000
##
```

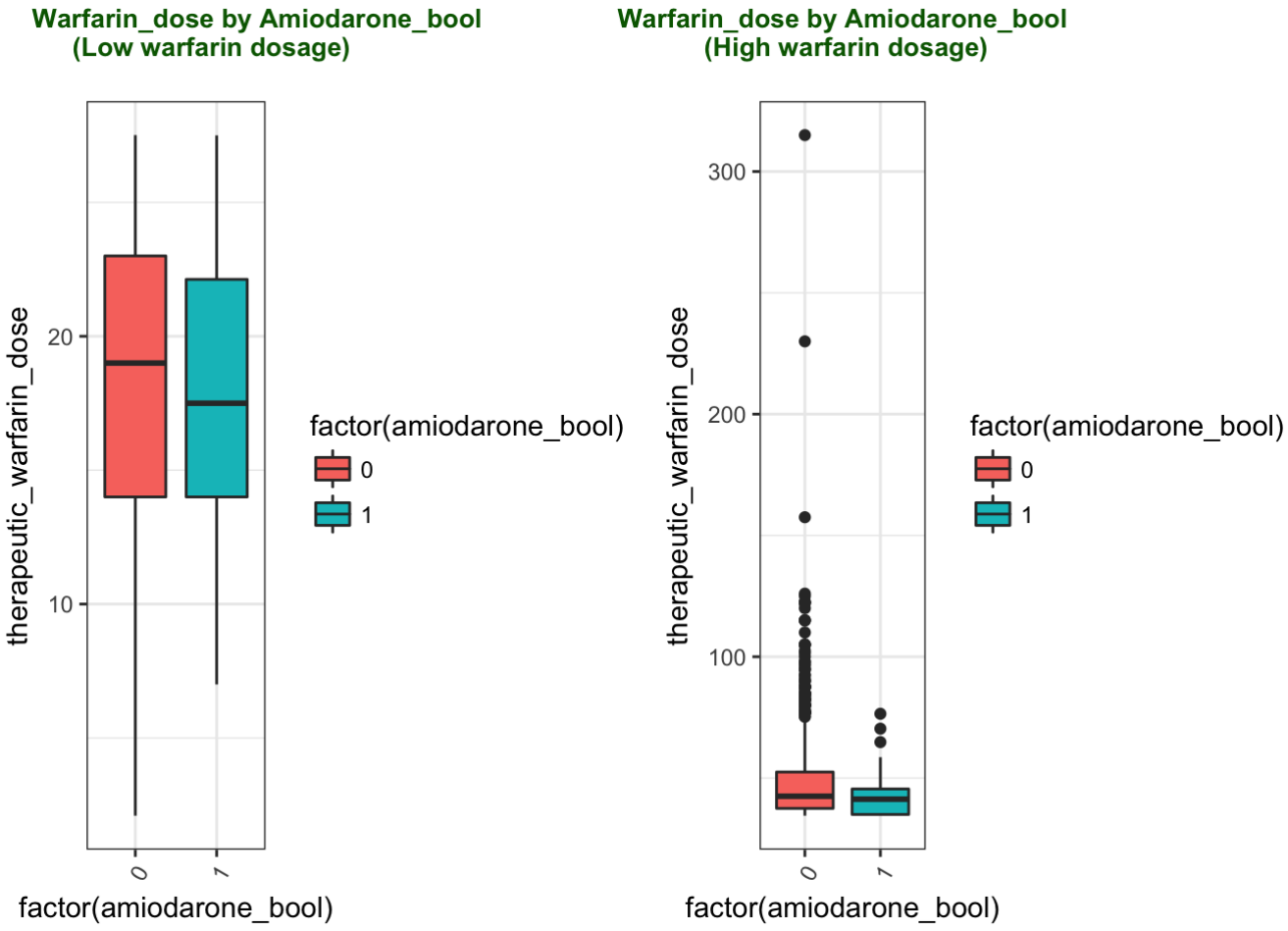
The high and low warfarin dosage subsets are then combined into one dataframe as iwpc_hl.

```
iwpc_hl <- filter(iwpc_df, dosage_level == "high" | dosage_level == "low")
```

Data Visualization

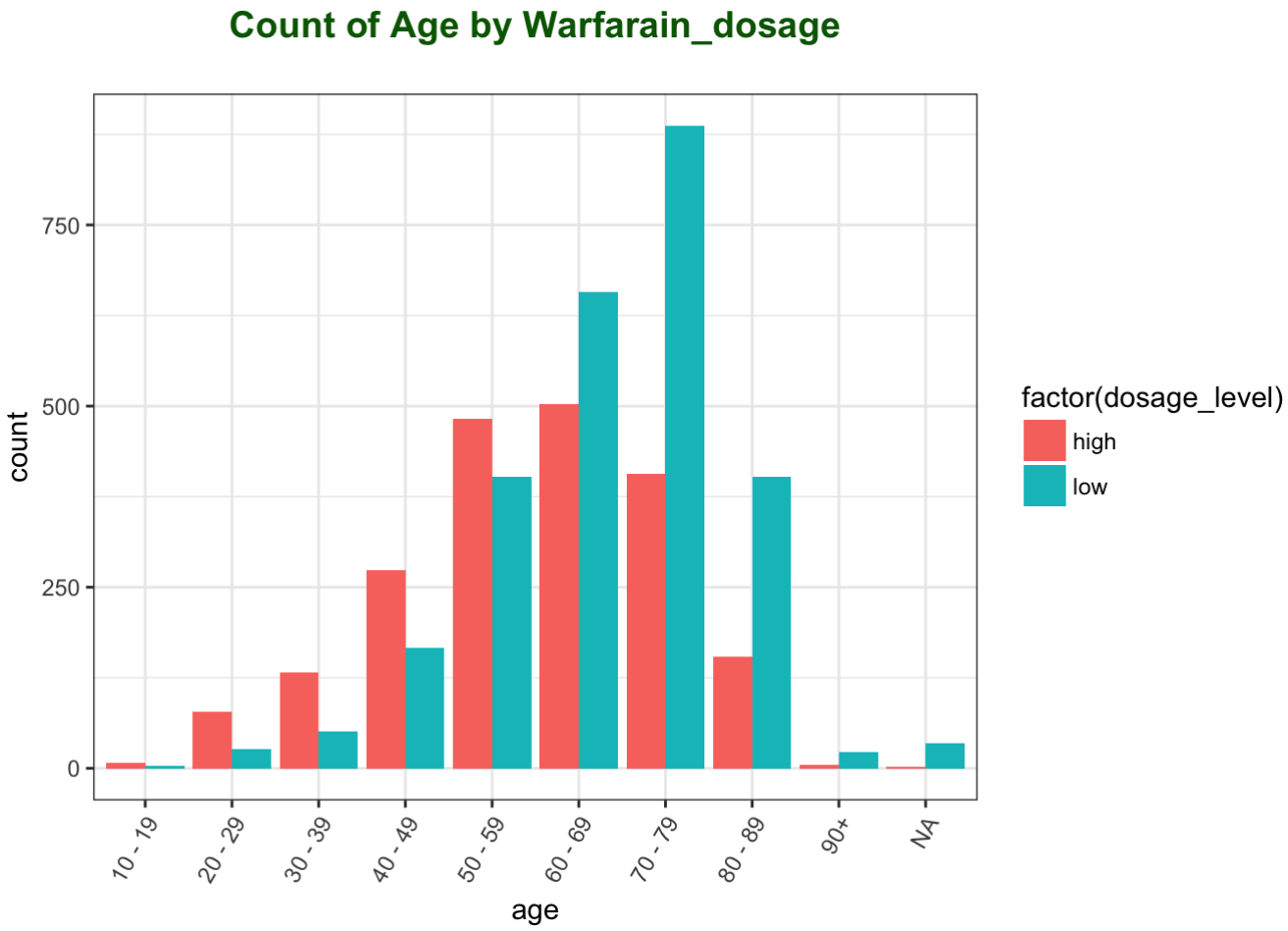
The following plots are created to visualize the data.

```
library(ggplot2)
library(cowplot)
theme_set(theme_cowplot(font_size=10))

# Boxplot of Current Smoker by Gender
high_dose_amio <- ggplot(high_wafarin_dosage, aes(x=factor(amiodarone_bool), y=therapeut
ic_warfarin_dose)) + # categorical variable on x-axis
  geom_boxplot(aes(fill = factor(amiodarone_bool))) +
  labs(title = "Warfarin_dose by Amiodarone_bool\n (High warfarin dosage)\n") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  theme(plot.title=element_text(hjust=0.5, size=10, face="bold", color="darkgreen"))
low_dose_amio <- ggplot(low_wafarin_dosage, aes(x=factor(amiodarone_bool), y=therapeutic
_warfarin_dose)) + # categorical variable on x-axis
  geom_boxplot(aes(fill = factor(amiodarone_bool))) +
  labs(title = "Warfarin_dose by Amiodarone_bool\n (Low warfarin dosage)\n") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  theme(plot.title=element_text(hjust=0.2, size=10, face="bold", color="darkgreen"))
plot_grid(low_dose_amio, high_dose_amio)
```



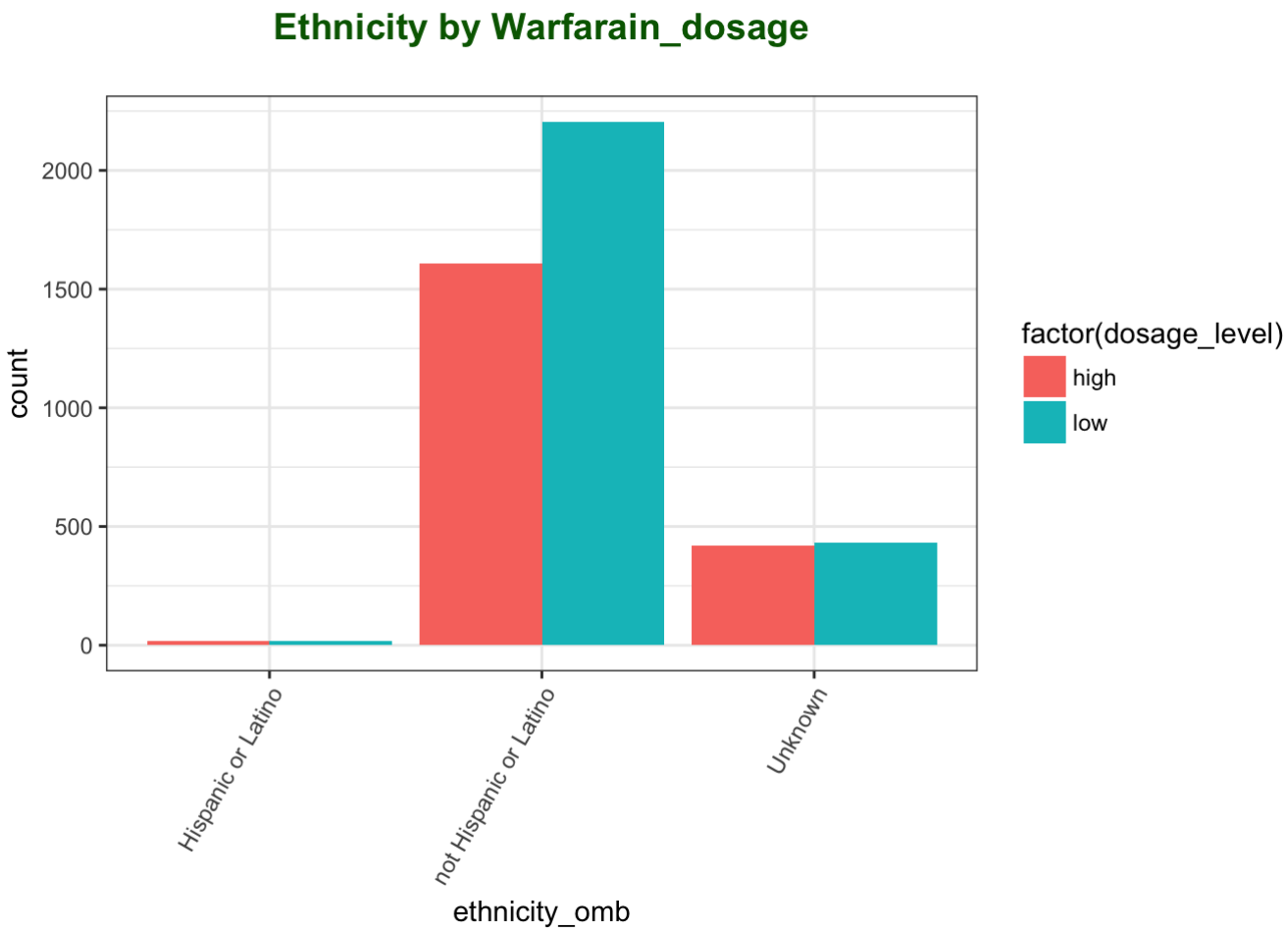
The patients not taking any Amiodarone seem to get higher dose for Warfarin.

```
# Count of Age by warfarin_dosage
age_dosage <- ggplot(iwpc_hl, aes(x=age)) +
  geom_bar(aes(fill = factor(dosage_level)), position = "dodge") +
  labs(title = "Count of Age by Warfarain_dosage\n") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  theme(plot.title=element_text(hjust=0.5, size=14, face="bold", color="darkgreen"))
age_dosage
```



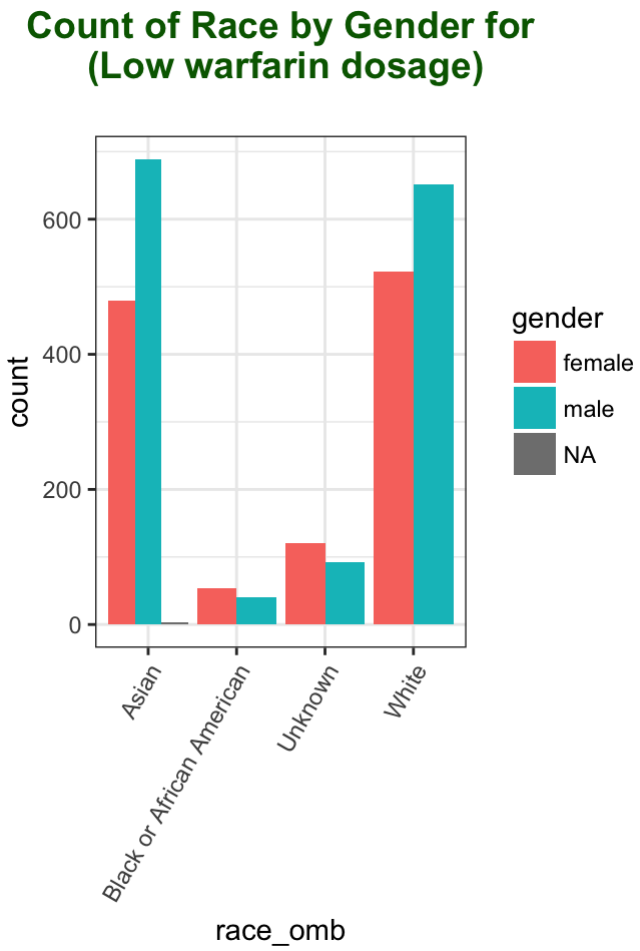
Generally, more older patients (>60-year-old) are prescribed lower dose of Warfarin, while more younger patients are prescribed higher dose of Warfarin.

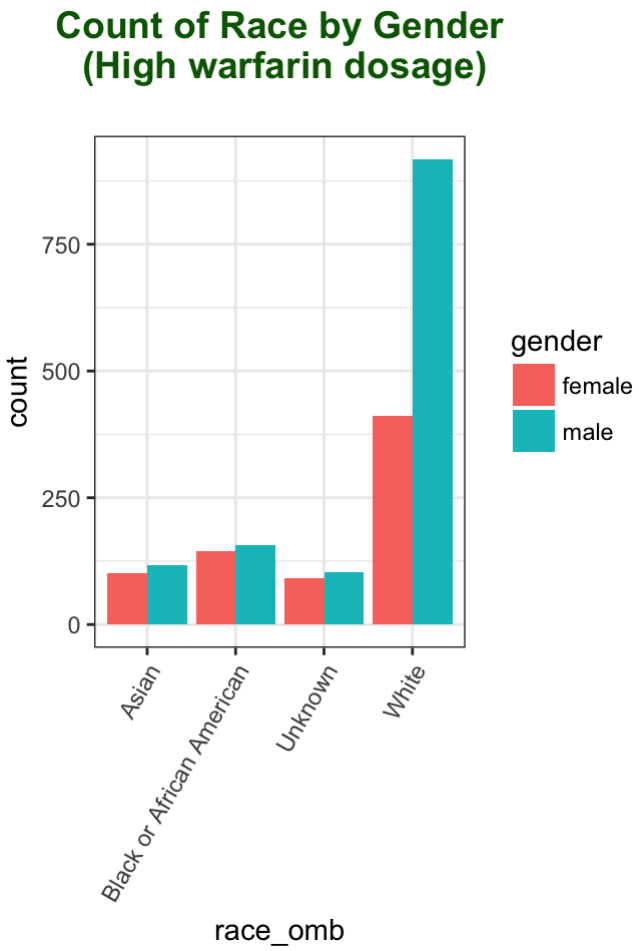
```
# Ethnicity by warfarin_dosage
ethnicity_dosage <- ggplot(iwpc_hl, aes(x=ethnicity_omb)) +
  geom_bar(aes(fill = factor(dosage_level)), position = "dodge") +
  labs(title = "Ethnicity by Warfarain_dosage\n") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  theme(plot.title=element_text(hjust=0.5, size=14, face="bold", color="darkgreen"))
ethnicity_dosage
```



More patients from Not Hispanic or Latino ethinicity has lower dose for Warfarin, while the other ethinicities generally have similar counts for patients for high and low dosage.

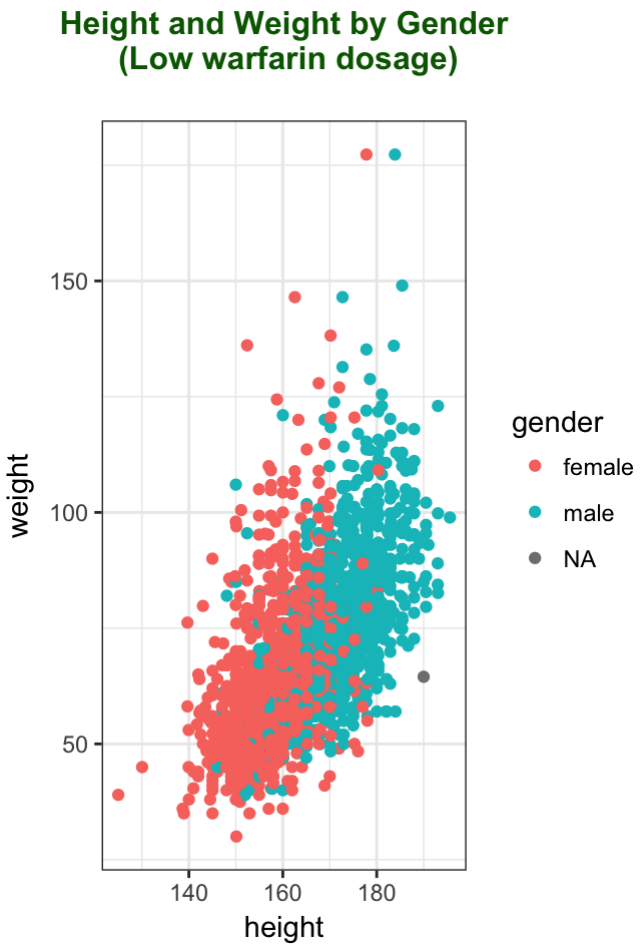
```
# Visualize the count of race by gender
high_gender <- ggplot(high_wafarin_dosage, aes(x=race_omb)) +
  geom_bar(aes(fill = gender), position = "dodge") +
  labs(title = "Count of Race by Gender\n (High warfarin dosage)\n") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  theme(plot.title=element_text(hjust=0.5, size=14, face="bold", color="darkgreen"))
low_gender <- ggplot(low_wafarin_dosage, aes(x=race_omb)) +
  geom_bar(aes(fill = gender), position = "dodge") +
  labs(title = "Count of Race by Gender for\n (Low warfarin dosage)\n") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  theme(plot.title=element_text(hjust=0.5, size=14, face="bold", color="darkgreen"))
plot_grid(low_gender, high_gender)
```

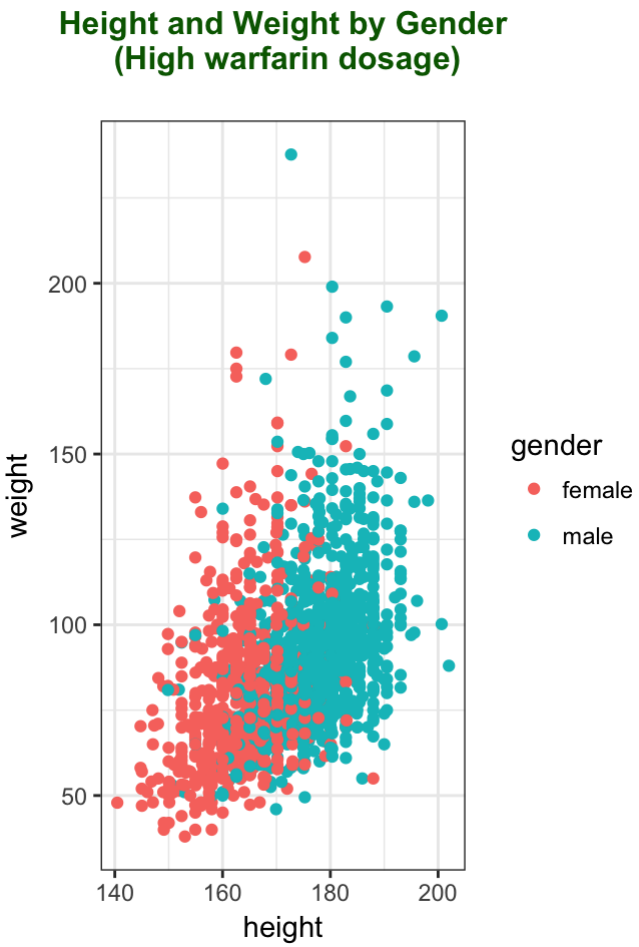




There are more males than females in this dataset, especially more Asian and White males in Low dosage population, and a lot more White males in High dosage population.

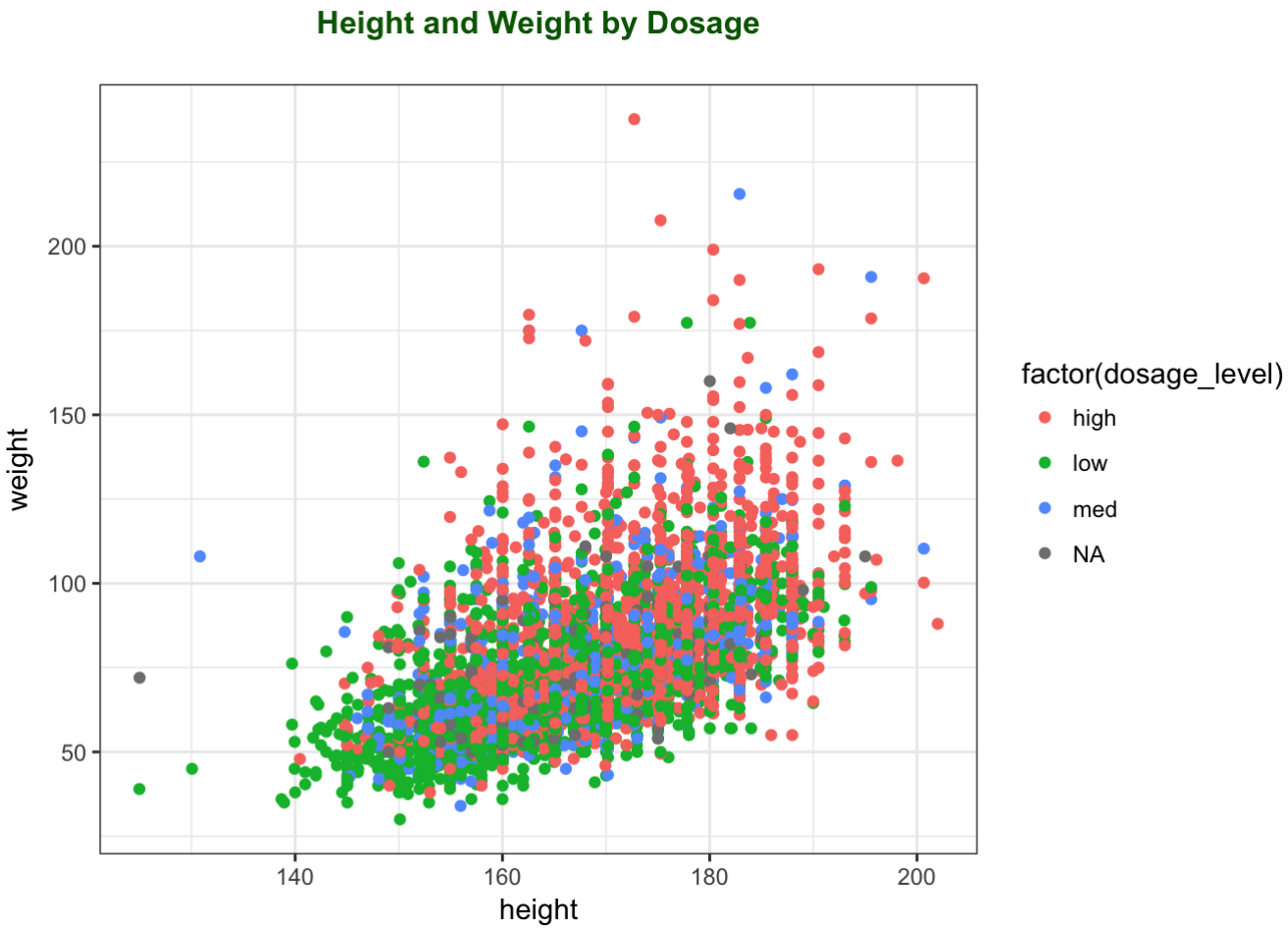
```
# Scatterplot of Height and Weight by Gender
high_weight <- ggplot(high_wafarin_dosage, aes(x=height,y=weight)) +
  geom_point(aes(fill=gender, color=gender), size=1.5) +
  labs(title="Height and Weight by Gender\n (High warfarin dosage)\n") +
  theme_bw() +
  theme(plot.title=element_text(hjust=0.5, size=12, face="bold", color="darkgreen"))
low_weight <- ggplot(low_wafarin_dosage, aes(x=height,y=weight)) +
  geom_point(aes(fill=gender, color=gender), size=1.5) +
  labs(title="Height and Weight by Gender\n (Low warfarin dosage)\n") +
  theme_bw() +
  theme(plot.title=element_text(hjust=0.5, size=12, face="bold", color="darkgreen"))
plot_grid(low_weight, high_weight)
```





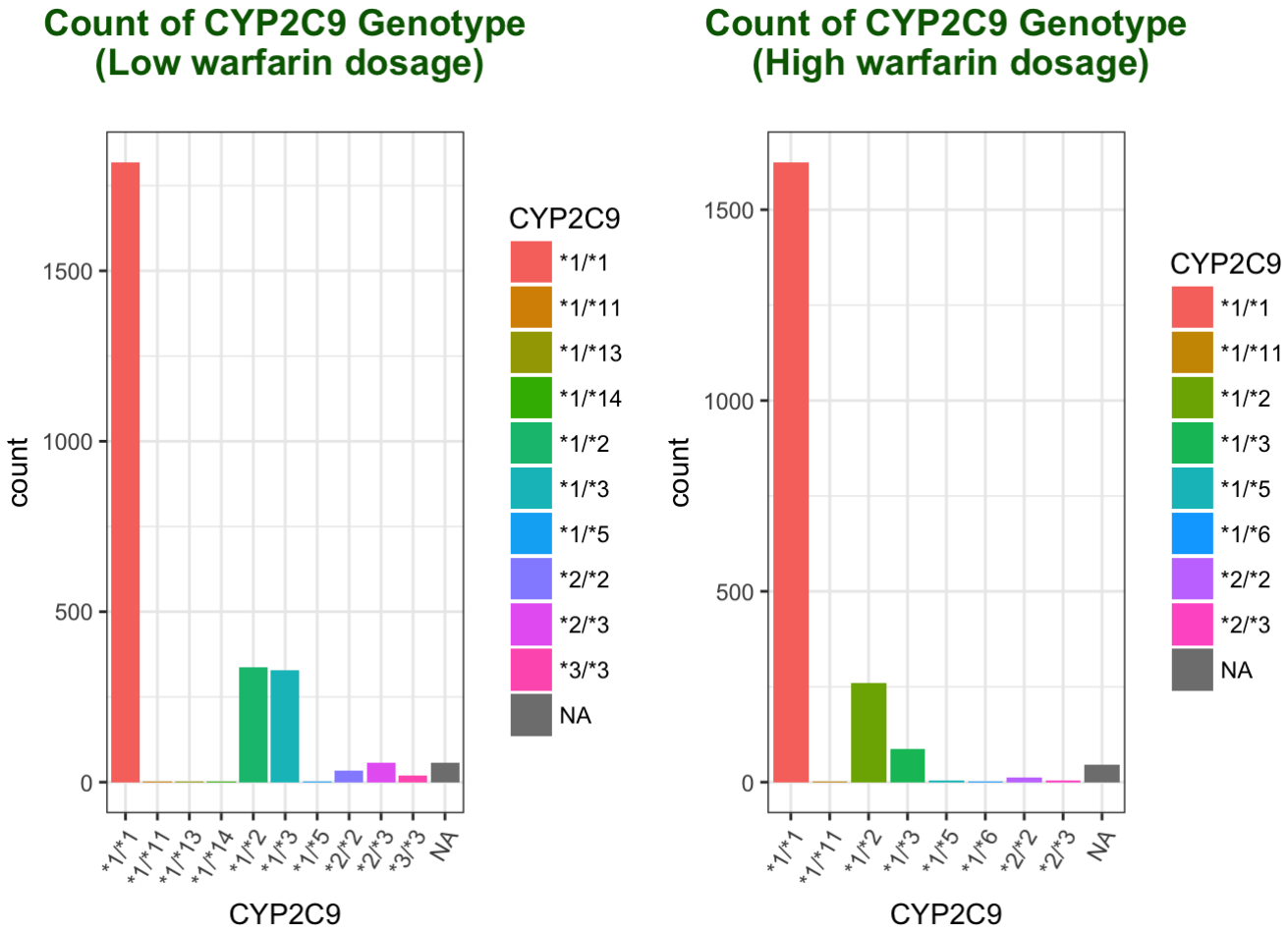
Males are generally taller and heavier in comparison to females.

```
# Scatter of Height and Weight by Dosage
height_weight <- ggplot(iwpc_df, aes(x=height,y=weight)) +
  geom_point(aes(fill=factor(dosage_level), color=factor(dosage_level)), size=1.5) +
  labs(title="Height and Weight by Dosage\n") +
  theme_bw() +
  theme(plot.title=element_text(hjust=0.5, size=12, face="bold", color="darkgreen"))
height_weight
```



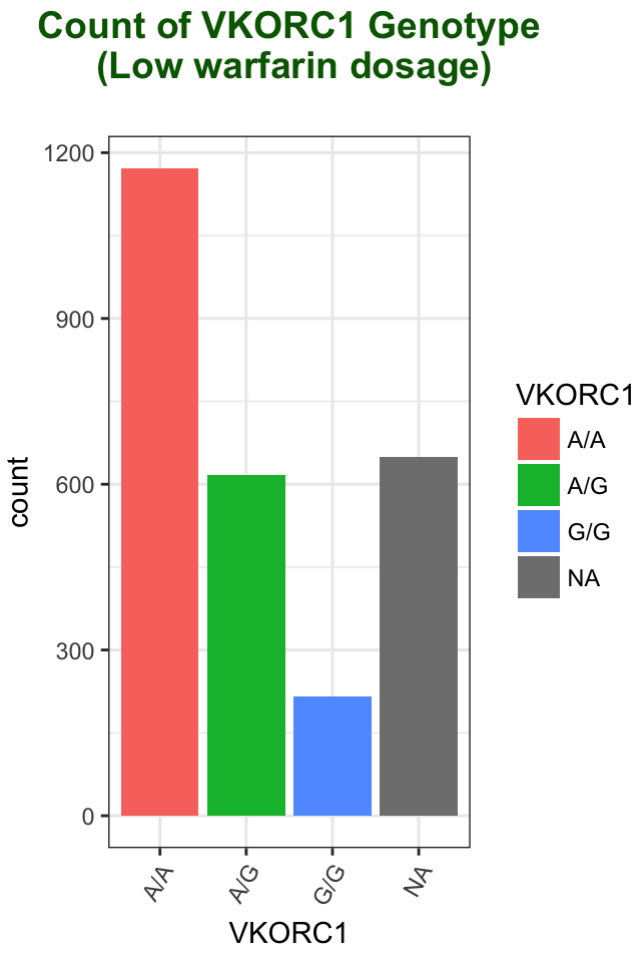
There is no clear pattern for Warfarin dosage by height and weight. There is only a slight pattern for those with shorter height and lower weight have lower Warfarin dosage and those with higher height and heavier weight have higher dosage.

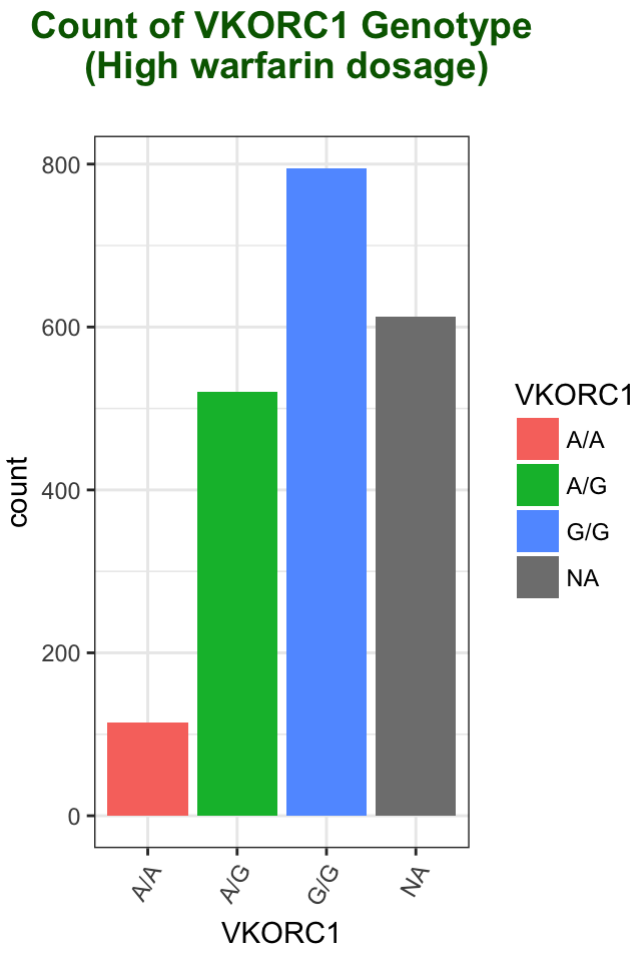
```
# Visualize the count of CYP2C9 Genotype
high_CYP2C9 <- ggplot(high_wafarin_dosage, aes(x=CYP2C9)) +
  geom_bar(aes(fill = CYP2C9)) +
  labs(title = "Count of CYP2C9 Genotype\n (High warfarin dosage)\n") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  theme(plot.title=element_text(hjust=0.5, size=14, face="bold", color="darkgreen"))
low_CYP2C9 <- ggplot(low_wafarin_dosage, aes(x=CYP2C9)) +
  geom_bar(aes(fill = CYP2C9)) +
  labs(title = "Count of CYP2C9 Genotype\n (Low warfarin dosage)\n") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  theme(plot.title=element_text(hjust=0.5, size=14, face="bold", color="darkgreen"))
plot_grid(low_CYP2C9, high_CYP2C9)
```



Patients who have 1 copy of the `CYP2C9*1/*2` SNP are slow metabolizers of S-warfarin; patients who are homozygous for `CYP2C9*2/*2` or who carry at least 1 copy of the `CYP2C9*3`, `CYP2C9*5` or `CYP2C9*6` SNP are very slow metabolizers. Some patients with `CYP2C9*1/*2` SNP are in High dosage population, but those with `CYP2C9*3`, `CYP2C9*5` or `CYP2C9*6` SNP are only in Low dosage population.

```
# Visualize the count of VKORC1 Genotype
high_VKORC1 <- ggplot(high_wafarin_dosage, aes(x=VKORC1)) +
  geom_bar(aes(fill = VKORC1)) +
  labs(title = "Count of VKORC1 Genotype\n (High warfarin dosage)\n") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  theme(plot.title=element_text(hjust=0.5, size=14, face="bold", color="darkgreen"))
low_VKORC1 <- ggplot(low_wafarin_dosage, aes(x=VKORC1)) +
  geom_bar(aes(fill = VKORC1)) +
  labs(title = "Count of VKORC1 Genotype\n (Low warfarin dosage)\n") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  theme(plot.title=element_text(hjust=0.5, size=14, face="bold", color="darkgreen"))
plot_grid(low_VKORC1, high_VKORC1)
```





Patients who have the AA genotype (or AA haplotype) are the most warfarin sensitive and therefore often require lower warfarin doses. The plots show that the majority of patients AA genotype are in Low dosage population.

Regression

Two regression models are fitted: one with therapeutic warfarin dose as the response variable, and the other one with square root of therapeutic warfarin dose. The dataframe for regression models is created with only the key variables.

```
iwpc <- select(iwpc_df, gender, race_omb, ethnicity_omb, age, height, weight,
                         therapeutic_warfarin_dose, CYP2C9, VKORC1, amiodarone_bool, en
zyme_inducer_bool)
summary(iwpc)
```

```
##     gender                            race_omb
##  female:2373   Asian                       :1634
##  male  :3323   Black or African American: 462
##  NA's  :   4   Unknown                     : 482
##                White                       :3122
##
##
##
##                  ethnicity_omb       age           height
##  Hispanic or Latino    :  45   70 - 79:1570   Min.   :125.0
##  not Hispanic or Latino:4524   60 - 69:1384   1st Qu.:160.0
##  Unknown               :1131   50 - 59:1085   Median :167.9
##                                80 - 89: 670   Mean   :168.0
##                                40 - 49: 540   3rd Qu.:176.0
##                                (Other): 409   Max.   :202.0
##                                NA's   :  42   NA's   :1146
##      weight       therapeutic_warfarin_dose     CYP2C9      VKORC1
##  Min.   : 30.00   Min.   :  2.10            *1/*1  :4157   A/A :1485
##  1st Qu.: 62.00   1st Qu.: 19.53            *1/*2  : 737   A/G :1470
##  Median : 75.00   Median : 28.00            *1/*3  : 498   G/G :1246
##  Mean   : 77.85   Mean   : 30.98            *2/*3  :  69   NA's:1499
##  3rd Qu.: 90.00   3rd Qu.: 38.50            *2/*2  :  56
##  Max.   :237.70   Max.   :315.00            (Other):  39
##  NA's   :287      NA's   :172               NA's   : 144
##  amiodarone_bool   enzyme_inducer_bool
##  Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000
##  Mean   :0.03333   Mean   :0.00421
##  3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000
##
```

```
iwpc <- na.omit(iwpc)
```

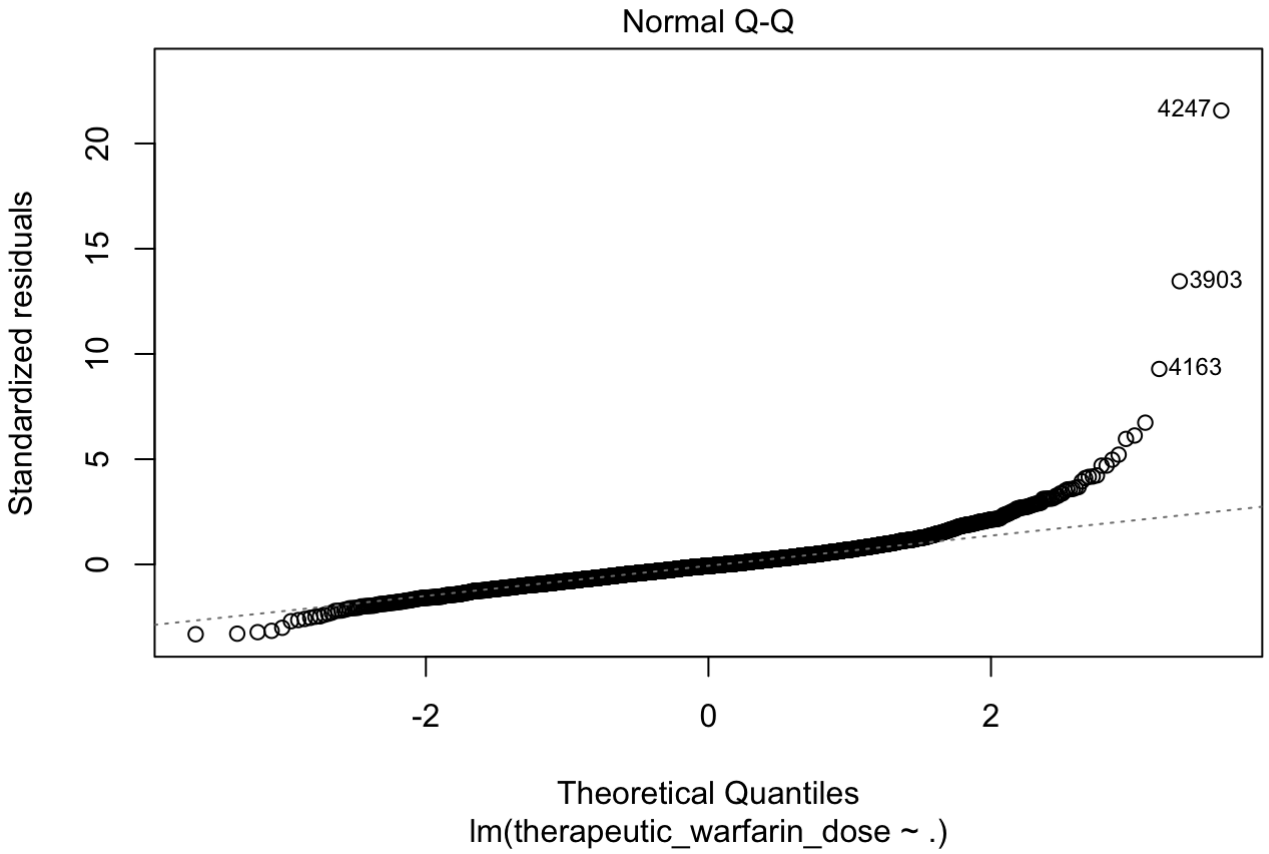
First, the full model is fitted with all variables as the predictor variables.

```
fit_full <- lm(therapeutic_warfarin_dose ~ ., data = iwpc)
summary(fit_full)
```

```
##
## Call:
## lm(formula = therapeutic_warfarin_dose ~ ., data = iwpc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.212  -7.001  -0.922   5.373 275.962
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        -1.76306    6.47293  -0.272 0.785350
## gendermale                         -1.69280    0.60737  -2.787 0.005347
## race_ombBlack or African American  -1.52040    1.13061  -1.345 0.178789
## race_ombUnknown                     1.19053    1.33452   0.892 0.372398
## race_ombWhite                       1.91253    0.79368   2.410 0.016017
## ethnicity_ombnot Hispanic or Latino  4.06255    2.24260   1.812 0.070143
## ethnicity_ombUnknown                0.17873    2.26184   0.079 0.937023
## age20 - 29                          0.20163    3.96301   0.051 0.959425
## age30 - 39                         -0.16644    3.86091  -0.043 0.965618
## age40 - 49                         -2.24948    3.78339  -0.595 0.552171
## age50 - 59                         -5.90574    3.74695  -1.576 0.115082
## age60 - 69                         -9.41832    3.74627  -2.514 0.011980
## age70 - 79                        -12.49768    3.74726  -3.335 0.000861
## age80 - 89                        -13.88232    3.78817  -3.665 0.000251
## age90+                            -14.97265    4.92848  -3.038 0.002399
## height                              0.10653    0.03294   3.234 0.001233
## weight                              0.17251    0.01356  12.723  < 2e-16
## CYP2C9*1/*11                      -18.20405    9.09558  -2.001 0.045425
## CYP2C9*1/*13                      -10.51073   12.86307  -0.817 0.413913
## CYP2C9*1/*14                      -11.75009   12.85606  -0.914 0.360794
## CYP2C9*1/*2                        -6.23668    0.69381  -8.989  < 2e-16
## CYP2C9*1/*3                        -9.33186    0.80397 -11.607  < 2e-16
## CYP2C9*1/*5                       -14.68013    5.79591  -2.533 0.011358
## CYP2C9*1/*6                        -7.69240    9.24668  -0.832 0.405517
## CYP2C9*2/*2                       -12.42650    2.17413  -5.716 1.18e-08
## CYP2C9*2/*3                       -19.80864    2.06847  -9.576  < 2e-16
## CYP2C9*3/*3                       -21.51590    4.08703  -5.264 1.49e-07
## VKORC1A/G                           9.05863    0.65243  13.884  < 2e-16
## VKORC1G/G                          19.52337    0.74745  26.120  < 2e-16
## amiodarone_bool                    -8.01914    1.20714  -6.643 3.55e-11
## enzyme_inducer_bool                13.72418    3.10312   4.423 1.00e-05
##
## (Intercept)
## gendermale                         **
## race_ombBlack or African American
## race_ombUnknown
## race_ombWhite                      *
## ethnicity_ombnot Hispanic or Latino .
## ethnicity_ombUnknown
## age20 - 29
## age30 - 39
## age40 - 49
## age50 - 59
```

```
## age60 - 69                          *
## age70 - 79                          ***
## age80 - 89                          ***
## age90+                              **
## height                              **
## weight                              ***
## CYP2C9*1/*11                        *
## CYP2C9*1/*13
## CYP2C9*1/*14
## CYP2C9*1/*2                         ***
## CYP2C9*1/*3                         ***
## CYP2C9*1/*5                         *
## CYP2C9*1/*6
## CYP2C9*2/*2                         ***
## CYP2C9*2/*3                         ***
## CYP2C9*3/*3                         ***
## VKORC1A/G                           ***
## VKORC1G/G                           ***
## amiodarone_bool                     ***
## enzyme_inducer_bool                 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.84 on 3491 degrees of freedom
## Multiple R-squared:  0.4573, Adjusted R-squared:  0.4527
## F-statistic: 98.06 on 30 and 3491 DF,  p-value: < 2.2e-16
```

```
plot(fit_full, 2)
```



The Normal Q-Q plot for fit_full model is a bit heavy-tailed, the points curve off in the extremities. It means the data have more extreme values than would be expected if they truly came from a Normal distribution.

The summary of fit_full does not show which variables are particularly insignificant, it only shows that some dummy variables are insignificant. So, the step function is used for feature selection.

```
fit_null <- lm(therapeutic_warfarin_dose ~ 1, data = iwpc)
# summary(fit_null)

# step function
fit_step = step(fit_null, scope=list(lower=fit_null, upper=fit_full),direction="both")
```

```
## Start:  AIC=20104.7
## therapeutic_warfarin_dose ~ 1
##
##                         Df Sum of Sq     RSS   AIC
## + VKORC1                 2    289645  771156 18986
## + race_omb               3    183246  877554 19443
## + weight                 1    178451  882350 19458
## + height                 1    103240  957561 19746
## + age                    8     86773  974028 19820
## + CYP2C9                10     26526 1034275 20036
## + enzyme_inducer_bool    1      7361 1053440 20082
## + amiodarone_bool        1      3339 1057462 20096
## + gender                 1      1549 1059252 20102
## <none>                               1060801 20105
## + ethnicity_omb          2       517 1060284 20107
##
## Step:  AIC=18985.57
## therapeutic_warfarin_dose ~ VKORC1
##
##                         Df Sum of Sq     RSS   AIC
## + age                    8     75538  695618 18638
## + weight                 1     66238  704918 18671
## + CYP2C9                10     43709  727447 18800
## + height                 1     30640  740516 18845
## + race_omb               3      9921  761235 18946
## + amiodarone_bool        1      6835  764321 18956
## + ethnicity_omb          2      5299  765856 18965
## + enzyme_inducer_bool    1      4122  767034 18969
## + gender                 1      1953  769203 18979
## <none>                               771156 18986
## - VKORC1                 2    289645 1060801 20105
##
## Step:  AIC=18638.49
## therapeutic_warfarin_dose ~ VKORC1 + age
##
##                         Df Sum of Sq    RSS   AIC
## + weight                 1     49871 645746 18378
## + CYP2C9                10     43434 652184 18431
## + height                 1     21074 674544 18532
## + race_omb               3     15537 680081 18565
## + ethnicity_omb          2      9534 686084 18594
## + amiodarone_bool        1      7249 688369 18604
## + gender                 1      3579 692039 18622
## + enzyme_inducer_bool    1      2283 693335 18629
## <none>                               695618 18638
## - age                    8     75538 771156 18986
## - VKORC1                 2    278410 974028 19820
##
## Step:  AIC=18378.47
## therapeutic_warfarin_dose ~ VKORC1 + age + weight
##
##                         Df Sum of Sq    RSS   AIC
## + CYP2C9                10     47967 597780 18127
```

```
## + amiodarone_bool      1      7614 638133 18339
## + ethnicity_omb        2      7025 638722 18344
## + race_omb             3      3849 641898 18363
## + enzyme_inducer_bool  1      2810 642937 18365
## + height               1      1153 644593 18374
## <none>                             645746 18378
## + gender               1        13 645734 18380
## - weight               1     49871 695618 18638
## - age                  8     59172 704918 18671
## - VKORC1               2    175452 821198 19221
##
## Step:  AIC=18126.63
## therapeutic_warfarin_dose ~ VKORC1 + age + weight + CYP2C9
##
##                       Df Sum of Sq    RSS   AIC
## + amiodarone_bool      1      7884 589895 18082
## + ethnicity_omb        2      6178 591601 18094
## + race_omb             3      5846 591933 18098
## + enzyme_inducer_bool  1      2416 595363 18114
## + height               1      1771 596008 18118
## <none>                             597780 18127
## + gender               1        54 597725 18128
## - CYP2C9              10     47967 645746 18378
## - weight               1     54404 652184 18431
## - age                  8     58530 656310 18440
## - VKORC1               2    188727 786507 19089
##
## Step:  AIC=18081.87
## therapeutic_warfarin_dose ~ VKORC1 + age + weight + CYP2C9 +
##     amiodarone_bool
##
##                       Df Sum of Sq    RSS   AIC
## + ethnicity_omb        2      4906 584989 18056
## + race_omb             3      4972 584923 18058
## + enzyme_inducer_bool  1      3131 586764 18065
## + height               1      1877 588018 18073
## <none>                             589895 18082
## + gender               1        39 589856 18084
## - amiodarone_bool      1      7884 597780 18127
## - CYP2C9              10     48237 638133 18339
## - weight               1     54726 644622 18392
## - age                  8     58670 648566 18400
## - VKORC1               2    191469 781365 19068
##
## Step:  AIC=18056.45
## therapeutic_warfarin_dose ~ VKORC1 + age + weight + CYP2C9 +
##     amiodarone_bool + ethnicity_omb
##
##                       Df Sum of Sq    RSS   AIC
## + race_omb             3      4209 580781 18037
## + enzyme_inducer_bool  1      3034 581956 18040
## + height               1      1391 583599 18050
## <none>                             584989 18056
## + gender               1        84 584905 18058
```

```
## - ethnicity_omb        2      4906 589895 18082
## - amiodarone_bool      1      6612 591601 18094
## - CYP2C9              10     47462 632451 18311
## - weight               1     52383 637373 18356
## - age                  8     61402 646391 18392
## - VKORC1               2    196283 781272 19072
##
## Step:  AIC=18037.02
## therapeutic_warfarin_dose ~ VKORC1 + age + weight + CYP2C9 +
##     amiodarone_bool + ethnicity_omb + race_omb
##
##                       Df Sum of Sq    RSS   AIC
## + enzyme_inducer_bool  1      3224 577556 18019
## + height               1       607 580174 18035
## <none>                             580781 18037
## + gender               1       150 580630 18038
## - race_omb             3      4209 584989 18056
## - ethnicity_omb        2      4143 584923 18058
## - amiodarone_bool      1      6624 587405 18075
## - weight               1     36532 617313 18250
## - CYP2C9              10     51470 632251 18316
## - age                  8     63108 643889 18384
## - VKORC1               2    118025 698805 18685
##
## Step:  AIC=18019.42
## therapeutic_warfarin_dose ~ VKORC1 + age + weight + CYP2C9 +
##     amiodarone_bool + ethnicity_omb + race_omb + enzyme_inducer_bool
##
##                       Df Sum of Sq    RSS   AIC
## + height               1       599 576957 18018
## <none>                             577556 18019
## + gender               1       156 577401 18020
## - enzyme_inducer_bool  1      3224 580781 18037
## - ethnicity_omb        2      4029 581585 18040
## - race_omb             3      4399 581956 18040
## - amiodarone_bool      1      7279 584835 18062
## - weight               1     37075 614632 18236
## - CYP2C9              10     51184 628740 18298
## - age                  8     61536 639092 18360
## - VKORC1               2    117525 695081 18668
##
## Step:  AIC=18017.76
## therapeutic_warfarin_dose ~ VKORC1 + age + weight + CYP2C9 +
##     amiodarone_bool + ethnicity_omb + race_omb + enzyme_inducer_bool +
##     height
##
##                       Df Sum of Sq    RSS   AIC
## + gender               1      1281 575677 18012
## <none>                             576957 18018
## - height               1       599 577556 18019
## - race_omb             3      3605 580563 18034
## - enzyme_inducer_bool  1      3216 580174 18035
## - ethnicity_omb        2      3739 580696 18036
## - amiodarone_bool      1      7363 584320 18060
```

```
## - weight               1     26532 603490 18174
## - CYP2C9              10     51023 627980 18296
## - age                  8     59840 636798 18349
## - VKORC1               2    116893 693851 18664
##
## Step:  AIC=18011.93
## therapeutic_warfarin_dose ~ VKORC1 + age + weight + CYP2C9 +
##     amiodarone_bool + ethnicity_omb + race_omb + enzyme_inducer_bool +
##     height + gender
##
##                       Df Sum of Sq    RSS   AIC
## <none>                             575677 18012
## - gender               1      1281 576957 18018
## - height               1      1724 577401 18020
## - race_omb             3      3304 578980 18026
## - enzyme_inducer_bool  1      3226 578902 18030
## - ethnicity_omb        2      3575 579252 18030
## - amiodarone_bool      1      7277 582954 18054
## - weight               1     26694 602370 18170
## - CYP2C9              10     51226 626902 18292
## - age                  8     54564 630241 18315
## - VKORC1               2    116581 692257 18657
```

The step function returns the following model as the best model.

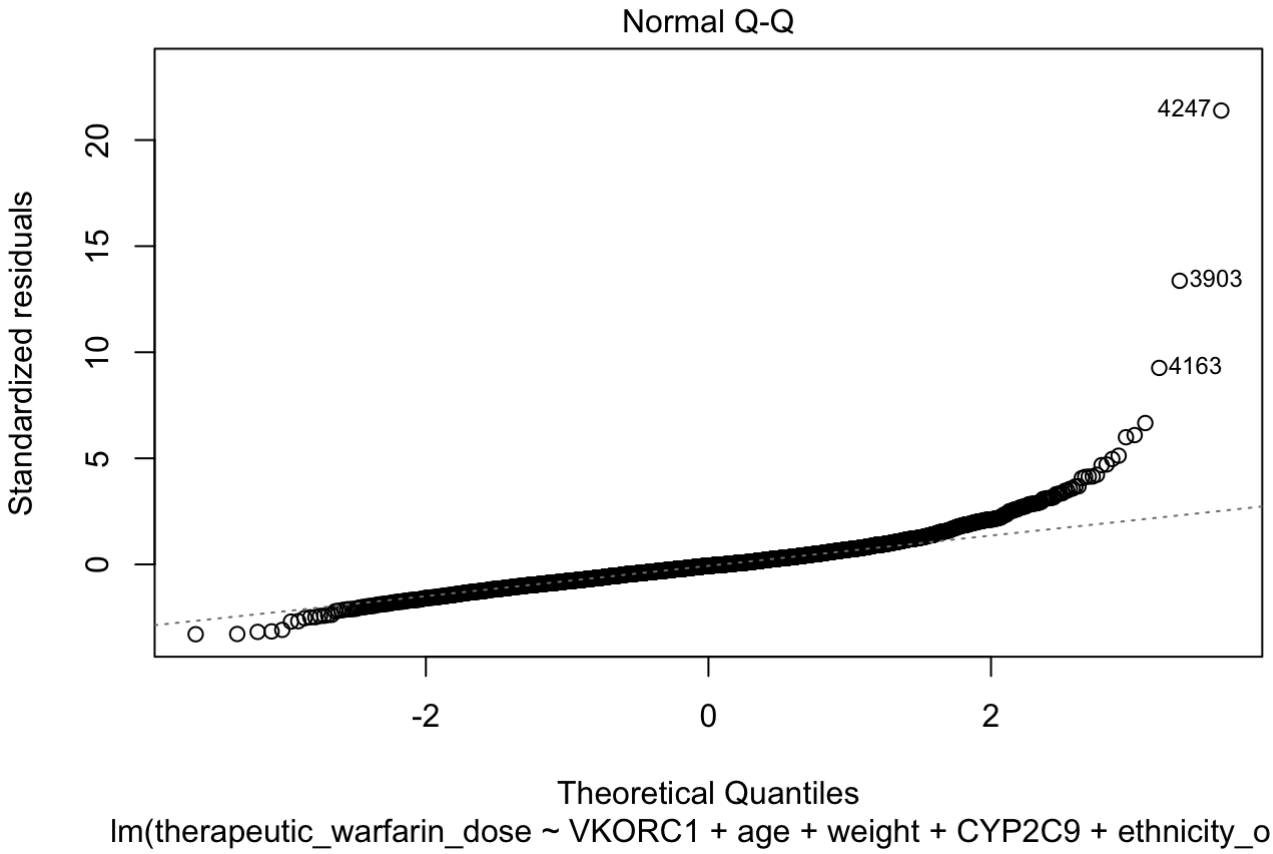
```
fit_final <- lm(therapeutic_warfarin_dose ~ VKORC1 + age + weight + CYP2C9 +
              ethnicity_omb + amiodarone_bool + race_omb + enzyme_inducer_bool, data = i
wpc)
summary(fit_final)
```

```
##
## Call:
## lm(formula = therapeutic_warfarin_dose ~ VKORC1 + age + weight +
##     CYP2C9 + ethnicity_omb + amiodarone_bool + race_omb + enzyme_inducer_bool,
##     data = iwpc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.788  -7.018  -0.836   5.304 274.414
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        13.48803    4.45672   3.026 0.002493
## VKORC1A/G                           9.09832    0.65309  13.931  < 2e-16
## VKORC1G/G                          19.59345    0.74801  26.194  < 2e-16
## age20 - 29                          0.20276    3.96807   0.051 0.959251
## age30 - 39                         -0.13226    3.86601  -0.034 0.972711
## age40 - 49                         -2.31223    3.78794  -0.610 0.541625
## age50 - 59                         -6.10325    3.75110  -1.627 0.103816
## age60 - 69                         -9.76439    3.74912  -2.604 0.009241
## age70 - 79                        -12.95544    3.74914  -3.456 0.000556
## age80 - 89                        -14.38170    3.79004  -3.795 0.000150
## age90+                            -15.58970    4.93130  -3.161 0.001584
## weight                              0.18318    0.01223  14.974  < 2e-16
## CYP2C9*1/*11                      -18.39345    9.10652  -2.020 0.043479
## CYP2C9*1/*13                       -9.05548   12.87188  -0.704 0.481786
## CYP2C9*1/*14                      -11.72137   12.87226  -0.911 0.362573
## CYP2C9*1/*2                        -6.13968    0.69411  -8.845  < 2e-16
## CYP2C9*1/*3                        -9.31487    0.80503 -11.571  < 2e-16
## CYP2C9*1/*5                       -14.75040    5.80365  -2.542 0.011078
## CYP2C9*1/*6                        -7.93539    9.25208  -0.858 0.391124
## CYP2C9*2/*2                       -12.39440    2.17553  -5.697 1.32e-08
## CYP2C9*2/*3                       -20.08957    2.06912  -9.709  < 2e-16
## CYP2C9*3/*3                       -21.33704    4.09184  -5.215 1.95e-07
## ethnicity_ombnot Hispanic or Latino   4.59821    2.23920   2.054 0.040098
## ethnicity_ombUnknown                  0.52414    2.26248   0.232 0.816811
## amiodarone_bool                      -8.01756    1.20839  -6.635 3.75e-11
## race_ombBlack or African American    -0.72794    1.10152  -0.661 0.508749
## race_ombUnknown                       1.75012    1.32554   1.320 0.186820
## race_ombWhite                         2.70777    0.75890   3.568 0.000364
## enzyme_inducer_bool                  13.72153    3.10727   4.416 1.04e-05
##
## (Intercept)                       **
## VKORC1A/G                         ***
## VKORC1G/G                         ***
## age20 - 29
## age30 - 39
## age40 - 49
## age50 - 59
## age60 - 69                        **
## age70 - 79                        ***
## age80 - 89                        ***
## age90+                            **
```

```
## weight                                *** 
## CYP2C9*1/*11                          *  
## CYP2C9*1/*13                             
## CYP2C9*1/*14                             
## CYP2C9*1/*2                           *** 
## CYP2C9*1/*3                           *** 
## CYP2C9*1/*5                           *  
## CYP2C9*1/*6                              
## CYP2C9*2/*2                           *** 
## CYP2C9*2/*3                           *** 
## CYP2C9*3/*3                           *** 
## ethnicity_ombnot Hispanic or Latino *  
## ethnicity_ombUnknown                     
## amiodarone_bool                       *** 
## race_ombBlack or African American        
## race_ombUnknown                          
## race_ombWhite                         *** 
## enzyme_inducer_bool                   *** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.86 on 3493 degrees of freedom
## Multiple R-squared:  0.4555, Adjusted R-squared:  0.4512 
## F-statistic: 104.4 on 28 and 3493 DF,  p-value: < 2.2e-16
```

The final model makes sense because gender is not a significant predictor for warfarin dosage and height is correlated to weight, so we only need to include either height or weight. Overall, the model is significant with p-value: < 2.2e-16 and decent Adjusted R-squared of 0.4512.

```
plot(fit_final, 2)
```



The Normal Q-Q plot for fit_final model is better compared to that of fit_full model, the points curve off lesser here.

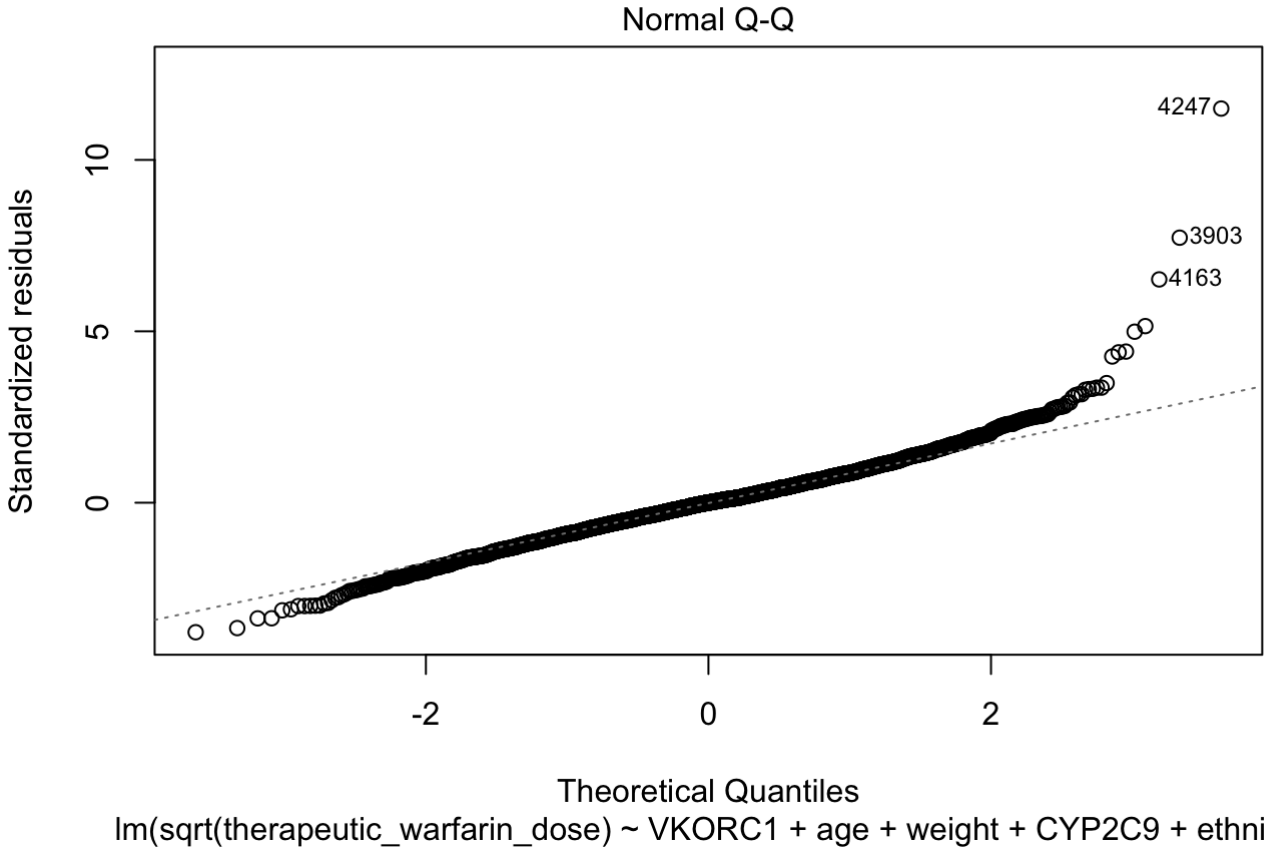
```
fit_final2 <- lm(sqrt(therapeutic_warfarin_dose) ~ VKORC1 + age + weight + CYP2C9 +
                   ethnicity_omb + amiodarone_bool + race_omb + enzyme_inducer_bool, dat
a = iwpc)
summary(fit_final2)
```

```
##
## Call:
## lm(formula = sqrt(therapeutic_warfarin_dose) ~ VKORC1 + age +
##     weight + CYP2C9 + ethnicity_omb + amiodarone_bool + race_omb +
##     enzyme_inducer_bool, data = iwpc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7800 -0.5984  0.0019  0.5727 11.5117
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       3.9644327  0.3478250  11.398  < 2e-16
## VKORC1A/G                         0.8715335  0.0509707  17.099  < 2e-16
## VKORC1G/G                         1.7285670  0.0583786  29.610  < 2e-16
## age20 - 29                       -0.0896919  0.3096882  -0.290  0.77212
## age30 - 39                       -0.1109488  0.3017225  -0.368  0.71311
## age40 - 49                       -0.3063858  0.2956301  -1.036  0.30010
## age50 - 59                       -0.5886390  0.2927550  -2.011  0.04444
## age60 - 69                       -0.9027325  0.2926000  -3.085  0.00205
## age70 - 79                       -1.2007250  0.2926013  -4.104 4.16e-05
## age80 - 89                       -1.3436105  0.2957939  -4.542 5.75e-06
## age90+                           -1.5449577  0.3848636  -4.014 6.09e-05
## weight                            0.0154572  0.0009547  16.190  < 2e-16
## CYP2C9*1/*11                     -1.5653362  0.7107187  -2.202  0.02770
## CYP2C9*1/*13                     -0.8261558  1.0045861  -0.822  0.41091
## CYP2C9*1/*14                     -1.1304113  1.0046153  -1.125  0.26057
## CYP2C9*1/*2                      -0.4881624  0.0541718  -9.011  < 2e-16
## CYP2C9*1/*3                      -0.8362249  0.0628287 -13.310  < 2e-16
## CYP2C9*1/*5                      -1.0808085  0.4529456  -2.386  0.01708
## CYP2C9*1/*6                      -0.3156868  0.7220786  -0.437  0.66200
## CYP2C9*2/*2                      -1.1448672  0.1697896  -6.743 1.81e-11
## CYP2C9*2/*3                      -1.9507296  0.1614845 -12.080  < 2e-16
## CYP2C9*3/*3                      -2.1861118  0.3193477  -6.846 8.96e-12
## ethnicity_ombnot Hispanic or Latino  0.3369866  0.1747582   1.928  0.05390
## ethnicity_ombUnknown             -0.0003370  0.1765749  -0.002  0.99848
## amiodarone_bool                  -0.7020865  0.0943084  -7.445 1.22e-13
## race_ombBlack or African American  -0.0395190  0.0859679  -0.460  0.64576
## race_ombUnknown                   0.2250253  0.1034521   2.175  0.02968
## race_ombWhite                     0.2725625  0.0592280   4.602 4.34e-06
## enzyme_inducer_bool               0.6684301  0.2425066   2.756  0.00588
##
## (Intercept)                      ***
## VKORC1A/G                        ***
## VKORC1G/G                        ***
## age20 - 29
## age30 - 39
## age40 - 49
## age50 - 59                       *
## age60 - 69                       **
## age70 - 79                       ***
## age80 - 89                       ***
## age90+                           ***
```

```
## weight                                *** 
## CYP2C9*1/*11                          *  
## CYP2C9*1/*13                             
## CYP2C9*1/*14                             
## CYP2C9*1/*2                           *** 
## CYP2C9*1/*3                           *** 
## CYP2C9*1/*5                           *  
## CYP2C9*1/*6                              
## CYP2C9*2/*2                           *** 
## CYP2C9*2/*3                           *** 
## CYP2C9*3/*3                           *** 
## ethnicity_ombnot Hispanic or Latino .  
## ethnicity_ombUnknown                     
## amiodarone_bool                       *** 
## race_ombBlack or African American        
## race_ombUnknown                       *  
## race_ombWhite                         *** 
## enzyme_inducer_bool                   ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.004 on 3493 degrees of freedom
## Multiple R-squared:  0.5161, Adjusted R-squared:  0.5122 
## F-statistic: 133.1 on 28 and 3493 DF,  p-value: < 2.2e-16
```

By using square root for the response variable, Adjusted R-squared value is improved to 0.5122.

```
plot(fit_final2, 2)
```



The

Normal Q-Q plot for fit_final2 model is the best out of all the models, the points are almost a striaght line.

Support Vector Machine (SVM)

First, iwpc_svm dataframe is created with only key variables required for classification. Then, the rows with NA's are removed and factor variables are converted from numeric to factor.

```
iwpc_svm <- select(iwpc_hl, gender, race_omb, ethnicity_omb, age, weight,
                   CYP2C9, VKORC1, amiodarone_bool, enzyme_inducer_bool, dosage_level)
iwpc_svm <- na.omit(iwpc_svm)
iwpc_svm$amiodarone_bool <- factor(iwpc_svm$amiodarone_bool)
iwpc_svm$enzyme_inducer_bool <- factor(iwpc_svm$enzyme_inducer_bool)
iwpc_svm$dosage_level <- factor(iwpc_svm$dosage_level)
```

The data is then split into training set (80%) and test set (20%).

```
# Split training and testing sets
dat <- iwpc_svm
set.seed(101)

train <- sample(nrow(dat), round(0.8*nrow(dat)))
trainset <- dat[train, ]
testset <- dat[-train, ]
```

The svm() function is used to train a model on the training set, and get a summary of the model.

```
library(e1071)
model <- svm(dosage_level ~ ., data=trainset, kernel = "radial", cost = 1, gamma = 1)
summary(model)
```

```
##
## Call:
## svm(formula = dosage_level ~ ., data = trainset, kernel = "radial",
##     cost = 1, gamma = 1)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  1
##       gamma:  1
##
## Number of Support Vectors:  1410
##
##  ( 720 690 )
##
##
## Number of Classes:  2
##
## Levels:
##  high low
```

Predict function is used to predict new values from the test set using the model.

```
predicted.values <- predict(model, testset[1:9])
# confusion matrix
table(true=testset$dosage_level, pred=predicted.values)
```

```
##       pred
## true   high low
##   high  206  63
##   low    57 310
```

The confusion matrix shows that there are 210 (33.0%) True Positives and 317 (49.8%) True Negatives. The accuracy comes out to be 82.9%, which is decent. Next, the parameters are tuned in an attempt to improve the model.

```
set.seed(101)
tune.results <- tune(svm, train.x=dosage_level~., data=trainset,kernel='radial',
                     ranges=list(cost=10^(-1:2), gamma=c(.125, .5,1,2)))
tune.results
```

```
##
## Parameter tuning of 'svm':
##
## - sampling method: 10-fold cross validation
##
## - best parameters:
##  cost gamma
##     1 0.125
##
## - best performance: 0.1749668
```

The best performance occurs with cost=10 and gamma=0.125. The model is trained again with these specific parameters.

```
model_tuned <- svm(dosage_level ~ .,data=trainset, kernel = "radial", cost=10, gamma =
0.125)

predicted.values.tuned <- predict(model_tuned,testset[1:9])
table(true=testset$dosage_level, pred=predicted.values.tuned)
```

```
##       pred
## true   high low
##   high  207  62
##   low    47 320
```

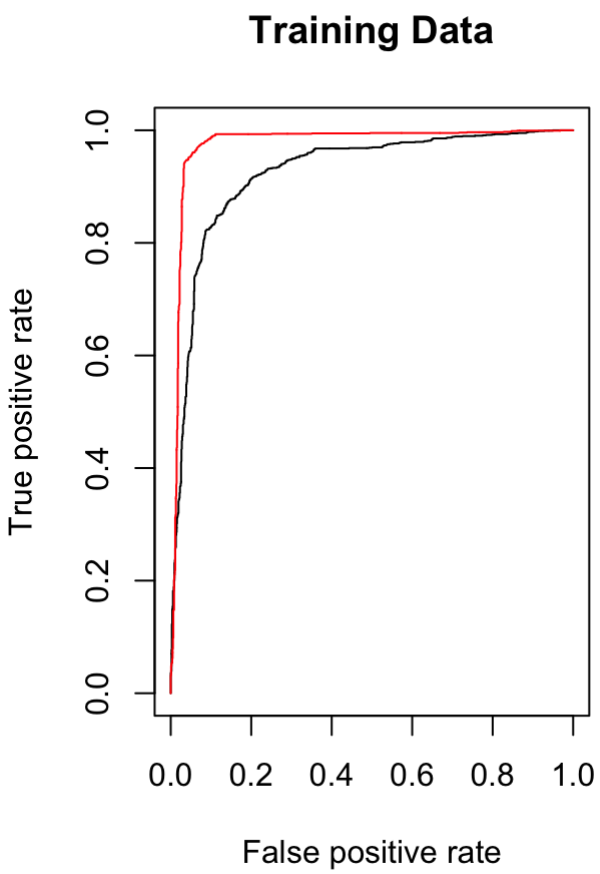
With tuned parameters, True Positives improved to 222 (34.9%), but True Negatives went down to 308. So the accuray for the tuned model is 83.3%, it only improved very little.

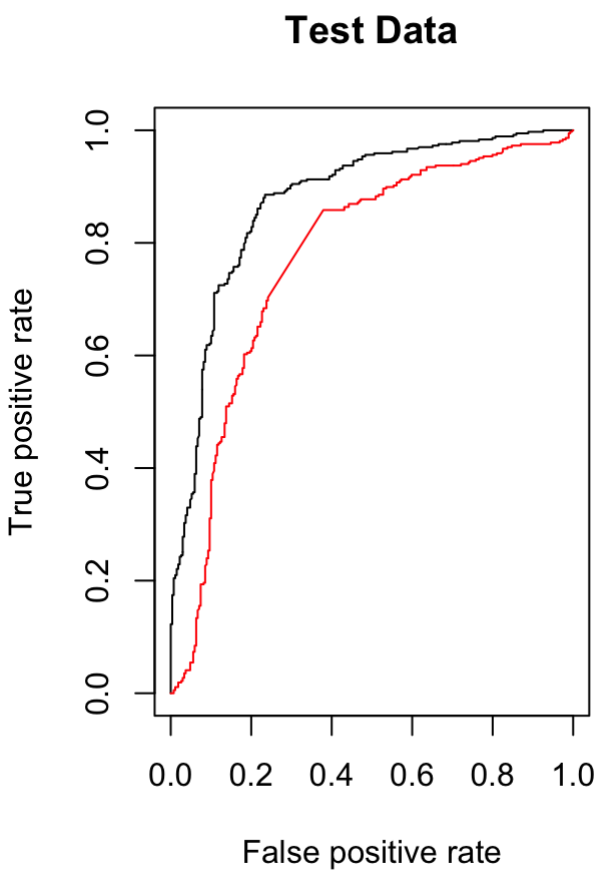
Receiver Operating Characteristic (ROC)

In a ROC curve, the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold.

```
library(ROCR)
# Creating function to plot ROC curve
rocplot = function(pred, truth, ...) {
  predob = prediction(pred, truth)
  perf = performance(predob, "tpr", "fpr")
  plot(perf, ...)
}

# Optimal model based on tuning
svmfit.opt = svm(dosage_level ~ .,
                 data = trainset,
                 kernel = "radial",
                 gamma = 0.125,
                 cost = 10,
                 decision.values = T) # to obtain the fitted values for a given SVM mode
l

# γ is increased to produce a more flexible fit and generate further improvements in acc
uracy
svmfit.flex = svm(dosage_level ~ .,
                  data = trainset,
                  kernel = "radial",
                  gamma = 50,
                  cost = 10,
                  decision.values = T)

fitted1 = attributes(predict(svmfit.opt, trainset,
                             decision.values = TRUE))$decision.values
fitted2 = attributes(predict(svmfit.flex, trainset,
                             decision.values =T))$decision.values
par(mfrow = c(1,2))

# ROC plot for optimal model
rocplot(fitted1,
        trainset[,'dosage_level'],
        main = "Training Data")
# ROC model of flexible model
rocplot(fitted2,
        trainset[,'dosage_level'],
        add = T,
        col = "red")

fitted3 = attributes(predict(svmfit.opt, testset,
                             decision.values = T))$decision.values
fitted4 = attributes(predict(svmfit.flex, testset,
                             decision.values =T))$decision.values

rocplot(fitted3,
        testset[,'dosage_level'],
        main = "Test Data")

rocplot(fitted4,
        testset[,'dosage_level'],
```

```
        add = T,
        col = "red")
```





In the ROC plots, the black lines are from the optimal model with tuned parameters, and the red lines are from the flexible models with bigger gamma value for a more flexible fit and accuracy. Even though the red line shows a better accuracy in the training data, it has worse accuracy in the test data when compared to the black line.