Data 630 9040

Machine Learning 2215

Professor Bati Firdu

Melissa Hunfalvay

Date: 6-22-2021

Assignment 2

Introduction

Objective

The dataset used for this project was the Acquired Immune Deficiency Syndrome (AIDS) clinical trials study group 320 data from the University of Massachusetts at Amherst.

The objective of the analysis was to explore the factors that influence CD4 cell levels in the hope that doctors may to prescribe more effective treatments at different stages of the AIDS disease and perhaps better understand the disease trajectory and health care options (Farhadian, Mohammadi, Mirzaei, & Shirmohammadi-Khorram, 2021). Furthermore, understanding factors that influence CD4 cells may help to improve pre-screening criteria for drug studies which in turn are designed to improve a patient's quality of life (Hirsch et al., 1999). Drug studies are very expensive and often many drugs are left on the shelf because pre-screening of patients was not sensitive or specific enough which ultimately affects the impact of the drugs on the disease (Fogel, 2018).

To answer the objective the analysis a binary logistic regression, was used to estimate the probability of an outcome of CD4 cells above and below a threshold of 50 cells per cubic millimeter (cells/mm3). Given the results of the logistic regression, a second analytical model, a Naive Bayesian Classification was used to estimate the class probability and specifically to determine if this model could improve, accuracy, sensitivity, and specificity (Appendix C).

Both logistic regression and Naïve Bayesian classification are a type of supervised learning classification techniques. Supervised learning refers to giving the dataset a predetermined, already known "right answer" or outcome (Ng, 2021). This dataset contained AIDS patients, in which, for every patient there was a known outcome, of CD4 cells (<=50 or >50 cells/mm3).

Problem Domain

Every year, on December 1st World AIDS Day highlights awareness and support for those living with AIDS. World AIDS day started in 1998. The current dataset was collected almost a decade later, in 1997. Yet, even today, there is still a lot of work to be done to understand and treat AIDS according to the World Health Organization (https://www.who.int/news-room/fact-sheets/detail/hiv-aids).

As of late 2020, HIV/AIDS has claimed 34.7 million lives globally and continue to be a worldwide public health issue (https://www.who.int/news-room/fact-sheets/detail/hiv-aids). AIDS disproportionately affects men who have sex with men, racial and ethnic minorities, and people in the Southern states. As of 2021, there are 1.2 million Americans living with HIV. Many are unaware they are infected (14%) and many children are infected (https://www.webmd.com/hiv-aids/guide/HIV-AIDS-living-managing).

There is still no cure for AIDS. Yet, through research, education, medical advancements in drug therapies and social education, AIDS has reduced in prevalence in the United States by 8% over the last five years (https://www.hiv.gov/hiv-basics/overview/data-and-trends/statistics).

Life expectancy of those with a positive diagnosis has also increased since the turn of the century from an average of 59 years in 2000-2003; to 77 in 2014-2016. By comparison however, people without AIDS still live longer, averaging 86 years (Helford, 2020).

When a person is diagnosed with HIV there are several options for treatment. One option is a drug “cocktail”. Understanding the right drug cocktail for the patient and stage of disease progression is challenging. However, prescribing the right cocktail, in an expeditious fashion,

can improve a patients' quality of life and potentially extend the patients' life (Gulick et al., 1997).

The level of CD4 cells in the blood is perhaps the most important criteria for understanding a patients' stage within the AIDS disease (Cichocki, 2020). CD4 cells are "helper cells" because they trigger the body's immune response to infection. Predicting and classifying patients with low (<=50 cells/mm3) or higher (>50 cells/mm3) is an important criterion in knowing what course of therapy may be appropriate.

Therefore, the purpose of this analysis is to explore predictors of CD4 cell levels. It is hopeful, that by understanding this information, we can be more informed about the specific factors (virology, epidemiology, clinical state, and drug therapies) that influence these levels. By doing so, we may be able to prescribe effective treatments at different stages of the AIDS disease and perhaps better understand the disease trajectory. We may also be better able to pre-screen patients into drug studies.

Method Rationale

The main methodology chosen is a logistic regression, which is a statistical classification techniques and form of supervised learning.

The rationale for *supervised learning* methodology includes:

a) A known outcome of CD4 cells (<=50 or >50 cells/mm3).

b) Variables within the dataset that may be used to predict the known outcomes

More specifically, the rationale for use *logistic regression* includes:

a) There is a predictive value, that is CD4 cells (<=50 or >50 cells/mm3), (0,1) that can be directly predicted in logistic regression.

b) There is a dichotomous dependent variable CD4 cells (<=50 or >50 cells/mm3) and the independent variables are continuous or categorical which is well suited to a binary logistic regression model.

As this dataset is medical and the condition is known (AIDS) the most useful information involves a predictive analysis. Patients, doctors, and educators want to ask questions of a dataset like this to include:

1. What factors affect CD4 levels that may assist healthcare professionals and patients with AIDS in the management and monitoring of health cares?
2. What combination of medication affect CD4 levels and are best for patients with certain types of virology, epidemiology, clinical state?
3. What factors effect disease trajectory and which medications may be used to intervene at the different stages?

Analysis

Data

This data set was collected in 1997 as part of The AIDS Clinical Trails Group 320 Study. Patients were recruited from 33 AIDS Clinical trial Units and 7 National Hemophilia Foundation sites located in the United States. The study was randomized, double-blind, placebo-controlled trials that comparted varying drug regiments. The three-drug regiment included indinavir (Crixivan), open-label zidovudine (Retrovir) or stavudine (Zerit), and lamivudine (Epivir). The two-drug regiment included zidovudine (or stavudine) and lamivudine.

Pre-screening criteria: Patients in this dataset had progressed from the Human Immunodeficiency Virus (HIV) to AIDS. A progression to a diagnosis of AIDS is when the

number of CD4 cells fall below 200 cells per cubic millimeter of blood (200 cells/mm3). This was one of the pre-screening criteria for patients in this study, whereby they needed to have the cell count within 60 days of entry into the study. As a comparison, people with a healthy immune system have CD4 counts between 500 and 1600 cells/mm3 (https://www.hiv.gov/hiv-basics/overview/about-hiv-and-aids/).

Patients had to be 16 years of age and have laboratory documentation of HIV. Patients also needed a Karnofsky performance score of at least 70 which states a condition of "Unable to work; able to live at home and core for most personal needs; varying amounts of assistance needed" (Schag, Heinrich, Ganz, 1984). Patients needed to have no more than 1week of prior lamivudine treatment prior to entering the study.

Interestingly, after a second review by a data and safety monitoring board on February 18th, 1997, a comparison of the groups reveled significant differences that met the prespecified guideline for stopping the study (Peto, Pike, Armitage, et al., 1976). Therefore, the board recommended that no additional patients be added, and the study was closed.

Variables in the data set included epidemiological factors including sex (1 = male, 2 = female); race/ethnicity (1 = White non-Hispanic, 2 = Black non-Hispanic, 3 = Hispanic regardless of race, 4 = Asian or Pacific Islander, 5 = American Indian, Alaskan Native, 6 = Other or unknown). Age of patients was also included in years.

Information regarding the patients' level of impairment at the time of entry into the study included a score of the Karnofsky scale between 70-100. This scale allows patients to be classified as to their functional impairment and can be used to compare the effectiveness of treatment (http://www.npcrc.org/files/news/karnofsky_performance_scale.pdf). The lower the scale the worse the functional state. Scores are reported in increments of 10 and patients minimal

acceptable score for this study was 70 on the Karnofsky scale. A score of 100 indicates the patient is "Able to carry on normal activity and to work; no special care needed" (http://www.npcrc.org/files/news/karnofsky_performance_scale.pdf).

Patients' medical information included if they were a hemophiliac (1 = yes, 0 = No), if they had IV drug use (1 = never, 2 = currently, 3 = previously) and a record of open label zidovudine (priorZDV), measured in months. The level of CD4 stratum testing results at the time of screening was included in the dataset (0 = CD4<= 50; 1 = CD4 >50). A baseline CD4 count was also obtained and recorded as cells per milliliter.

Treatment indicator (1 = Treatment includes IDV, 0 = Control group treatment without IDV); and treatment group indicator (2 = ZDV + 3TC + IDV, 4 = d4T + 3TC + IDV) versus control group indicator (1 = ZDV+3TC, 3 = d4T +3TC) were included in the dataset.

Finally, time to AIDS diagnosis or death, measured in days; and the event indicator for AIDS defining diagnosis of death (1= AIDS defining diagnosis or death; 0 = Otherwise) were recorded in the dataset.

Exploratory Analysis

There are 16 variables in the data set, as shown by the str function (Figure 1). The attributes in this dataset include both continuous data (such as age) and categorical data (such as sex or race). Variable types include numeric variables, including "int" or integers which are discrete (as opposed to continuous) categorical variables, examples include censor_d (death or otherwise) and Treatment (control group or treatment group).

```
> # a) Provide basic description of the data
> str(actg320)
'data.frame':   1151 obs. of  16 variables:
 $ id           : int  1 2 3 4 5 6 7 8 9 10 ...
 $ time         : int  189 287 242 199 286 285 270 285 276 306 ...
 $ censor       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ time_d       : int  189 287 242 199 286 285 270 285 276 306 ...
 $ censor_d     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Treatment    : int  0 0 1 0 1 1 0 1 0 0 ...
 $ Treatment_grp: int  1 1 2 1 2 2 1 2 1 1 ...
 $ CD4_Stratum  : int  1 1 0 1 0 0 1 1 1 1 ...
 $ sex          : int  1 2 1 1 1 1 1 1 1 1 ...
 $ Race         : int  1 2 1 1 1 1 2 2 1 1 ...
 $ IV_drug      : int  1 1 1 1 3 1 1 3 1 1 ...
 $ hemophil     : int  0 0 1 0 0 0 0 0 0 0 ...
 $ karnof       : int  100 90 100 90 90 70 100 80 100 90 ...
 $ CD4_base     : num  169 149.5 23.5 46 10 ...
 $ priorzdv     : num  39 15 9 53 12 24 6 24 7 7 ...
 $ age          : int  34 34 20 48 46 51 51 40 34 38 ...
>
```

Figure 1: str function output without any pre-processing

A second variable type in this dataset are "num" variables, which are real numbers, that have a value of a continuous quantity, that can represent a distance along a line (or alternatively, a quantity that can be represented as an infinite decimal expansion; Feferman, 1989). An example of a num variable in the data set is CD4_base and time.

There are 1151 rows (or observations) of data. Each row represents one unique patient. Missing data values were checked. None were found in this data set.

An initial review of the data revealed one variable (id) to be removed. This variable was not additive to the dataset. Furthermore, some of the numeric "int" variables needed to be transformed to factors for the analysis as required by the Naïve Bayesian Classification method. These variables included: censor, censor_d, treatment,, treatment_grp, CD4_Stratum, sec, Race, IV-Drug, hemophil and karnof. Labels were added to the levels for easy of understanding.

The str function was then re-run showing the changes in the dataset (see Figure 2). At this point in the exploratory analysis there were 15 variables, five were numeric (2 = num, 3 = int)

and ten factors. Factors are discrete, categorical variables with pre-determined levels such as yes or no for status as a hemophiliac.

```
> str(actg320)
'data.frame':   1144 obs. of  15 variables:
 $ time         : int  189 287 242 199 286 285 270 285 276 306 ...
 $ censor       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ time_d       : int  189 287 242 199 286 285 270 285 276 306 ...
 $ censor_d     : Factor w/ 2 levels "Otherwise","Death": 1 1 1 1 1 1 1 1 1 1 ...
 $ Treatment    : Factor w/ 2 levels "Control","Treatment wIDV": 1 1 2 1 2 2 1 2 1 1 ...
 $ Treatment_grp: Factor w/ 2 levels "1","2": 1 1 2 1 2 2 1 2 1 1 ...
 $ CD4_Stratum  : Factor w/ 2 levels "<=50",">50": 2 2 1 2 1 1 2 2 2 2 ...
 $ sex          : Factor w/ 2 levels "Male","Female": 1 2 1 1 1 1 1 1 1 1 ...
 $ Race         : Factor w/ 6 levels "White","Black",..: 1 2 1 1 1 1 2 2 1 1 ...
 $ IV_drug      : Factor w/ 2 levels "1","3": 1 1 1 1 2 1 1 2 1 1 ...
 $ hemophil     : Factor w/ 2 levels "No","Yes": 1 1 2 1 1 1 1 1 1 1 ...
 $ karnof       : Factor w/ 4 levels "70","80","90",..: 4 3 4 3 3 1 4 2 4 3 ...
 $ CD4_base     : num  169 149.5 23.5 46 10 ...
 $ priorzdv     : num  39 15 9 53 12 24 6 24 7 7 ...
 $ age          : int  34 34 20 48 46 51 51 40 34 38 ...
```

Figure 2: str function output after renaming and transforming variables

To further explore the data, the summary command was run for all variables (Figure 3).

```
> summary (actg320)
      time         censor       time_d          censor_d            Treatment   Treatment_grp CD4_Stratum     sex                Race     IV_drug hemophil   karnof
 Min.   :  1.0   0:1049   Min.   :  1.0   Otherwise:1119   Control      :575   1:575         <=50:437    Male  :946   White         :594   1:966   No :1109   70 : 31
 1st Qu.:174.8   1:  95   1st Qu.:196.0   Death    :  25   Treatment wIDV:569   2:569         >50 :707    Female:198   Black         :324   3:178   Yes:  35   80 :181
 Median :257.0            Median :265.0                                                                              Hispanic      :201                      90 :539
 Mean   :230.6            Mean   :242.8                                                                              Asian/PI      : 14                      100:393
 3rd Qu.:300.0            3rd Qu.:306.0                                                                              Indian/Alaskan: 11
 Max.   :364.0            Max.   :364.0                                                                              Other/Unknown :  0
    CD4_base        priorzdv          age
 Min.   :  0.00   Min.   :  3.00   Min.   :15.00
 1st Qu.: 22.88   1st Qu.: 10.00   1st Qu.:33.00
 Median : 75.00   Median : 21.00   Median :38.00
 Mean   : 86.57   Mean   : 30.41   Mean   :38.61
 3rd Qu.:136.50   3rd Qu.: 42.00   3rd Qu.:44.00
 Max.   :392.00   Max.   :312.00   Max.   :73.00
```

Figure 3: Summary Command showing descriptive statistics for all variables in the data set.

For each factor variable additional exploration was conducted to include using the head command for showing the first 100 rows of data to explore frequency of outputs; a count of the levels within the factors to determine distribution across the levels within the variable; and a percentage of each level within the factor (Figure 4).

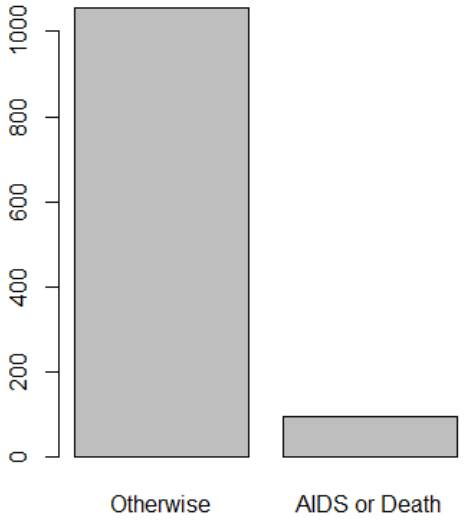
```
> actg320$censor<-as.factor(actg320$censor)  #Transform int to factor
> actg320$censor<-factor(actg320$censor, levels = 0:1, labels = c("Otherwise", "AIDS or Death")) # Relabel levels for easy understanding
> # show first 100 raw variables
> head(actg320$censor, 100)
  [1] Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise
 [13] Otherwise     AIDS or Death Otherwise     Otherwise     AIDS or Death Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     AIDS or Death Otherwise
 [25] Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     AIDS or Death Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise
 [37] Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise
 [49] Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     AIDS or Death Otherwise     Otherwise     Otherwise
 [61] Otherwise     Otherwise     AIDS or Death Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     AIDS or Deat
 [73] Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise
 [85] AIDS or Death Otherwise     Otherwise     AIDS or Death Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise     Otherwise
 [97] Otherwise     Otherwise     Otherwise     Otherwise
Levels: Otherwise AIDS or Death
> # Count the number of values within each category
> table(actg320$censor)

    Otherwise AIDS or Death
         1055           96
> # Percentage
> table(actg320$censor)/length(actg320$censor)

    Otherwise AIDS or Death
   0.91659427    0.08340573
```

Figure 4: Additional exploratory analysis for the Censor variable

Factor variables were also visualized using both a bar chart and a pie chart (Figure 5 and 6 respectively). These visualizations helped see the data and assist with further processing and decision making.



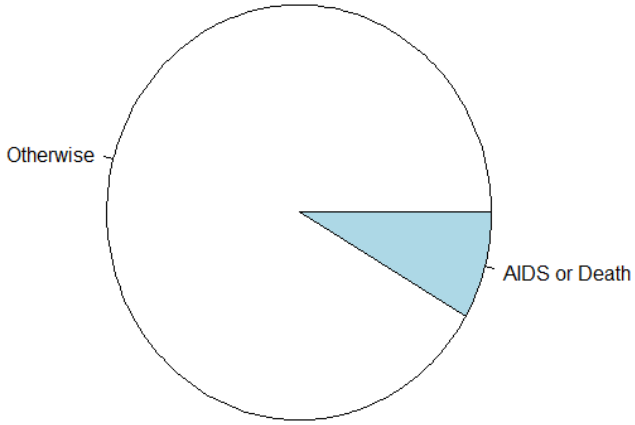


Figure 5: Bar chart; Figure 6: Pie chart for Censor variable

It can be observed via the visualizations and descriptive statistics that there were more males than females in the dataset (83%, 17% respectively). The predominant racial makeup included 51% white persons; then second most frequent race/ethnicity was black at 28%.

Information regarding the patients' level of impairment at the time of entry into the study as recorded from the Karnofsky scale revealed patients at each level of the scale between 70-100

at increments of 10. There were 31 patients with a score of 70; 181 with a score of 80; 539 with a score of 90 and 393 with a score of 100. The higher the score the better the functioning. 100 is normal functioning. Most patients in the study were therefore on the higher level of function in their daily lives.

There were 35 patients who were hemophiliacs (3%). Four patients who were currently using IV drugs (.3%) and 179 who were previously IV drug users (15.55%) and 968 people who had never used drugs (84.10%).

Blood levels (CD4 Stratum) of patients included 439 who were less than or equal to 50 cells per milliliter (38.14%). 712 patients with CD4 levels above 50 cells per millimeter (61.86%).

The control group consisted of 577 patients (50.13%) and treatment group was 574 patients (49.87%; Figure 6). Sadly, twenty-six patients died during the study and the average time to death was 242.30 days.

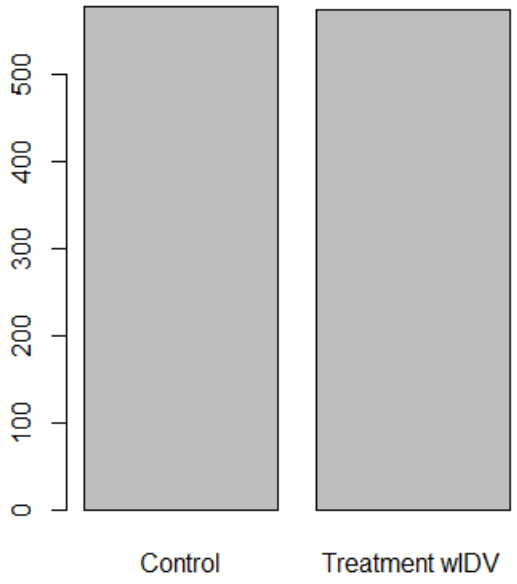


Figure 6: Bar plot showing treatment and control groups had even numbers

For all numeric variables (int and num) the summary command and standard deviation and mode were calculated (Figure 7). As it is important to determine the variance and skewness of the data for a regression as this influences the coefficients, it was deemed prudent to include these extra calculations during the exploratory analysis.

```
> # age
> # Descriptive Statistics
> summary (actg320$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.00   33.00   38.00   38.65   44.00   73.00
> # Standard deviation
> sd(actg320$age)
[1] 8.811395
> # Mode
> Data_age <- actg320
> names(sort(-table(Data_age$age)))[1]
[1] "33"
```

Figure 7: Summary command, standard deviation, and calculation of the mode for Age

Visualizations for numeric variables included histograms and boxplots (Figure 8 & 9). These were considered appropriate for the continuous nature of the numeric variables to show both distribution and outliers. These visualizations along with the descriptive calculations were used to make further decisions regarding data cleaning and processing.

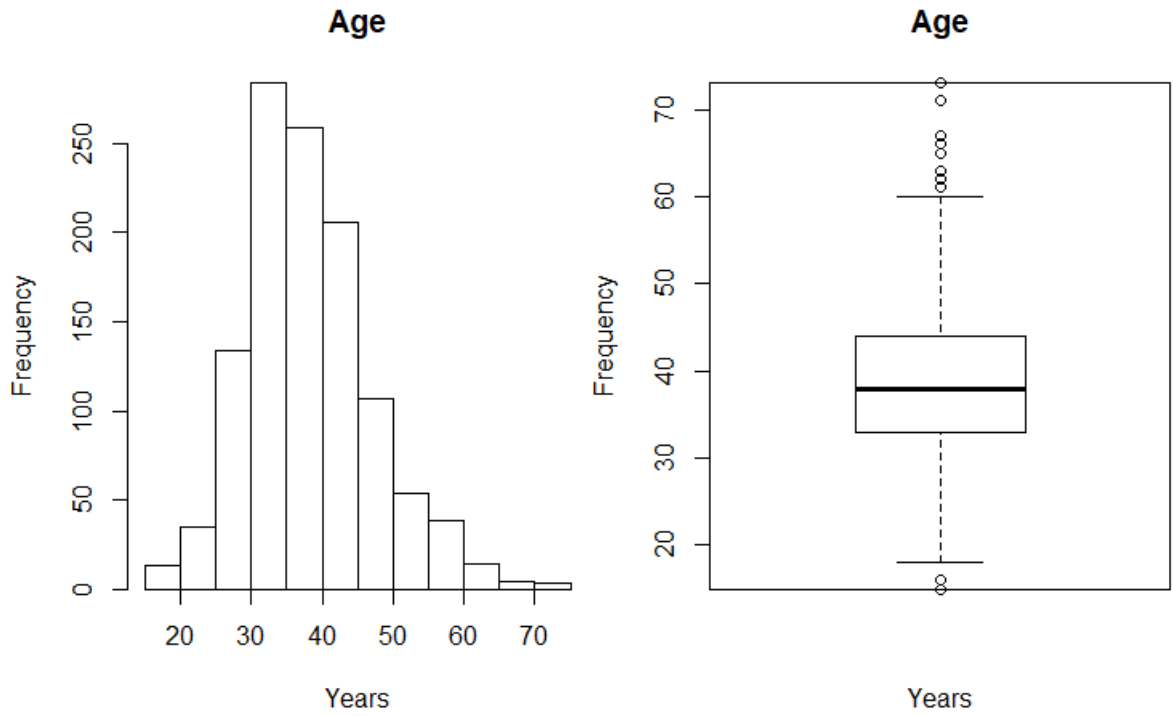


Figure 8: Histogram; Figure 9: Boxplot for the age variable.

For the numeric variables it can be observed that age was normally distributed throughout the dataset (N = 1151; Figure 8). However, there were outliers in the age group (Figure 9) to include both young and elderly persons. Prescreening preventing people under 16 from participating, nevertheless there were two outliers in the 25th percentile between 16 to 20 years of age.

Patients pre-screening required a minimum of 3 months of ZDV use. The mean was 30.42 months with a standard deviation of 6 months, meaning that most people fell within approximately one year (24.42 months) up to two years (36.42 months) of prior use. We can suspect based on the minimum, maximum and range (3, 312, 309 months respectively) that there are some outliers and skewness of this variable. The boxplot (Figure 9) and histogram (Figure 10) confirms this suspicion and shows the data is right skewed.

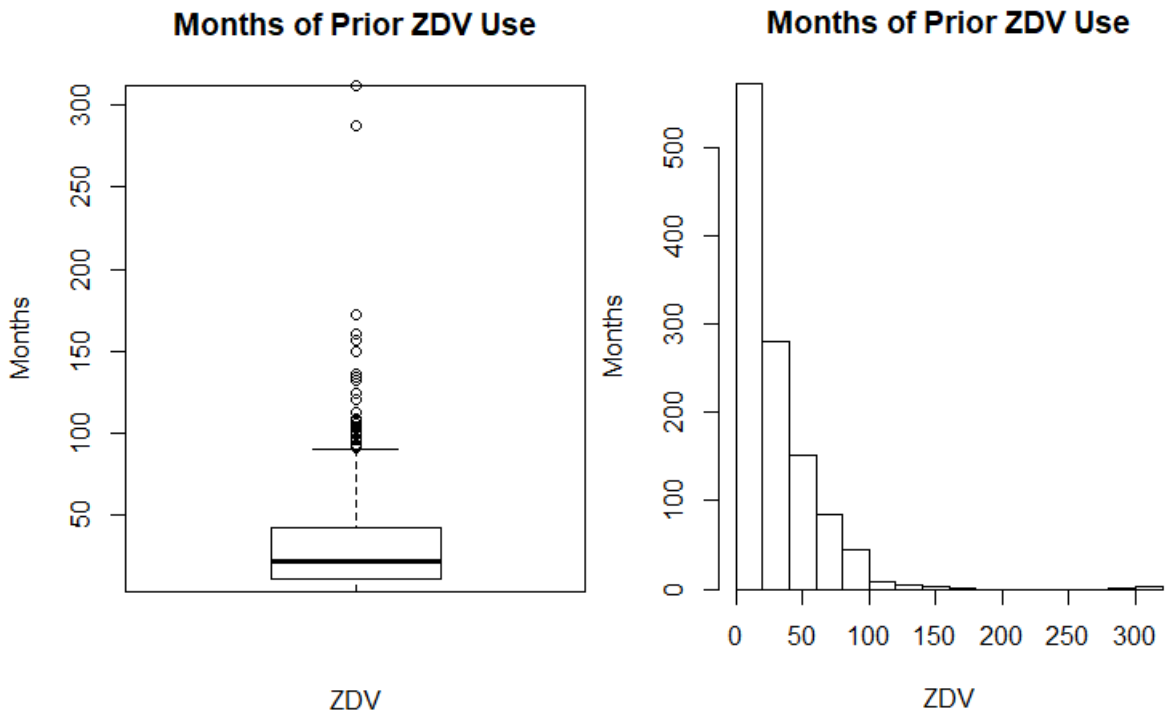


Figure 9: Boxplot; Figure 10: Histogram of prior ZDV use measured in months

Similar results are found for CD4 base. This is the baseline count of CD4 blood levels at the time of initiation into the study. Unlike the previous CD4 factor variable which delineated CD4 levels as either less than or equal to 50 cells per milliliter or above 50 cells per milliliter, the CD4 base variable is continuous and therefore provides a more precise count of CD4 in the blood (Figure 11 & 12).

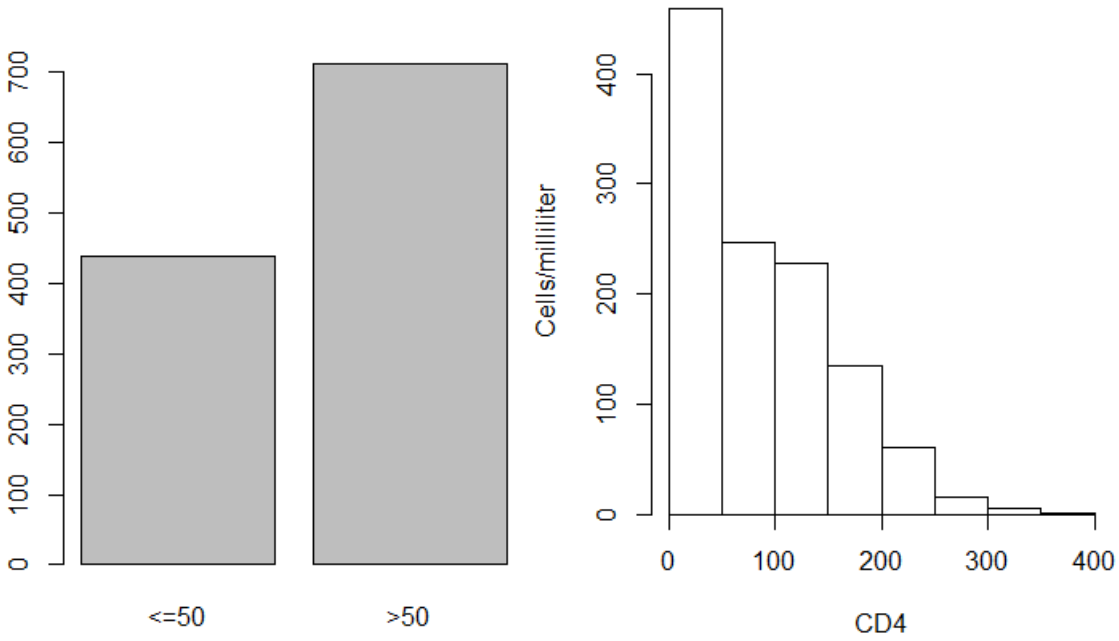


Figure 11: Bar plot of dichotomous factor variable CD4 stratum screening; Figure 12: Histogram of same variable on a continuous scale, showing more precise blood levels.

Time to death variable show definite skewness as seen by an average of 242.3 days with a mode of 293 days. When viewing the histogram, we can appreciate that it is left skewed, indicating that most patients lived longer than the minority of patients who died earlier in the study (Figure. There are some early (outlier) deaths seen in Figure 14.

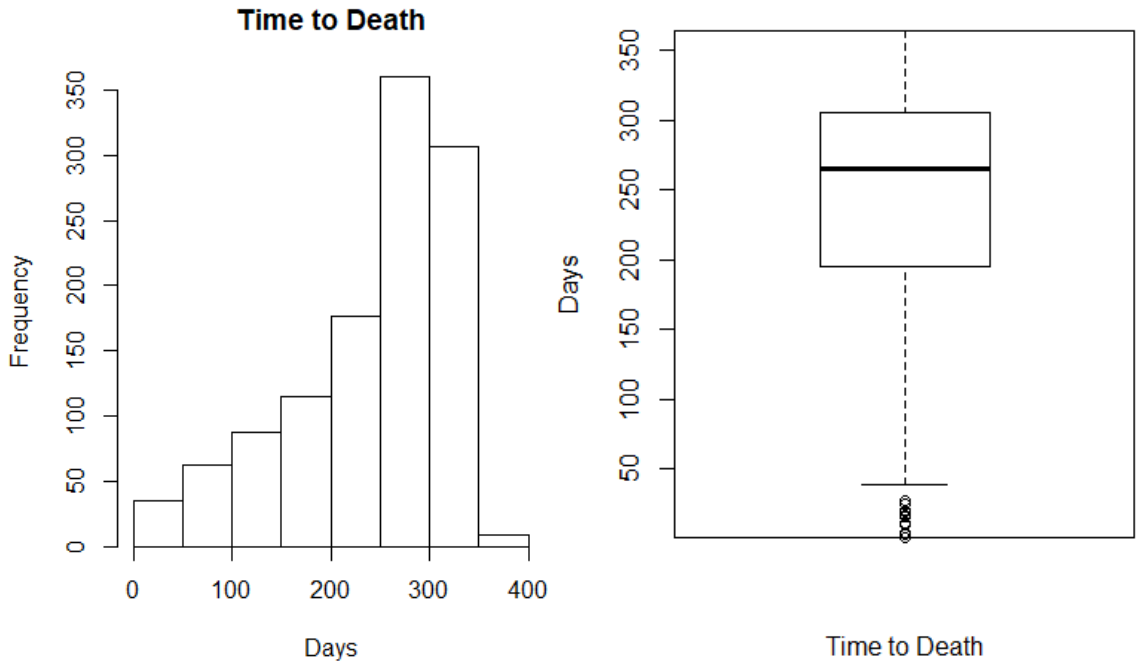


Figure 13: Histogram; Figure 14: Boxplot of time to death

Preprocessing

Armed with the objective and exploratory analysis, as well as the model type preprocessing was conducted. This involved removal of some rows of data, removal of outliers from numeric variables, removal of unwanted variables due to collinearity.

Removal of Rows. Rows were removed from two variables (Figure 14). The purpose of removing these rows was because the number of instances (or patients) who had these factors was very low. For instance, the were four people who were currently IV drug users in the study, this represented such a small number that it was deemed not additive to the analysis and could be removed. After removal of rows the data set was reduced from 1151 rows (observations) to 1144. A total of seven rows were removed.

```
> #Preprocessing
>
> #a) remove rows
>
> # Treatment group removed levels 3 & 4
> table(actg320$Treatment_grp) # before

  1   2   3   4
576 572   1   2
> actg320=actg320[-which(actg320$Treatment_grp==3 | actg320$Treatment_grp== 4),]
> actg320$Treatment_grp<-as.factor(actg320$Treatment_grp)
> # Count the number of values within each category
> table(actg320$Treatment_grp) # after

  1   2
576 572
>
> # IV_drug remove current drug use
> table(actg320$IV_drug) # before

  1   2   3
966   4 178
> actg320=actg320[-which(actg320$IV_drug==2),]
> actg320$IV_drug<-as.factor(actg320$IV_drug)
> # Count the number of values within each category
> table(actg320$IV_drug) # after

  1   3
966 178
```

Figure 14: Removal of rows from Treatment Group and IV_drug

Outlier Removal. Outliers were removed from all numeric variables. Outliers were defined as those falling outside the 25th and 75th percentiles (Figure 15 & 16).

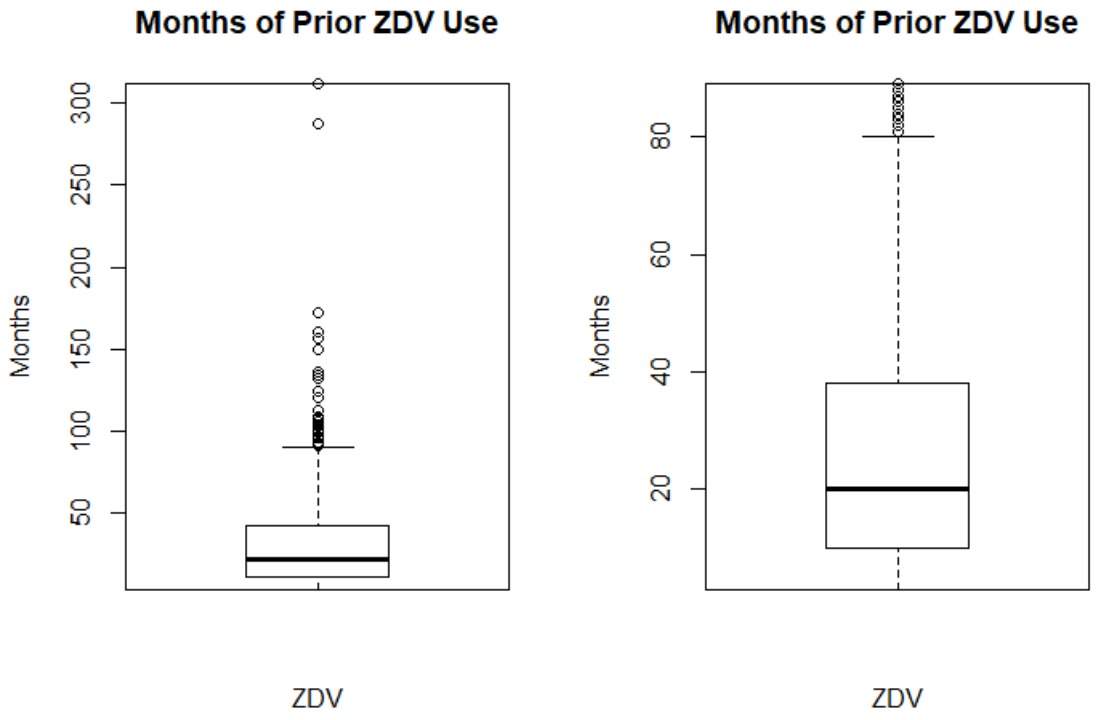


Figure 15: Box plot for prior ZDV use with outliers. Figure 16: Box plot for prior ZDV after outlier removal

Removal of variables. Throughout the exploratory analysis it was observed that several variables were colinear (Appendix A). The Treatment_grp variable had collinearity with the Treatment variable and was removed. The censor_d variable had collinearity with the censor variable and was removed. The CD4_base variable had collinearity with the CD4_stratum variables and was removed. When correlation between independent variables in the same regression model are high, they cannot independently predict the value of the dependent variable and therefore should be removed (Enders, 2020).

Algorithm Intuition

Logistic regression method is a classification and regression models. The binary logistic regression uses a 0/1 code (CD4 cells <=50 or >50 cells/mm3) to model the probability that *Y* belongs to one of these categories. This method estimates the probability of an outcome.

To understand logistic regression, it is first important to know how it differs from linear regression. An important distinction between logistic and linear regression is that logistic regression uses log function, whereas linear regression uses regression by least squares. The linear regression plots a straight "best-fit" straight line through the data. The straight line is derived from the linear regression model $\beta 1$ that gives the average change in *Y* associated with a one-unit change in *X*.

In contrast, in a logistic regression model, increases X by a one unit changes the *log odds* by $\beta 1$. Therefore, because the relationship between p(X) and X is not a straight line $\beta 1$ does not correspond to the one-unit change increase in X as is associated with linear regression. Hence a

logistic regression will always produce an S-shaped curve. This allows logistic regression to capture a larger range of possibility than linear regression.

The principle of the regression modeling is that we can always predict a value for p(X). However, sometimes the value for p(X) falls outside 0/1 and the range of X is unlimited. In this case we must model p(X) using a function (such as a log function used in logistic regression) that gives outputs between 0 and 1 for all values of X.

Logistic regression models the probability value (Figure 16). For example, the probability of the CD4 cells >50 cells/mm3 given the predictor, such as Prior ZDV use.

Pr(CD4_Stratum = >50|independent variable)

Figure 16: Pr = probability.

The logistic function (Figure 17) models p(X) and gives outputs between 0 and 1 for all values of X. The right-hand side illustrates the fit of the logistic function model to the data, never below 0 or above 1 (James, Witten, Hastie, Tibshirani, 2013).

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}.$$

Figure 17: Logistic function model

Logistic regression uses a method called *maximum likelihood* to fit the logistic regression model to the data (Figure 18). Therefore, the predicted probabilities will always fall between 0/1 allowing a sensible prediction. The quantity p(X)/[1-p(X)] is called the *odds*. Odds can take on any values between 0 and infinity. Values of the odds close to 0 indicate low probability. Higher values indicate a higher probability (James, Witten, Hastie, Tibshirani, 2013).

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

Figure 18: Logistic regression model formula. Left-hand side shows the log odds and the right-hand side shows increasing X by one-unit changes the log odds by $\beta 1$.

Model Fitting

Logistic Regression

The key steps used to fit the model were:

Step 1: To make sure the results were reproducible by using the set.seed command

Step 2: To split the data into 70% training data and 30% test data. Inspect the results via the str. Command (Figure 21).

```
> str(train.data)                          > str(test.data)
'data.frame':   751 obs. of  11 variables:'data.frame':   323 obs. of  11 variables:
```

Figure 21: Results of splitting data into training data (left hand side) and test data (right hand side)

Step 3: Build the model using the dependent (target) variable all the independent variables (Figure 21). CD4_Stratum is a dichotomous (binary) variable of (<=50 or >50 cells/mm3).

model<-glm(CD4_Stratum~., family=binomial, data=train.data)

Figure 21: shows the dependent variable as CD4_Stratum and all independent variables

Step 4: Print and inspect the coefficient and residual deviance from this first iteration of training set data

Step 5: Output and inspect a summary of the first iteration of the model of training set data

Step 6: Output and inspect the coefficients and intercept of the first iteration the model of training set data

Step 7: Output and inspect the first 10 estimated values of training set data

Step 8: Output and inspect the confusion matrix for the training set data

Step 9: Visualize the results for the first iteration of the training data using the:

a) Plot model: use the residual plot (diagnostic plot_ to plot the predicted values versus the residuals

b) Lines: predict the function line in the plot

c) Abline: add the horizontal line where residuals = 0

Step 10: Create and inspect the minimal adequate model. Review key output metrics

Step 11: Check for issues within the model such as collinearity of variables.

Step 12: Experiment and iterate over the input (independent variables). Use decision making logic as it pertains to the objective or question being asked of the data.

Step 13: Do a final summary and inspection of the model in relation to the stated objective.

Results

Output

The default parameters of the model included all independent variables, and the target (dependent) variable was CD4_Stratum. CD4_Stratum is a dichotomous (binary) variable of (<=50 or >50 cells/mm3).

Results of the summary model can be seen in Figure 22. Results reveal significant difference via the Pr(>|z|) two tailed p-values for several variables. At the p<.05 level patients whose race was Hispanic (coefficient) showed significant differences in CD4_stratum levels (dependent/target variable). Significant differences are found at >.05 level for time; censor 1, which is AIDS defining diagnosis or death; race of black, patients with hemophilia, patients with Karnofsky Performance scale scores of 90 and 100, and those with prior ZDV drug usage.

Null deviance is the deviance when all independent variables are zero. If the model had just an intercept the deviance is 978.69 on 750 degrees of freedom. Null deviance is also Chi squared. *Residual deviance* is the deviance after adding the independent variables to the model. The difference between the null and residual deviance needs to account for the *sample size*. The Akaike Information Criterion (AIC) function is 2K – 2 (log-likelihood). A lower AIC is a better-fit model.

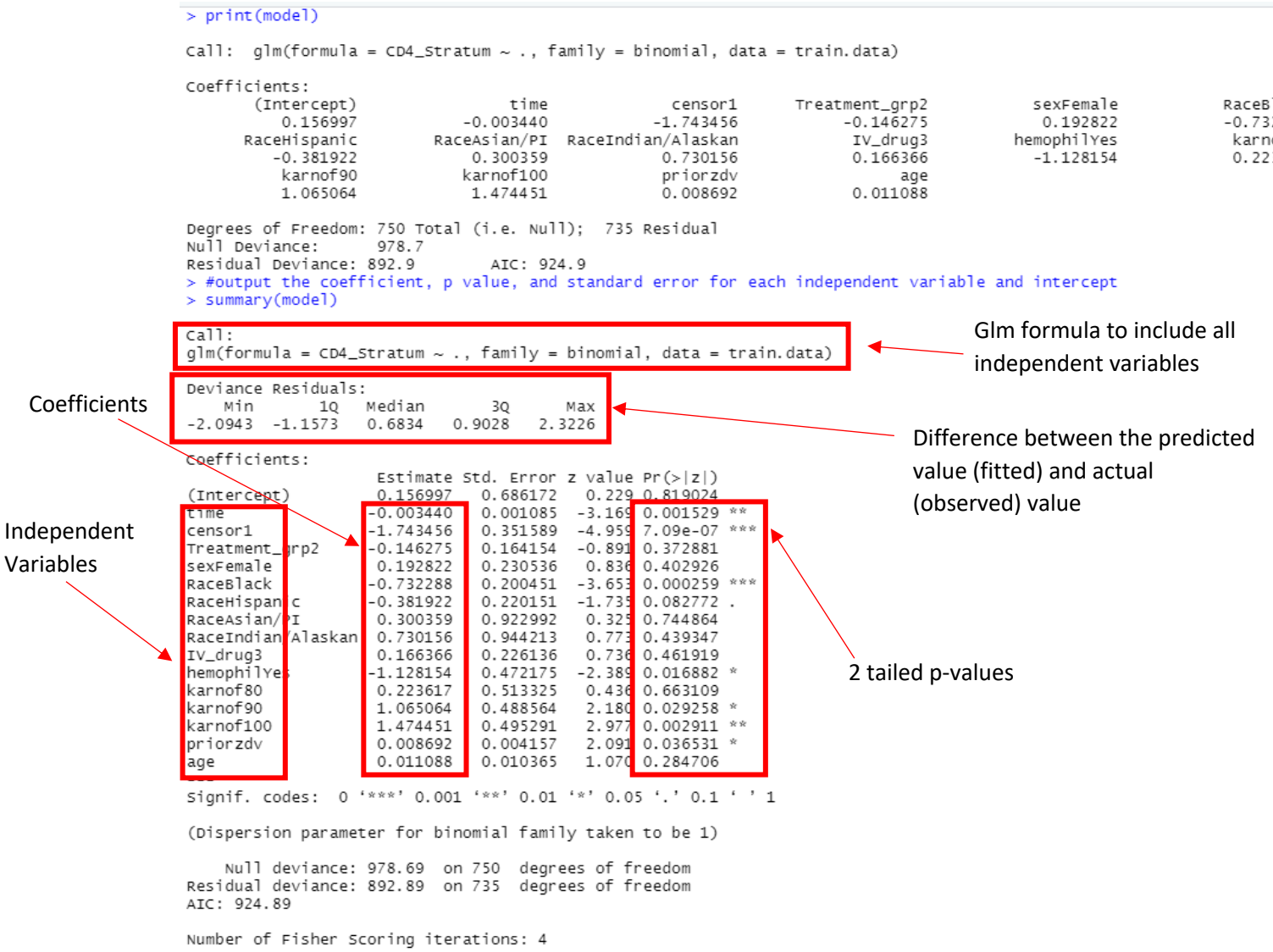


Figure 22: Initial model output results

Figure 23 shows the computed intercept and coefficients for the odds ratio. For example. The coefficient for time means that increasing time by 1 increases the odds ration by 0.9965663 if other variables remain constant.

Figure 23: Output of the coefficients and intercept

In summary, there are significant differences between the dependent variable, CD4 cell levels and various independent predictor variables such as race (e.g. black, Hispanic), health status (e.g. hemophiliac) and level of functioning (Karnofsky scale values). This is important for medical personnel in determining prescreening factors for studies with AIDS patients. The results may also inform time and disease trajectory based on CD4 levels which can inform medications and improve quality of life for patients with AIDS. Therefore, the results of the logistic regression, via the summary analysis, met the stated objective.

Model Properties

After much iteration and inspection of results, the stepwise regression was conducted to determine the minimum adequate model. Five steps were completed by the regression, whereby, at each step the variable with the lowest AIC is removed until only significant variables remained.

Results of the minimum adequate model (Figure 24) show the remaining independent variables after the final iteration to include time, censor1 (AIDS diagnosis or death), Race of back, Hispanic, Asian, Pacific Islander, Indian or Alaskan, positive hemophiliacs, Karnofsky Performance scale scores of 80, 90, 100 and prior use of the DZV drug.

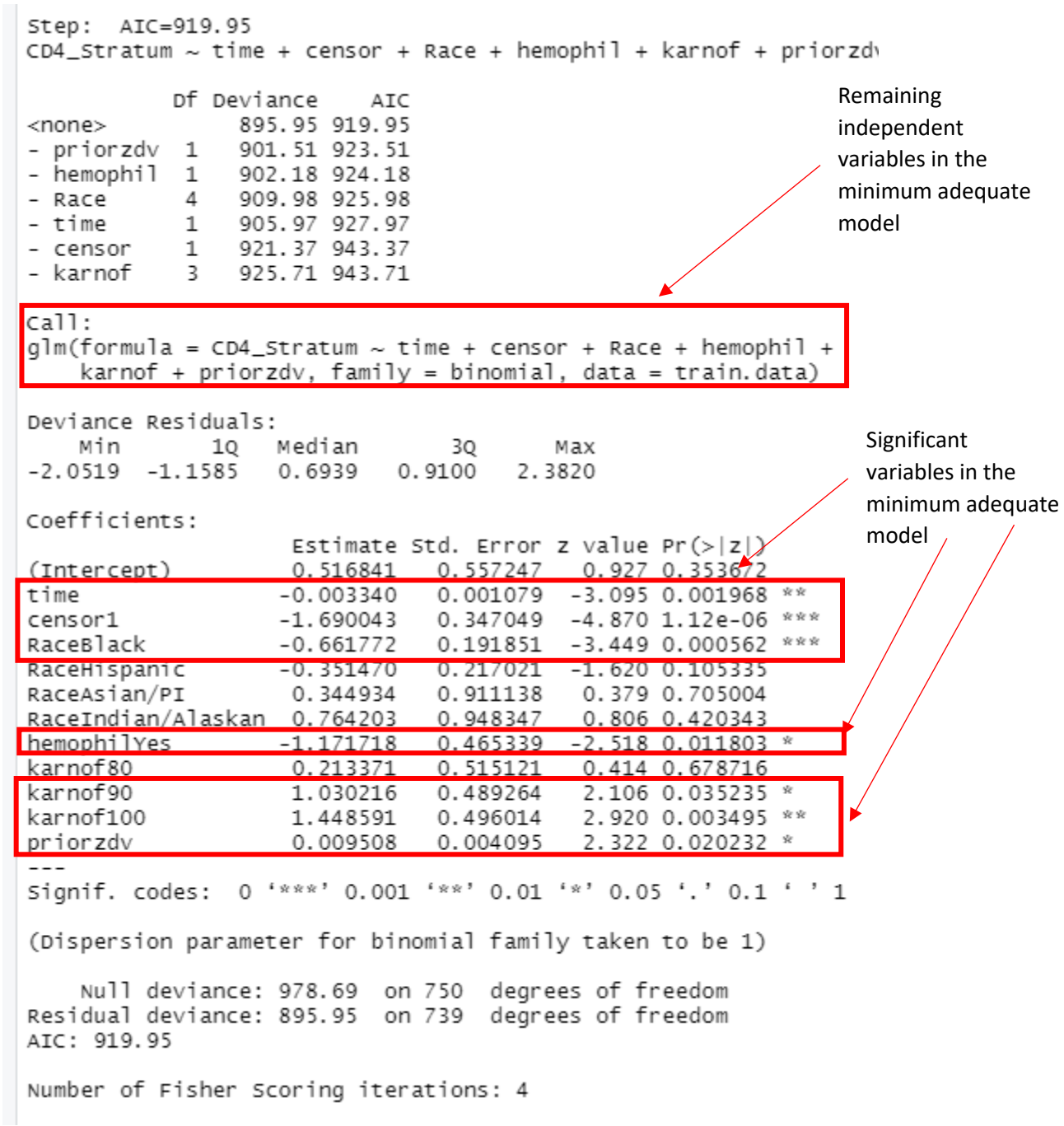


Figure 24: Minimum adequate model

Coefficients in logistic regression are unknown and are estimated based on the training data. The maximum likelihood approach is used to evaluate the coefficients. To evaluate the usefulness of a coefficient we use the p-value. The p value is a quantity that allows us to reject the null hypothesis. The null hypothesis states there are no relationship between the dependent

and independent variables. The threshold for p value is the lowest significant level in which the null hypothesis can be rejected. A two-tailed p-value simply means that the critical area of the distribution is two sided, meaning each side of the distribution is cut in half at 2.5% rather than a one tailed where the distribution is one way and accepted at 5%. Results revealed significant differences and therefore we must reject the null hypothesis.

Evaluation

Evaluate the model on the test data: Using the predict command the probability was predicted for the test data. Figure 25 shows the first 10 instances in the test data set. These values are then rounded to the largest integer and tabulated in the confusion matrix. For instance, number 5 in the dataset would reflect a probability of 1 as the value 0.7053250 is closest to 1 and not 0.

```
> #display the first 10 estimated values for the test data
> predict (model, test.data, type="response")[1:10]
        5        14        16        27        29        30        37        41        42        52
0.7053250 0.1154384 0.7644021 0.7815220 0.8608671 0.2520530 0.7060583 0.7094705 0.6804571 0.4934555
> |
```

Figure 25: Predicted probabilities of the first 10 instances of test data

Confusion Matrix (Figure 26): results from the confusion matrix command show how many AIDS patient records in the test data have each predicted CD4_stratum levels of either <=50 or >50 cells/mm3. The number of correctly classified instances in the test data set = 47 + 151 = 198. The number of misclassified instances in the test data set = 99 + 26 = 125. The total number of instances in the test dataset = 198 + 125 = 323 (which corresponds to the str command output in Figure 21 of test data observations). The classification accuracy is the sum of numbers on diagonal/sum of all numbers = 198/323 = ***61.3%*** *classification accuracy.*

```
> #confusion matrix for the test data
> table (mypredictions, test.data$CD4_Stratum)

mypredictions <=50 >50
            0   47   26
            1   99  151
```

Figure 26: Confusion Matrix

True Positives: these are values that correctly classify the patients belonging to the positive CD4_Stratum (i.e. <=50). The number of true positives in the dataset is 47 and is found at the top left cell of the confusion matrix. True positive values are also known as a *sensitivity* measure.

True Negative: these are the values that correctly classify patients with negative (or >50) CD4_Stratum levels. A result that appears negative (i.e. in the >50) when it should not. The number of true negatives in the dataset is 151 and is found at the bottom right of the confusion matrix. True negatives are also known as a *specificity* measure.

False Positive: also known as a *Type 1 error* occur when the null hypothesis is incorrectly rejected. The creates a "false positive" that leads to a conclusion that the alternate hypothesis is true when it is not. The number of false positives in the dataset is 99 and is found in the bottom left of the confusion matric. Therefore, 99 patients in this dataset may be misclassified as having a significant difference when there is not one.

False Negative: also known as a *type II error* is the non-rejection of a false null hypothesis. Whereby a true difference is not found. The number of false negatives in the dataset is 26 and is found at the top right of the confusion matrix.

Receiver Operating Curve (ROC): is a visualization that displays the false positive rate (specificity; X-axis) and the true positive rate, (sensitivity; Y-axis). The ROC shows the trade-off

that occurs between these two measures. The closer the curve to the top left corner the higher the probability that the model correctly predicts the >50 cells/mm3 CD4 levels. When the curve is close to the 45-degree (red) diagonal line the probability that the model correctly predicts the >50 cells/mm3 is low.

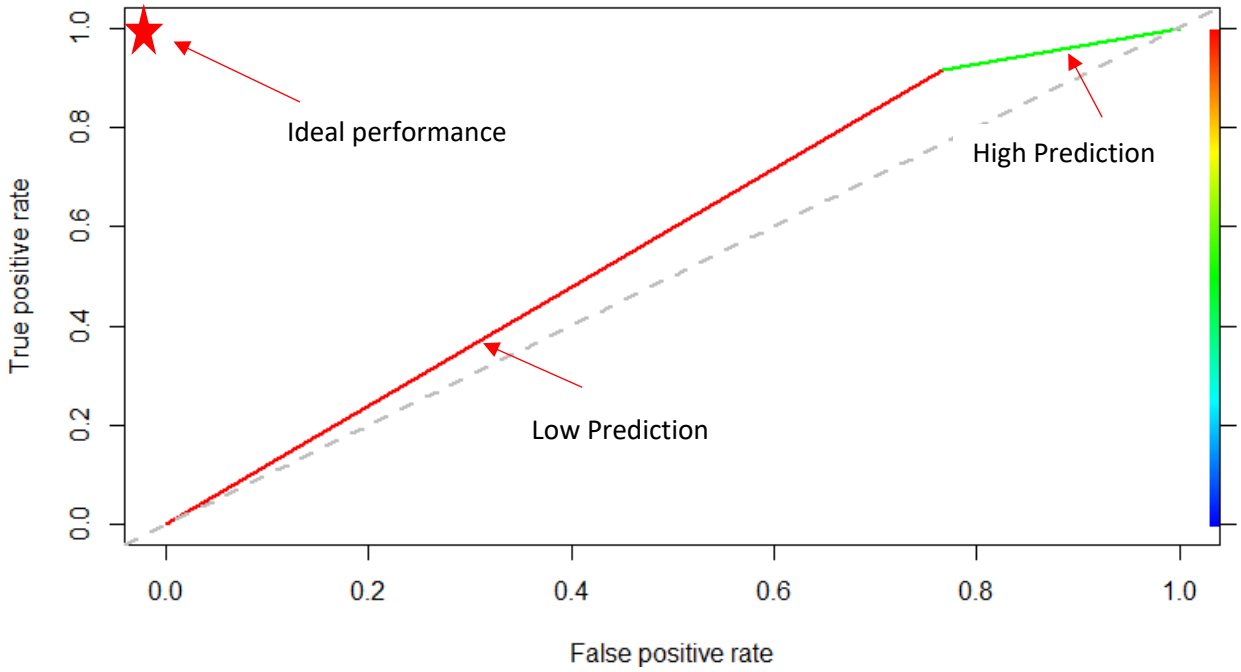


Figure 27: Receiver Operating Curve (ROC) illustrating trade-off between false positive rate and true positive rate.

Area Under the Curve (AUC): The AUC is used to summarize the performance of each classifier (<=50 and >50) into a single measure. The AUC is a measure that is equivalent to the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance (Chan, 2020). It is equivalent to a Wilcoxon rank sum statistic. The high AUC score indicates better classification. AUC score for this dataset was 0.57 which is moderate.

Effect plots (Figure 28): are used to visualize the change in probability of the dependent variable (CD4_Startum) s the value of the independent variable changes. If the independent

variables are numeric then the light blue bands show the 95% confidence interval. For instance, the time effect plot shows that as days increase from 200 to 300 and then 400 days the CD4_Stratum probability value decreases from .7 to .6 to .5. If the independent variable is categorical, then the pink bar shows the 95% confidence interval. For instance, the censor effect plot shows a confidence interval between 0.3 and 0.9 for responses categorized as 0 (i.e. "otherwise") and 0.0-0.6 for responses categorized as 1 ("AIDS diagnosis or death). These plots are a good way to show how the probability of the dependent variable changes across the distribution of the independent variables.

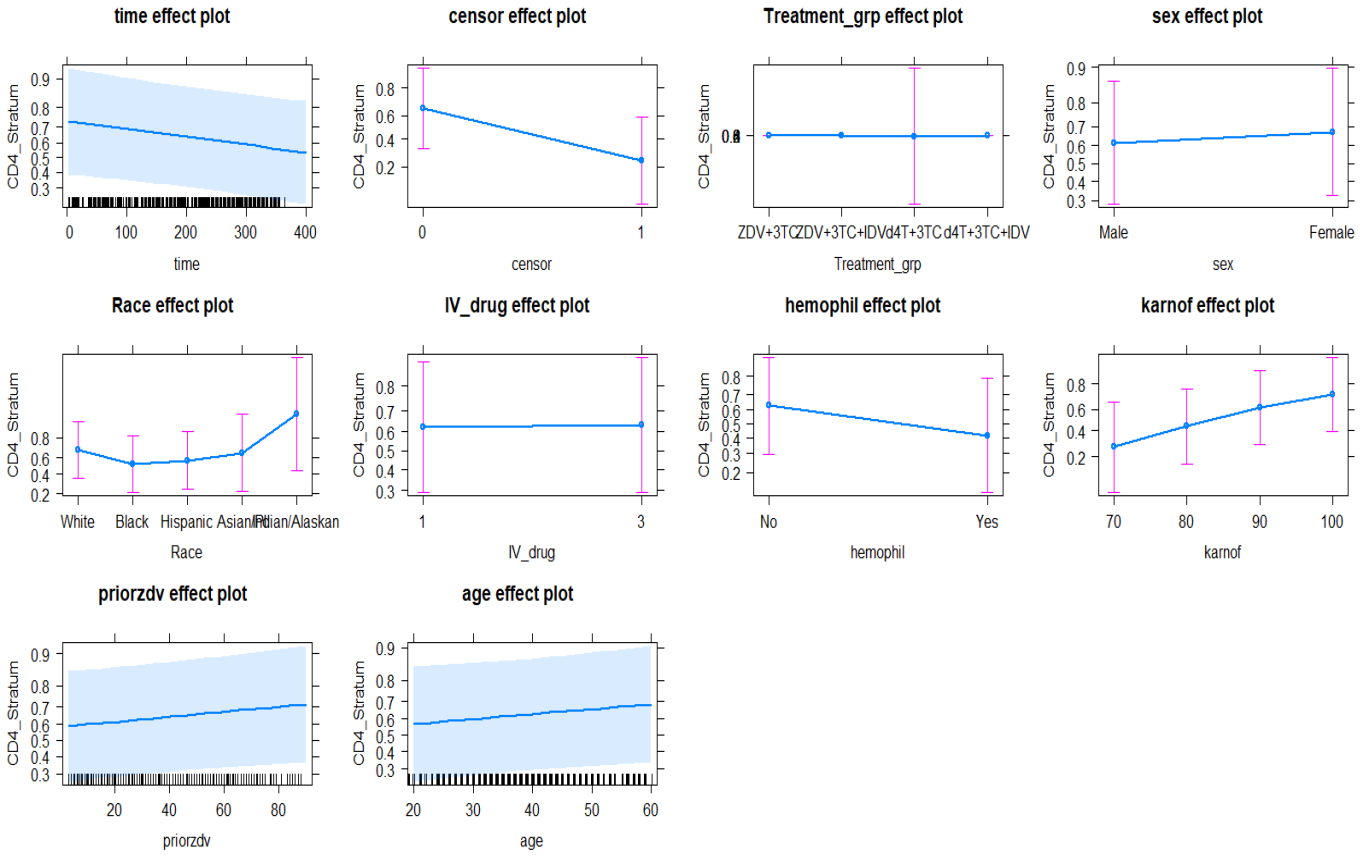


Figure 28: Effect plots showing levels of confidence for each independent variable

Diagnostics

The residual plot also known as the diagnostic plot is shown in Figure 29. Predicted values versus residual are represented by blue dots. The prediction line (or prediction function) is the solid black line, and the horizontal line is where the residuals = 0. The points form two curves because the CD4-Stratum levels (dependent variable) have two possible outcomes 0 (<=50 cells/mm3) and 1 (>50 cells/mm3). The residuals for the <=50 are negative which form the bottom line. The residual values for >50 are positive which form the top line.

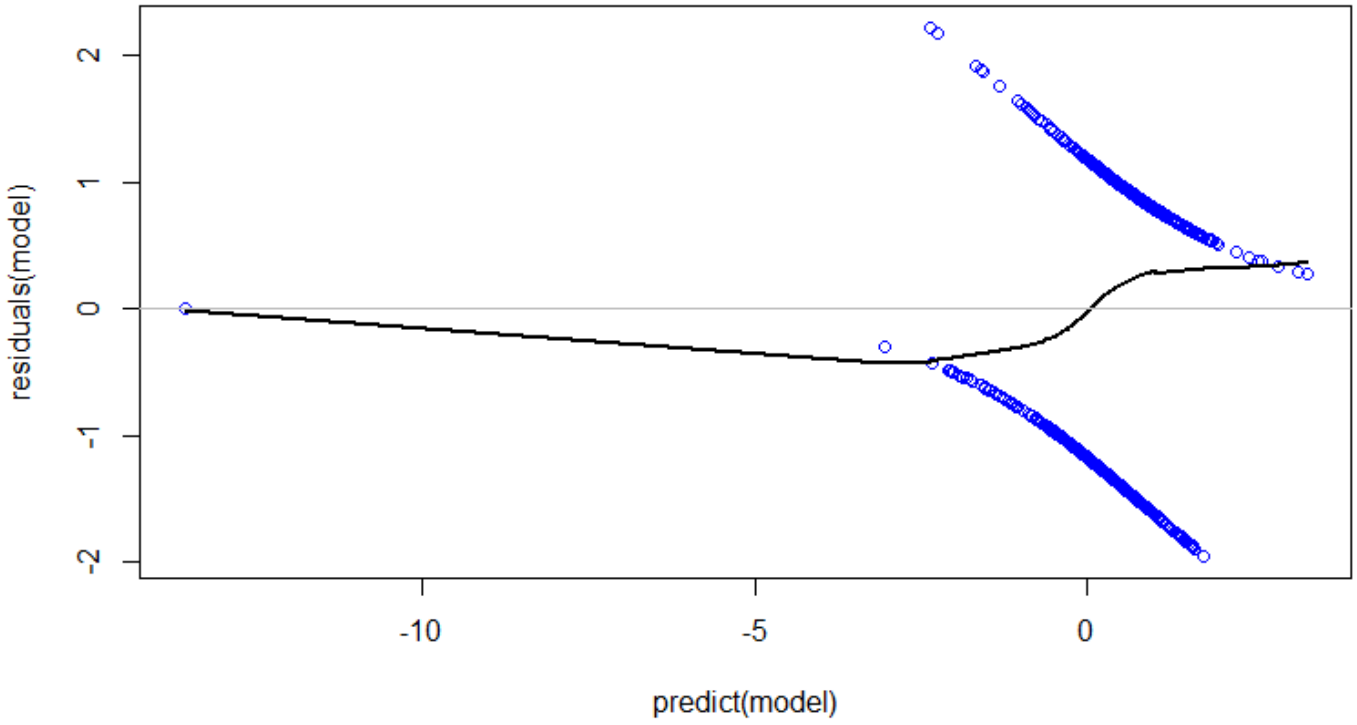


Figure 29: Residual or Diagnostic plot for the predicted versus residual values.

Logistic regression has several assumptions including the need for the dependent variable structure to fall between 0 and 1. This assumption is met in the 0,1 categorization of the CD4_Stratum dependent variable. Second the observation independence, or absence of multicollinearity, in the independent variables. This was examined and all variables with high

colinear values were removed (See Appendix A). Therefore, this assumption was met (Schreiber-Gregory, 2018).

Finally, logistic regression typically requires a large sample size. A general guideline is that there is a minimum of 10 cases with the least frequent outcome for each independent variable (Schreiber-Gregory, 2018). For each independent variable, a count was reviewed, if the cases were below 10, they were removed. This was found to be the case for IV_drug and Treatment_grp. Therefore, these factor levels were reduced, and the assumption was met.

Conclusion

Summary

The objective of the analysis was to determine factors that influence the CD4 blood levels, specifically whether CD4 levels are above or below 50 cells/mm3 based on various input (independent) variables. If CD4 blood levels could be explained, this may help medical personnel in determining prescreening factors for studies with AIDS patients. Drug studies are very expensive and often many drugs are left on the shelf because pre-screening of patients was not satisfactory. This affects the accuracy of the results of the drugs (Fogel, 2018). Effectively predicting CD4 levels may also inform disease trajectory, medications and ultimately quality of life of AIDS patients (Hirsch et al., 1999).

Key findings revealed significant differences in several of the input variables in relation to CD4 levels above and below 50 cells/mm3. Specifically, race (e.g. black, Hispanic), health status (e.g. hemophiliac) and level of functioning (Karnofsky scale values). The minimum adequate model revealed an accuracy of 63% with an AUC of .57. These results were moderate in determining the sensitivity and specificity of the model.

With these results it was deemed appropriate to investigate another model for comparison

to see if the accuracy, sensitivity, or specificity could be improved (see Appendix B). The Naive Bayesian model was chosen as a comparative model because it is also a supervised classification method. It should be used when the independent variables are independent, as is the case in this dataset. Furthermore, the Naive Bayesian model estimates class probability that could easily be compared to the logistic regression output.

The results from the Naïve Bayesian model were very similar to those of the binary logistic regression. Accuracy was only slightly higher at 63.5%. Sensitivity in the Naïve Bayesian model (33) was slightly lower than the logistic regression (47) and vice versa for specificity values (151 for logistic regression and 172 for Naïve Bayesian model).

In conclusion, the results guide medical professionals by being aware of certain significant differences that play a role in blood levels to include current health status for example. However, it should be cautioned that these factors provide the sensitivity and specificity needed in medical research to definitively confirm pre-screening requirements for a study that may use drugs to improve CD4 levels in the blood.

Limitations

Several limitations were found while conducting the analysis. First, the dataset, a logistic regression is not well suited to factor variables with too many levels. Race had six levels, the Karnofsky scale had four levels.

A second limitation is that the categorical variables within the dataset could have been categorized in a more meaningful manner. Specifically, the Karnofsky scale could likely be categorized into two levels, low scores (70 & 80) and high scores (90 & 100).

A third limitation of the dataset was the lack of information on the stage of the disease. There are four stages to AIDS and these stages have been associated with different CD4 cell

levels (Farhadian, Mohammadi, Mirzaei, & Shirmohammadi-Khorram, 2021).

A fourth limitation of the dataset was the lack of information regarding the length of time since the HIV diagnosis. Length of time living with HIV/AIDS has been shown to impact CD4 cells (Farhadian, Mohammadi, Mirzaei, & Shirmohammadi-Khorram, 2021).

A fifth limitation was found in the skewness of some variables. For instance, variables like censor were heavily skewed (in proportions) and therefore could not be utilized to a great extent for the analysis. Collinearity amongst variables also limited the variables that could have been used for the analysis.

A sixth limitation includes the evaluation of the model. The CD4_Strutum target variable was uneven with 437 in the <=50 cells/mm3 (38%) and 707 in the >50 group (63%). The uneven outcome of the target variable is a limitation in the data. Therefore, evaluation of the model via accuracy may not be the most effective way to examine the ability of the model to meet the objective.

Improvement Areas

This data was collected more than thirty years ago, should the study be redone, the input factors (independent variables) for the study should be re-evaluated. Since the study additional input factors such as CD8 cell counts have been associated with greater increases in CD4 cell counts (Smith et al, 2004). Co-infections such as Tuberculosis and clinical disease stage have also been found to influence CD4 cells since this data was collected (Farhadian, Mohammadi, Mirzaei, & Shirmohammadi-Khorram, 2021). These additional variables, along with improvements in the factor levels could influence the accuracy, sensitivity, and specificity of the model outcomes. It would be worthwhile to see if these variables could be retrospectively found in medical database to add to the datasheet, if possible.

Given the data retrieved, and the significant differences in variables, but only moderate accuracy, it may be worth exploring other analytical approaches. One example may be a random forest method for classification.

One of the most significant improvements can be in the evaluation of the model. This can be done by looking not only at accuracy, but also examining F1 score the Positive Predictive Value (PPV) and Negative Predictive Value (NPV) and confidence intervals. Calculating and evaluating the precision and recall would help in furthering the understanding of the model output, its benefits for application and limitations.

Another improvement can be made by comparing the test and training data for the purpose of examining overfitting and understanding the generalized error in the model.

Finally, having more data for the model to train on may yield significant improvements in the results.

Appendix A

Collinearity Table

```
> # Check for collinearity
> cor(actg320)
                      id         time       censor       time_d     censor_d     Treatment Treatment_grp   CD4_Stratum
id            1.0000000000  0.004364678  0.015221704  0.014354707  0.042349932  0.0380942215   0.052504930  0.0230692398
time          0.0043646784  1.000000000 -0.420894538  0.827008972 -0.198361937  0.0739778408   0.051143155  0.0283809916
censor        0.0152217045 -0.420894538  1.000000000  0.027994476  0.503965754 -0.0934810169  -0.094360268 -0.1965616480
time_d        0.0143547072  0.827008972  0.027994476  1.000000000 -0.190062703  0.0388435897   0.013478100 -0.0656531272
censor_d      0.0423499323 -0.198361937  0.503965754 -0.190062703  1.000000000 -0.0580744937  -0.058260535 -0.0973048549
Treatment     0.0380942215  0.073977841 -0.093481017  0.038843590 -0.058074494  1.0000000000   0.979971948 -0.0002579657
Treatment_grp 0.0525049298  0.051143155 -0.094360268  0.013478100 -0.058260535  0.979971948   1.000000000 -0.0062311478
CD4_Stratum   0.0230692398  0.028380992 -0.196561648 -0.065653127 -0.097304855 -0.0002579657  -0.006231148  1.0000000000
sex           -0.0156627024 -0.059536939 -0.013941640 -0.079197157  0.007440888  0.0287107904   0.032325302  0.0107700373
Race          0.0046310842 -0.072551887  0.020708009 -0.063057284  0.030467854 -0.0149566750  -0.011205021 -0.0634410553
IV_drug       0.0314816803  0.030672548 -0.039805032  0.005223925  0.006629185 -0.0084481622  -0.003309372  0.0100303881
hemophil      0.0006142293 -0.015004933  0.001478615 -0.015386270  0.007129975 -0.0349571810  -0.035940529 -0.0276123793
karnof        0.0154150553  0.053139977 -0.189245329 -0.021818049 -0.154366109 -0.0085602887  -0.005683129  0.1973588766
CD4_base      0.0467346431  0.004297130 -0.212640125 -0.106558569 -0.093809296  0.0308142854   0.027141182  0.7188255723
priorzdv      0.0127378303  0.035778886 -0.016702474  0.040222684 -0.033914632  0.0024634129  -0.004538436  0.0481084120
age           -0.0058296200  0.074172494  0.064168722  0.080586251  0.119609696  0.0028549309  -0.004125405  0.0528315363
                       sex        Race      IV_drug      hemophil       karnof     CD4_base     priorzdv          age
id            -0.015662702  0.004631084  0.031481680  0.0006142293  0.015415055  0.04673464  0.012737830 -0.005829620
time          -0.059536939 -0.072551887  0.030672548 -0.0150049330  0.053139977  0.00429713  0.035778886  0.074172494
censor        -0.013941640  0.020708009 -0.039805032  0.0014786146 -0.189245329 -0.21264012 -0.016702474  0.064168722
time_d        -0.079197157 -0.063057284  0.005223925 -0.0153862700 -0.021818049 -0.10655857  0.040222684  0.080586251
censor_d       0.007440888  0.030467854  0.006629185  0.0071299747 -0.154366109 -0.09380930 -0.033914632  0.119609696
Treatment      0.028710790 -0.014956675 -0.008448162 -0.0349571810 -0.008560289  0.03081429  0.002463413  0.002854931
Treatment_grp  0.032325302 -0.011205021 -0.003309372 -0.0359405289 -0.005683129  0.02714118 -0.004538436 -0.004125405
CD4_Stratum    0.010770037 -0.063441055  0.010030388 -0.0276123793  0.197358877  0.71882557  0.048108412  0.052831536
sex            1.000000000  0.113356022  0.025587648 -0.0678594360  0.041376755  0.01770048 -0.008623092 -0.126902810
Race           0.113356022  1.000000000  0.136451268 -0.0052996050 -0.036132058 -0.04134113 -0.043652370 -0.074836980
IV_drug        0.025587648  0.136451268  1.000000000 -0.0628061218 -0.056082933  0.04122276 -0.009991540  0.071185581
hemophil      -0.067859436 -0.005299605 -0.062806122  1.0000000000  0.055289745 -0.01868535  0.121679195 -0.100336052
karnof         0.041376755 -0.036132058 -0.056082933  0.0552897453  1.000000000  0.16495685 -0.022406716 -0.121612761
CD4_base       0.017700484 -0.041341135  0.041222762 -0.0186853475  0.164956851  1.00000000  0.091585168  0.038543780
priorzdv      -0.008623092 -0.043652370 -0.009991540  0.1216791955 -0.022406716  0.09158517  1.000000000  0.126978191
age           -0.126902810 -0.074836980  0.071185581 -0.1003360519 -0.121612761  0.03854378  0.126978191  1.000000000
>
```

Appendix B

Naïve Bayesian Classification model

The rationale for the use of a Naïve Bayesian methodology includes:

a) An objective is to predict class membership CD4 cells (<=50 or >50 cells/mm3)

b) The classifiers are continuous and categorical independent variables which are well suited to a Naïve Bayesian Classification methodology.

c) A second methodology to compare to the moderate findings int eh logistic regression output

Preprocessing

Discretization of variables. The Naïve Bayesian method requires all variables in the dataset to be discrete or factors. Therefore, priorzdv and CD4_base variable were discretized in six equal breaks. This changed these variables from continuous "num" variables to factors.

Algorithm Intuition

The Naive Bayesian Classification method is used to estimate class probability when all the independent variables are independent. This method assume that the independent variables have an equal weight on the dependent variable.

The purpose of this model is to examine the probability of a patient falling into the "<=50 cells/mm3" category or the ">50 cells/mm3" category. This classification is based on the input, predictive variables. The Naïve Bayesian method is a supervised learning method as we have a known outcome (<=50 or >50 cells/mm3) within the dataset.

The principle of the Naïve Bayesian classification is modeling the probability of classification into a group based on a probability algorithm (Figure 19).

$$p(cj \mid d)$$

Figure 19: *cj* is probability of class, *d* is the observed data

Bayesian Classifiers use the Bayes theorem (Figure 20). Whereby, the method tries to compute the probability of an instance being in class.

$$p(cj \mid d) = \frac{p(d \mid cj)\, p(cj)}{p(d)}$$

Figure 20: Bayes theorem. $p(cj \mid d)$ where p = probability, *cj* is class, d is instance.

The logic behind the Naive Bayesian Classification method is an expectation that we can estimate class membership when all the independent variables are independent, *and* the method assume that the independent variables have an equal weight on the dependent variable.

Model Fitting

The key steps used to fit the model were:

Step 1: Load additional library packages

Step 2: Discretize any num variables

Step 3: To make sure the results were reproducible by using the set.seed command

Step 4: To split the data into 70% training data and 30% test data. Inspect the results via the str. Command.

Step 5: Build the model using the dependent (target) variable all the independent variables.

Step 6: Output and inspect the default version of the model to include the A-priori probabilities and conditional probabilities

Step 7: Output and inspect the default version of the model confusion matrix for the training set of data

Step 8: Output and inspect the default version of the model confusion matrix for the test set of data

Step 9: Visualize the output for the first iteration of the data using a mosaic plot

Step 10: Experiment and iterate over the input (independent variables). Use decision making logic as it pertains to the objective or question being asked of the data.

Step 11: Do a final summary and inspection of the model. Inspect the model in relation to the objective.

Step 12: Compare the two models and determine which one better explains the stated objective. logistic regression

Results (in summary)

Please note: the following is a summary of the results from the Naïve Bayesian model. It is only for the purpose of comparison to the logistic regression to see if accuracy, sensitivity, and specificity could be improved. It is not to replace the analysis via the logistic regression from the report. It is just for my own edification of how to use and interpret this model. Thank you.

```
> #build the model and store in a variable model
> model<-naiveBayes(CD4_Stratum~., train.data)
> #output the model
> print(model)

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
     <=50       >50
0.3812665 0.6187335

Conditional probabilities:
      time
Y          [,1]     [,2]
  <=50 225.6055 95.73529
  >50  233.8465 85.22683

      censor
Y                0          1
  <=50 0.83737024 0.16262976
  >50  0.96375267 0.03624733

      Treatment_grp
Y          ZDV+3TC ZDV+3TC+IDV     d4T+3TC d4T+3TC+IDV
  <=50 0.525951557 0.467128028 0.003460208 0.003460208
  >50  0.526652452 0.471215352 0.000000000 0.002132196

      sex
Y           Male    Female
  <=50 0.8477509 0.1522491
  >50  0.8294243 0.1705757

      Race
Y            White       Black    Hispanic    Asian/PI Indian/Alaskan Other/Unknown
  <=50 0.439446367 0.339100346 0.200692042 0.017301038    0.003460208   0.000000000
  >50  0.558635394 0.253731343 0.159914712 0.012793177    0.014925373   0.000000000

      IV_drug
Y              1         3
  <=50 0.8477509 0.1522491
  >50  0.8464819 0.1535181

      hemophil
Y             No        Yes
  <=50 0.96193772 0.03806228
  >50  0.97654584 0.02345416
```

Shows all independent variables are factors

Probability that the dependent variable is yes (or <=50) and the probability of No >50

Probabilities for broken down by variable within each factor

```
      karnof
Y              70         80         90        100
  <=50 0.04844291 0.23529412 0.47404844 0.24221453
  >50  0.01066098 0.11513859 0.47547974 0.39872068

      priorzdv
Y          [3,7)     [7,12)    [12,20)    [20,31)    [31,48)    [48,89]
  <=50 0.1107266 0.1833910 0.2179931 0.1833910 0.1660900 0.1384083
  >50  0.1236674 0.1748401 0.2046908 0.1385928 0.1513859 0.2068230

      age
Y          [,1]     [,2]
  <=50 37.93772 8.167279
  >50  38.34755 8.100642
```

Confusion Matrix (Figure 1): results from the confusion matrix command show how many AIDS patient records in the test data have each predicted CD4_stratum levels of either <=50 or >50 cells/mm3. The number of correctly classified instances in the test data set = 33 + 172 = 198. The number of misclassified instances in the test data set = 99 + 26 = 125. The total number of instances in the test dataset = 198 + 125 = 323 (which corresponds to the str command output in Figure 21 of test data observations). The classification accuracy is the sum of numbers on diagonal/sum of all numbers = 198/323 = ***63.5%*** *classification accuracy.*

```
> #confusion matrix for the test data
> table(predict(model, test.data), test.data$CD4_Stratum)

      <=50 >50
<=50    33   24
>50     94  172
```

Figure 1: Confusion Matrix for Naïve Bayesian Classification model

True Positives: these are values that correctly classify the patients belonging to the positive CD4_Stratum (i.e. <=50). The number of true positives in the dataset is 33 and is found at the top left cell of the confusion matrix. True positive values are also known as a *sensitivity* measure.

True Negative: these are the values that correctly classify patients with negative (or >50) CD4_Stratum levels. A result that appears negative (i.e. in the >50) when it should not. The number of true negatives in the dataset is 172 and is found at the bottom right of the confusion matrix. True negatives are also known as a *specificity* measure.

False Positive: also known as a *Type 1 error* occur when the null hypothesis is incorrectly rejected. The creates a "false positive" that leads to a conclusion that the alternate hypothesis is true when it is not. The number of false positives in the dataset is 94 and is found in the bottom left of the confusion matric. Therefore, 94 patients in this dataset may be misclassified as having a significant difference when there is not one.

False Negative: also known as a *type II error* is the non-rejection of a false null hypothesis. Whereby a true difference is not found. The number of false negatives in the dataset is 24 and is found at the top right of the confusion matrix.

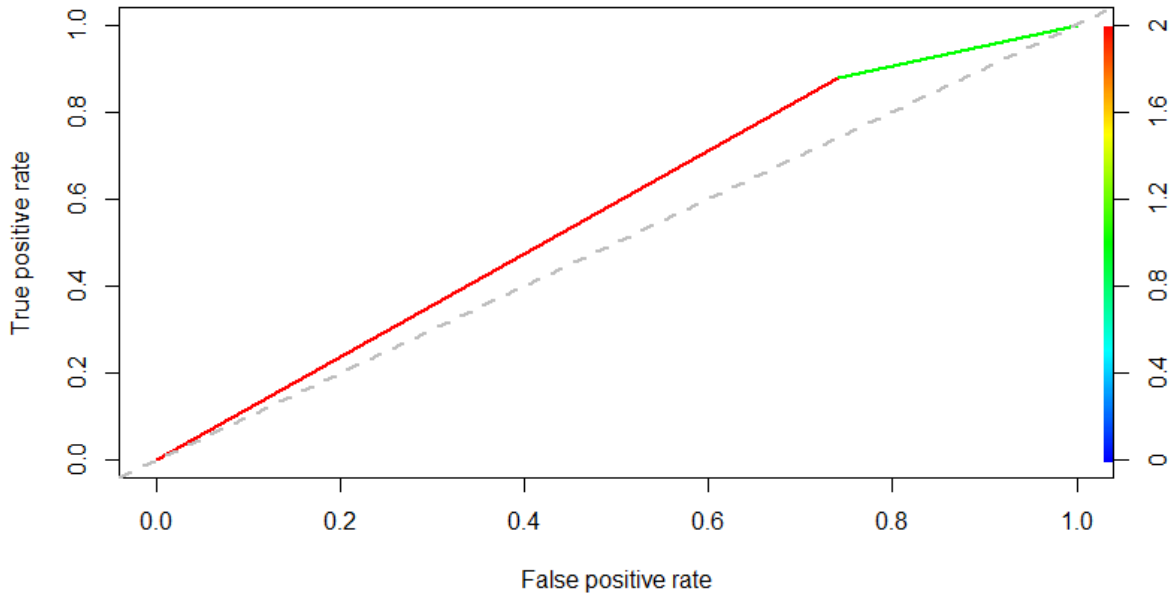


Figure 2: Receiver Operating Curve (ROC) illustrating trade-off between false positive rate and true positive rate.

Create a mosaic plot (Figure 3) to determine the predicted versus actual CD4 values. The mosaic function takes the confusion matric table as the first parameter. Blue represents the proportion of instances with the predicted class = actual class, that is true positives and true negatives. The red color represents the proportion of the misclassified instances, that is false positive and false negative.

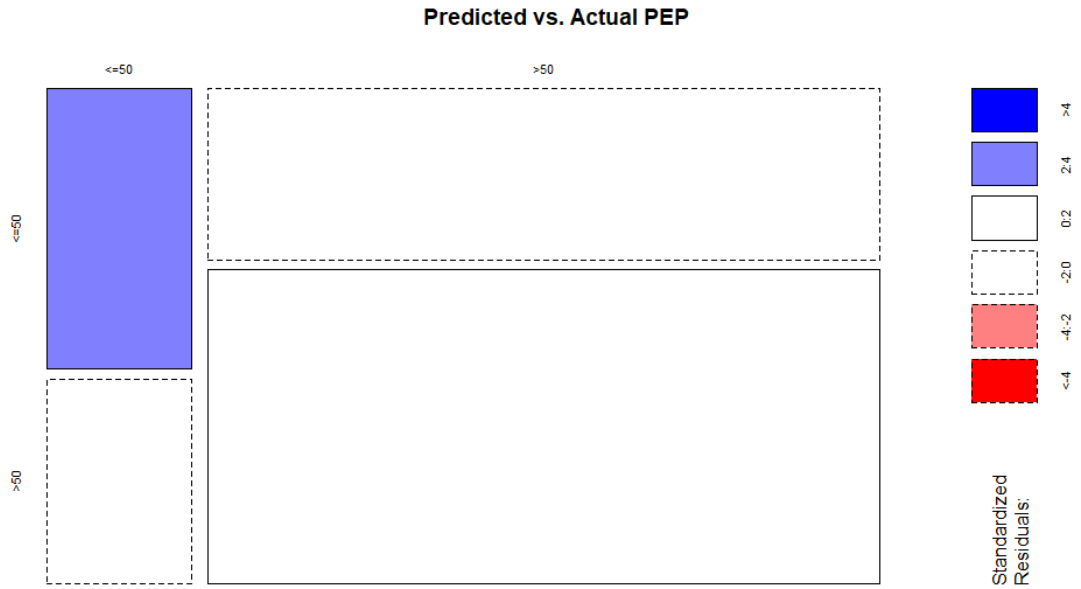


Figure 3: Mosaic plot for predicted versus actual CD4 values.

Conclusion (in summary)

The Naïve Bayesian model did not produce results that we distinctly different from the logistic regression. Accuracy, sensitivity, and specificity were similar.

References (Bayesian classifiers)

Bishop, C., Springer-Verlag, P. (2006). *Pattern Recognition and Machine Learning.* Wiley and Sons, New York, NY.

Duda, R.O., Hart, D. *Pattern Classification by Stork*. Wiley and Sons, New York, NY.

Appendix C

References

Cichocki, M. (2020). *CD4 T-Cells and Why They Are Important.* Retrieved from: https://www.verywellhealth.com/what-are-cd4-t-cells-49354

Coplan, P.M., Cook, J.R., Carides, G.W., et al. (2004). Impact of indinavir on the quality of life in patients with advanced HIV infection treated with zidovudine and lamivudine. *Clinical infectious diseases: An official publication of the Infectious Diseases Society of America*. 39(3):426-433. doi:10.1086/422520

Demeter, L.M., Hughes, M.D., Coombs, R.W., Jackson, J.B., Grimes, J.M., Bosch, R.J., Fiscus, S.A., Spector, S.A., Squires, K.E., Fischl, M.A., Hammer, S.M. (2001). Predictors of virologic and clinical outcomes in HIV-1-infected patients receiving concurrent treatment with indinavir, zidovudine, and lamivudine. AIDS Clinical Trials Group Protocol 320. *Annals of Internal Medicine 4;*135(11):954-64.

Chan, C. (2021). *What is a ROC Curve and How to Interpret It.* Retrieved from: Display R Blog: https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/

Enders, F.B. (2020). *Collinearity.* Retrieved from: https://www.britannica.com/topic/collinearity-statistics

Farhadian, M., Mohammadi, Y., Mirzaei, M. *et al.* (2021). Factors related to baseline CD4 cell counts in HIV/AIDS patients: comparison of poisson, generalized poisson and negative binomial regression models. *BMC Research Notes 14,* 114 https://doi.org/10.1186/s13104-021-05523-w

Feferman, S. (1989). *The Number Systems: Foundations of Algebra and Analysis*, AMS Chelsea, ISBN 0-8218-2915-7.

Fogel D. B. (2018). Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. *Contemporary clinical trials communications*, *11*, 156–164. https://doi.org/10.1016/j.conctc.2018.08.001

Gulick, R.M,, Mellors, J.W., Havlir, D., Eron, J.J., Gonzalez, C., McMahon, D., et al. (1997). Treatment with indinavir, zidovudine, and lamivudine in adults with human immunodeficiency virus infection and prior antiretroviral therapy. *New England Journal of Medicine; 337*:734-9.

Hammer, S.M., Squires, K.E., Hughes, M.D., Grimes, J.M., Demeter, L.M., Currier, J.S., Eron, J.J. Jr, Feinberg, J.E., Balfour, H.H. Jr, Deyton, L.R., Chodakewitz, J.A., Fischl, M.A. (1997). AIDS Clinical Trials Group 320 Study Team. A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. *New England Journal of Medicine, 11*;337(11):725-33.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining Concepts and Technique, 3rd Ed.* Ch.3. Elsevier Science. New York, NY.

Helford, M. (2020). *General Population – With Some Big "Buts".* Retrieved from: https://www.thebodypro.com/article/hiv-life-expectancy-in-u-s-matches-general-population-with-some-differences

Hirsch. M., Steigbigel, R., Staszewski, S., Mellors, J., Scerpella, E., Hirschel, B., et al. (1999). A randomized, controlled trial of indinavir, zidovudine, and lamivudine in adults with

advanced human immunodeficiency virus type 1 infection and prior antiretroviral therapy. *Journal of Infectious Disease,180*:659-65.

HIV.gov (2021). *U.S. Statistics. Fast Facts.* Retrieved from: https://www.hiv.gov/hiv-basics/overview/data-and-trends/statistics

HIV.gov (2021). *What is AIDS?* Retrieved from: https://www.hiv.gov/hiv-basics/overview/about-hiv-and-aids/what-are-hiv-and-aids#:~:text=A%20person%20with%20HIV%20is,OR

Hosmer, D.W. and Lemeshow, S. and May, S. (2008). *Applied Survival Analysis: Regression Modeling of Time to Event Data: Second Edition,* John Wiley and Sons Inc., New York, NY.

Karnosky Performance Scale. Retrieved from: http://www.npcrc.org/files/news/karnofsky_performance_scale.pdf

Ng, A. (2021). *Introduction to Supervised Learning.* Retrieved from: https://www.coursera.org/lecture/machine-learning/supervised-learning-1VkCb

Peto, R., Pike, M.C., Armitage, P., et al. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. British Journal of Cancer; 34:585-612.

Schag, C.C., Heinrich, R.L., Ganz, P.A. (1984). Karnofsky performance status revisited: Reliability, validity, and guidelines. *Journal of Clinical Oncology; 2*:187-193.

Schreiber-Gregory, D. (2018). *Logistic and Linear Regression Assumptions: Violation Recognition and Control.* Retrieved from: https://www.lexjansen.com/wuss/2018/130_Final_Paper_PDF.pdf

Smith, C.j., Sabin, C.A., Youle, M.S., Kinloch-de Loes, S., Lampe, F.C., Madge, S., Cropley, I., Johnson, M.A., Phillips, A.N. (2004). Factors Influencing Increases in CD4 Cell Counts of HIV-Positive Persons Receiving Long-Term Highly Active Antiretroviral Therapy, *The Journal of Infectious Diseases*, *190,* https://doi.org/10.1086/425075

World Health Organization (2020). *HIV/AIDS Key Facts.* Retrieved from: https://www.who.int/news-room/fact-sheets/detail/hiv-aids

World Health Organization (2020). *HIV/AIDS Living & Managing.* Retrieved from: https://www.webmd.com/hiv-aids/guide/HIV-AIDS-living-managing