“I hereby declare that the work represented in this report represents my individual effort. I

understand that I am encouraged to seek advice and guidance on any course material from my

professor, teaching assistants and fellow classmates, however, I am responsible, solely, for this

written document and my oral presentation of components of this report.”

\_\_\_\_\_\_\_\_\_Sean Vieau\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_10/06/2024\_\_\_\_\_\_\_

(name) (date)

**Project 2**

**Introduction**

The aim of the current study is to assess how treatment response differs for HIV+ patients 2 years after initiating Highly Active Antiretroviral Therapy (HAART) based on hard drug usage (such as heroin or cocaine). This study is of particular scientific interest because it is unclear whether the use of hard drugs inhibits the immune system in humans; treatment strategies may differ based on these results. The researchers are interested in comparing subjects who never used hard drugs to current hard drug users (those that use hard drugs at year 2) or previous hard drug users (those who used drugs at year 0 or 1). Outcomes of interest are: viral load (HIV copies in a mL of blood), CD4+ T cell count (a measure of immunologic health), and aggregate physical and quality of life scores from the SF-36.

The clinical hypothesis is that, if hard drugs inhibit the immune system in humans, subjects who currently or previously used hard drugs will have higher viral load and lower CD4+ T cell counts than those who never used hard drugs. Additionally, the researchers are interested in knowing if potential differences between the drug use groups can be explained by differences in adherence to the treatment regimen. The researchers are agnostic on how quality of life changes after treatment, since side effects of the treatment are significant.

**Method**

**Study Design**

This is a secondary data analysis of the Multicenter AIDS Cohort Study, an ongoing prospective cohort study investigating the natural and treated disease progression of HIV-1 in bisexual men in 4 major cities in the U.S. Measurements for all variables were taken once per year over an 8-year time period; however, the current analysis is only concerned with treatment outcomes after 2 years of HAART. Data was received as a longform .csv file containing 33 columns along with a data dictionary. The main outcomes of interest are viral load, CD4+ T cell count, and aggregate physical and quality of life scores. Adherence to treatment regiment will be investigated as a potential confounder.

Potential covariates of interest include: marijuana usage since last visit and frequency of usage, income, BMI, high blood pressure, diabetes, liver disease stage 3 / 4, kidney disease, frailty related phenotype, total cholesterol, triglycerides, fasting LDL, dyslipidemia, depression score, smoking status, alcohol use since last visit, heroin or opiate use since last visit, intravenous drug use since last visit, race, education at baseline, age, if they took ART at the visit or if they have ever taken it before, and years since initiating ART.

**Statistical Hypotheses**

* Null Hypothesis: The average change in viral load, CD4+ T cell count, and aggregate physical and mental QOL scores will be equal for current, previous, and never hard drug users.
* Alternative Hypothesis: At least one of the hard drug use groups will differ from the never hard drug users in viral load, CD4+ T cell count, or aggregate physical or mental QOL scores.

**Data Management**

All data analysis and management was performed in R version 4.4.1

*Outliers****.*** Boxplots and histograms of viral load, CD4+ T cell count, and physical and mental QOL measures were made to assess for potential outliers at baseline and 2 years. Outliers were then identified using jackknife residuals. Data points +/- 3 SD of the mean with high leverage and influence were considered outliers and removed from the data set as determined by Cook’s D, DFITS, DFBETAS, and hat-values.

*Missingness.* Missing data analysis was conducted for all primary variables of interest. The missing data mechanisms were evaluated to determine whether they adhered to the conditions of Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR). Complete case analysis was performed if data was MCAR. Later, multiple imputation was performed on variables determined to be Missing at Random (MAR). Final model coefficients on both data sets were compared.

*Variable Creation.* Change scores were created for viral load, CD4+ T cell count, and aggregate physical and mental QOL scores by subtracting the baseline (year 0) measurements from the year 2 measurements. This yielded scores where higher values signify an increase, and negative values signify a decrease from baseline in each outcome measure. For viral load, lower or negative values are desirable and denote improvement. For CD4+ T cell count and the aggregate physical and mental scores, larger values are desirable and denote improvement. Dummy codes were created to place subjects into hard drug use categories of never users, current hard drug users (use at year 2), and previous hard drug users (use at year 0 or 1).

*Filtering.*

**Preliminary Data Analysis and Descriptive Statistics**

While the researchers were not explicitly interested in covariates for this analysis, they provided a robust data set with many variables ostensibly linked to immune function, such as kidney disease. In the interest of providing the most accurate model, the relationship between these potential covariates were explored through scatterplots and correlation matrices to assess for potential relationships. Promising variables were included as candidates to be included in the final model.

**Data Analysis**

Four linear regressions were performed with either viral load, CD4+ T cell count, aggregate physical score, or aggregate mental score as the dependent variable. The simplest model was selected at the start to determine baseline relationships between hard drug use and the primary outcome variables. Overall F-tests were performed to assess for an overall effect in the model, with an alpha of 0.05 as the cutoff for significance. Partial F-tests were then performed to assess whether individual independent variables are significant predictors of viral load, CD4+ T cell count, and physical or mental QOL scores while accounting for the other variables in the model (p < 0.05). To mitigate the risk of Type I errors due to multiple pairwise comparisons, the Bonferroni correction was employed. Final models were selected based on AIC and BIC.

Several plots were made to ensure the data met the assumptions for a linear regression. Q-Q plots and histograms of the jackknife residuals were created to assess normality. A scatterplot of the jackknife residuals was made to ensure errors were centered around zero. Homoscedasticity was assessed by plotting the jackknife residuals for each treatment group,, along with Bartlett’s test of Homogeneity of Variances.