Project 1: Final Report

**I: Introduction**

This study utilized data from the Multicenter AIDS Cohort study, an ongoing prospective study of the histories of HIV in homosexual men from four U.S. cities. The data we received contained eight years’ worth of study data including four outcome measures and a multitude of other measured variables. We focused only on baseline and year two data. Our study was investigating the impact of the use of hard drugs (i.e. Cocaine, Heroin, etc.) on the effectiveness of Highly active antiretroviral treatment (HAART) on treating HIV/AIDS. The effectiveness of HAART was measured using four separate outcomes (two laboratory measures and two quality of life measures). The laboratory measures were viral load and CD4+ T cell count (LEU3N). The two quality of life measures were aggregate physical quality of life and aggregate mental quality of life, these were measured on a scale of 0 – 100 with higher scores equaling higher quality of life. Based on previous animal studies we hypothesized that people who were taking hard drugs at baseline would see smaller or no change in outcomes after two years. We created four new outcomes that were the difference between the outcomes at two years and baseline (i.e. viral load at year 2 – viral load at baseline). If HAART was working we should see positive differences for LEU3N, and the Quality of life scores, Viral load should have a negative difference if HAART is working.

**II: Methods**

We started the analysis by cleaning the data to make it fit for analysis, and thus ensuring that the results could be accurately interpreted. The first step in cleaning the data was to produce a set of descriptive statistics for the various variables. For the continuous variables this meant generating such statistics as the mean, minimum, maximum and standard deviation. For categorical variables we focused on the number and percent of each level of the variables. These statistics were used to check for missing data and any values that were out of the range of possibility. Any values outside the range of possibility were then set to missing. Next, we ran univariate statistics on the four outcome measures using SAS v.9.4. The univariate statistics and graphs (Histogram and normal plots) were used to check that the outcomes meet the assumptions needed to run the models. If violations of the assumptions were discovered, we discussed them with the investigator and tried to rectify them using the investigators recommendations.

Once the data was cleaned and finalized we then generated a new set of descriptive statistics, this time we divided the data between those subjects that had hard drug use and those that did not have hard drug use. These descriptive statistics were used to see if there were any major differences between the two groups of subjects. If there were major differences between the two groups those difference could mask the effect of hard drug use. In this stage of the analysis we also looked at the relationship between the independent variables. For continuous variables this was done by calculating correlation coefficients, and for categorical variables this was done by looking at two-way tables. The purpose of this step was to check for any significant relationships between the independent variables. A significant relationship between the independent variables could cause issues with the model, mainly masking the effects of other variables on the outcome.

Once we were satisfied with the distributions and relationships among the independent variables and were satisfied that the outcomes did not seriously violate any of the necessary assumptions we began checking the relationships between the outcomes and the independent variables. This was accomplished in two different ways, graphically and statistically. Graphically this was done by either creating scatter plots of the independent variable versus the outcome (for continuous variables) or creating boxplots of the outcome variable by the different levels of the independent variable (for categorical variables). In the scatter plots, if a relationship between the independent variables and the outcomes existed it would be seen as a distinct pattern in the plot, if there was no distinct pattern than most likely there was no relationship. In the boxplot, a relationship between the independent variables and the outcomes would be represented as a difference in the boxes represented the middle 75% of the data. These were than confirmed or denied by fitting crude models of the outcomes versus each independent variable. An independent variable was determined to be associated with the outcome if it had a p-value of less than 0.05. Only variables that were considered associated with the outcomes of interest were kept in the final model. Except for hard drug use, because it was the main variable of interest.

Once we determined the variables that were associated with the outcomes of interest we created four separate models, one for each outcome, that included all the associated variables. We than removed any variables that were not significant in the presence of the other, until all variables in the model were significant. The results of these final models were then interpreted in terms of the question of interest.

**III: Results**

Based on talks with the investigator we focused the attention of our analysis on the following variables; baseline outcome, baseline age, baseline BMI, race (Non-Hispanic white vs. Other), baseline marijuana use, baseline alcohol use (> 13 drinks per week vs. <= 13), baseline smoking status (current smoker vs. non/former smoker), baseline income (<$10,000, $10,000-$40,000, >$40,000), Education (>High School vs. <= High School) and Adherence level (>95% vs. <= 95%). Baseline BMI had values that exceed 100, these values were turned into missing values.

The outcome viral load had a non-normal distribution as seen in figure 1. The histogram and normal probability plots showed that the viral load outcome could not be used in its current form. After discussions with the investigator it was determined that the best solution was to use a log10 transformation on the viral load outcome. The distribution of the viral load outcome was greatly improved by the transformation (fig. 2).

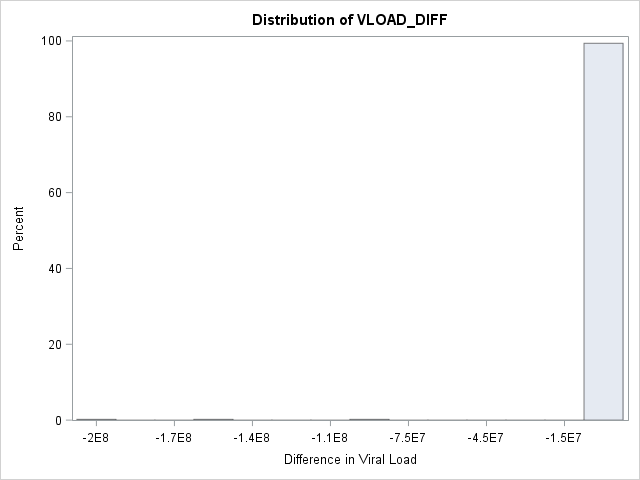


Fig.1: Histogram of frequency of Differences in Viral load

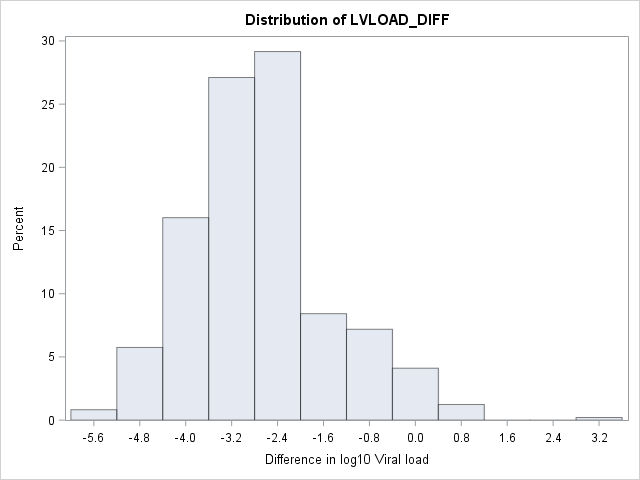


Fig 2: Histogram of frequency of differences in log10 Viral load

Table one shows the number and percent of missing data for each of the variables that has missing data. As table one shows about 30% of the quality of life measurements are missing, while 32% of the laboratory results are missing. The missing data was removed from the final data set before analysis. Since there was a difference in amount of missing data for the quality of life measures and laboratory measures we create a separate data set for each. Descriptive statistics for each dataset is shown in table 2. The number of drug users was the same between the quality of life and laboratory datasets (N=36), but the laboratory data set had fewer non-hard drug users (N=421) then the quality of life data sets (N=428). This results in a final sample size of 464 for the quality of life measurements and 457 for the laboratory measurements. The key differences between the hard drug users and the non-hard drug users are that hard drug users had a higher proportion in the low income (< $10,000) bracket (39% vs. 20%) and more hard drug users had less than a high school education compared to non-hard drug users (36% vs. 20%). We decided to control for these variables because of the large differences in them between the hard drug use group and the non-hard drug use group. Some interesting features are seen in the outcome variables between the hard drug users and non-hard drug users. Both groups saw an increase in mean aggregate mental quality of life (3.89 and 2.19), with the hard drug group see the larger increase. Aggregate physical quality of life saw the opposite, both groups saw a decrease in physical quality of life, with the hard drug group seeing a large drop (-4.80 and -1.48). We saw the right movement in the laboratory values with a decrease in viral load and increase in LEU3N counts. The hard drug group saw a much smaller increase in LEU3N counts. Table 1:

|  |  |  |
| --- | --- | --- |
| Variable | Number missing | Percent missing |
| Difference in Aggregate mental | 216 | 30.21% |
| Difference in Aggregate Physical | 216 | 30.21% |
| Difference in log10 Viral Load | 228 | 31.89% |
| Difference in LEU3N | 228 | 31.89% |
| BMI | 33 | 4.62% |
| Adherence | 209 | 29.23% |
| Income | 34 | 4.76% |

Table 2: Descriptive statistics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Hard Drug Use- QL & Lab [N = 36] | No Hard Drug Use – QL  [N = 428] | No Hard Drug Use – Lab  [N = 421] | Total – QL  [N = 464] | Total – Labs  [N = 457] |
| Mental Score Diff  [Mean (SD)] | 3.89 (15.65) | 2.19 (11.74) | NA | 2.32 (12.08) | NA |
| Physical Score Diff  [Mean (SD)] | -4.80 (8.38) | -1.48 (8.16) | NA | -1.73 (8.21) | NA |
| Log10 Viral Load Diff  [Mean (SD)] | -2.72 (1.32) | NA | -2.73 (1.22) | NA | -2.73 (1.22) |
| LEU3N count Diff  [Mean (SD)] | 11.18 (203.63) | NA | 182.59 (173.87) | NA | 169.08 (182.11) |
| Age [Mean (SD)] | 43.89 (9.52) | 43.21 (8.75) | 43.20 (8.78) | 43.26 (8.81) | 43.25 (8.83) |
| BMI [Mean (SD)] | 23.62 (3.45) | 25.38 (4.47) | 25.40 (4.49) | 25.24 (4.42) | 25.26 (4.44) |
| Race: NHW [N (%)] | 19 (52.78%) | 276 (64.49%) | 275 (65.32%) | 295 (63.58%) | 294 (64.33%) |
| Marijuana Use [N (%) – Yes] | 12 (33.33%) | 179 (41.82%) | 175 (41.57%) | 191 (41.16%) | 187 (40.92%) |
| >13 Drinks per Week [N (%)] | 2 (5.56%) | 31 (7.2%) | 32 (7.60%) | 33 (7.11%) | 34 (7.44%) |
| Current Smoker [N (%)] | 27 (75%) | 151 (35.28%) | 148 (35.15%) | 178 (38.36%) | 175 (38.29%) |
| Income <$10,000 [N (%)] | 14 (38.89%) | 84 (19.63%) | 85 (20.19%) | 98 (21.12%) | 99 (21.66%) |
| Income $10,000 - $40,000 [N (%)] | 13 (36.11%) | 183 (42.76%) | 179 (42.52%) | 196 (42.24%) | 192 (42.01%) |
| Income >$40,000 [N (%)] | 9 (25.00%) | 161 (37.62) | 157 (37.29%) | 170 (36.64%) | 166 (36.32%) |
| Education: < High School [N (%)] | 13 (36.11%) | 85 (19.86%) | 82 (19.48%) | 98 (21.12%) | 95 (20.79%) |
| Adherence: >95% [N (%)] | 35 (97.22%) | 384 (89.72%) | 377 (89.55%) | 419 (90.30%) | 412 (90.15%) |

We ran crude models to look for associations between the independent variables and the outcomes of interest. The crude models looked at the association between the outcomes and baseline outcome, age, BMI, Race, baseline marijuana use, Alcohol use, and adherence rates. Education level, income level and smoking status. The only variable that was found to have a significant association with aggerate mental quality of life scores was the score at baseline (p <0.001). Hard drug use was found to not have an association with changes in aggregate mental quality of life at year two (p = 0.4182).

Baseline physical quality of life and Hard drug use were both found to have an association with changes in physical quality of life at year 2 (p <0.001 and p=0.0194 respectfully). With all the other variables included in the model, hard drug use is still significantly associated with changes in physical quality of life scores between baseline and year two (p = 0.0243).

Hard drug use (p < 0.001), BMI (p = 0.0291), marijuana use at baseline (p = 0.0251) and medication adherence (p = 0.0412) were all significantly associated with changes in LEU3N counts in the crude models. BMI (p = 0.0583) became not significant when placed in the model with the other variables, and was subsequently dropped form the final model. With the other variables already in the model hard drug use was still significantly associated with changes in LEU3N counts between baseline and year 2 (p<0.001).

Log10 Baseline viral load (p<0.001), Race(p=0.0015), education level (p=0.0183), and medication adherence rates (p=0.0029) were the only variables found to be significantly associated with changes in log10 Viral load between baseline and year 2.

**IV: Conclusion**

The analysis revealed that hard drug use was only significantly associated with changes in physical quality of life scores and changes in LEU3N counts. Hard drug use was not significantly associated with changes in mental quality of life scores and changes in log10 viral load. The results show that on average people who are hard drug users have a difference in aggregate physical quality of life score -3.15 (95% CI: -5.90, -0.41) points lower than people who are not hard drug users. The results also show that on average people who are hard drug users have a difference in LEU3N counts -168.89 (95% CI: -230.33, -107.45) cells lower than people who are not hard drug users. The results seem strange in the fact that hard drug use appears to lower the recovery of CD4+ T cells (the LEU3N measure) but has no effect on the viral load. This will need to be explored further. This could depend on how the HAART medication works, and how quickly CD4+ T cells can replenish themselves. This points to the possibly that hard drug is affecting the immune system’s ability to recover but not effecting HAARTs ability to reduce the viral load.

**V: Reproducible Research**

https://github.com/BIOS6623-UCD/bios6623-MCuffney/tree/master/Project1/Code