BIOS 6623 Project 1 Written Report

Michaela Palumbo

October 8, 2017

Introduction

This data analysis project was done in collaboration with the Multicenter AIDS Cohort study, a prospective cohort study that follows HIV-1 positive bisexual or homosexual men over time and collects data on a number of biological, psychological, and social outcomes. The study began with collecting measurements at baseline just before the men began receiving HAART treatment, the standard of care for treating HIV infection. The dataset provided by the investigators included data for a number of different types of outcomes that had been collected at baseline and annually for the following 8 years. The goal of this analysis was to determine if treatment response at 2 years after baseline was different between men who reported hard drug use at baseline and men who did not report hard drug use. Treatment response was quantified using four variables: mental quality of life score, physcial quality of life score, viral load, and CD4+ T cell count. To answer this question we formulated the following statistical hypothesis: does hard drug use significantly effect the change from baseline to 2 years for the four outcomes of interest? In order to answer this question the data was analyzed as described below using a Bayesian approach per the investigators' request.

Methods

Upon receiving the dataset from the investigators, we began with data checking and cleaning. After communication with the investigators we narrowed down a list of variables that were of primary interest for this analysis. We then removed all data irrelevant to our analysis. We re-coded values for missing data, re-categorized levels of categorical variables to match

the investigators' requests, removed outliers, and removed subjects with missing data. To determine if there were any outliers or unrealistic values we created histograms for continuous variables and frequency tables for categorical variables. We noticed that viral load appeared to be skewed so we did a log10 transformation of viral load. The investigators informed us that log10 viral load is how that variable is clinically discussed. Therefore we decided that we would do analysis on log10 viral load and not back transform afterwards because discussing changes in log10 viral load was meaningful to clinicians and the investigators. Difference variables were then created for the 4 outcomes of interest (mental quality of life score, physical quality of life score, log10 viral load, and CD4+ T cell count). For our analyses we defined the change from baseline to 2 years as the measurement at year 2 minus the baseline measurement. The difference variables were created because we planned to use them as the outcome in the models we fit. The final, cleaned analytic dataset contained measurements on 414 subjects. All descriptive statistics and analyses were carried out using this dataset. Boxplots were created to compare the relationships between the 4 outcomes and hard drug use. Plots of the outcomes versus the covariates were also made to assess model assumptions like linearity. Boxplots and two-way tables were also used to compare demographic characteristics between subjects who reported hard drug use at baseline versus those who did not. The same was also done to compare demographic characteristics between those whose adherence at year 2 was 95% or higher versus subjects with less than 95% adherence at year 2.

Next, we outlined our model fitting strategy. Separate univariate models were fit for each of the four outcomes (change in mental quality of life score, change in physical quality of life score, change in log10 viral load, and change in CD4+ T cell count). We began by fitting two models. We fit a crude model containing only hard drug use and baseline value of the outcome as predictors. We also fit a full model that included the same predictors as the crude model plus all other covariates requested to be controlled for by the investigators. We would compare the DIC of these models and unless the DIC was drastically lower (more than 20 points) for the crude model, we would choose to interpret the full model because

the investigator was interested in also looking at associations for all the covariates they requested. We would also compare the effect size between the crude and full model to see if the interpretation was dramatically different between the models. If we saw this we planned to further investigate which additional covariates may be responsible for this change in effect. For whichever model (crude or full) was selected as the final model to make inference on for each outcome, we would then fit a reduced model that removed only the primary predictor, hard drug use. This was done so that DIC between the reduced model and the model including hard drug use could be compared in order to have a metric to assess if hard drug use was a significantly important predictor for the outcome. This approach was used to determine significance because there are no p-values in the Bayesian framework.

The Bayesian models were carried out using *PROC MCMC* in SAS. After discussing with the investigators, we decided to use uninformative priors. To have an uninformative prior for the variance parameter, sigma squared, we specified an inverse gamma distribution with a shape parameter equal to 2.001 and a scale parameter equal to 1.001. The starting value provided for the variance parameter was 1. This was the case for models for all of the outcomes. In order to have uninformative priors for the beta parameters correpsonding to the intercept and predictors in the models, the prior distributions were different depending on the outcome. For the the models with change in mental quality of life score, change in physical quality of life score, and change in log10 viral load as the outcome the prior distribution specified for the beta parameters was a normal distribution with mean 0 and variance 1000. For the model with change in CD4+ T cell count as the outcome we had to increase the variance in order to provide an uninformative prior for the beta parameters. For this outcome, the prior distribution specified for the beta parameters was a normal distribution with mean 0 and variance 100000. The starting value for beta parameters for all models was 0. In order for the models to converge the following were specified when running the models. The number of burn-ins was 1000 and the thinning was equal to 20 for all the models. For the crude models the number of simulations was equal to 300000 and for the full and reduced

models the number of simulations was equal to 350000.

Results

When comparing the desctiptive statistics between those who did not report hard drug use at baseline and those who did report drug use at baseline (appendix, table 1), we saw that there were many more subjects who did not report hard drug use at baseline (n = 382) compared to subjects who reported hard drug use (n = 32). We saw that 71.9% of hard drug users were also current smokers which is a significantly larger proportion compared to the 35.9% of non hard drug users that are current smokers ($p < 0.001$). Despite this difference in smoking status between those who do and do not use hard drugs, there was not a significant difference in marijuana or alcohol use. Of those who do not use hard drugs, 80.9% have greater than high school level education, which is significantly higher than the 59.4% of hard drug users that have greater than high school education ($p = 0.008$). The average BMI of hard drug users was 23.36 and significantly lower than the average BMI of those who did not use hard drugs which was 25.27 ($p = 0.015$). 50% of hard drug users were a race other than non-hispanic white while 35.1% of non hard drug users were a race other than non-hispanic white. However, this difference was not signifcant between the two groups. When looking at the boxplots comparing changes in the four outcomes between those who do and do not use hard drugs (appendix, figure 1) it is hard to tell if there are differences in outcomes between the two groups. The difference between the groups looks largest for the change in CD4+ T cell count. In the descriptive statistics table (appendix, table 1) we can also note that the frequentist tests showed there was a significant difference in the average change in physical quality of life score ($p = 0.012$) and CD4+T cell count ($p < 0.001$) between those who did and did not use hard drugs.

The results in the frequentist setting matches the conclusions that were drawn from our Bayesian models (appendix, table 2). Change in mental quality of life score was on average 1.1225 (95% HPD: -4.9229 to 2.5846) units lower for hard drug users compared to subjects

who did not report hard drug use. This difference is not statistically signifcant because the DIC was lower when hard drug use was removed from the model indicating that it was not an important predictor for the model. Change in physical quality of life score was on average 3.3322 (95% HPD: -6.2292 to -0.4980) units lower for hard drug users. This difference was statistically significant because removing hard drug use from the model increased DIC indicating that hard drug use was a significant predictor for this model. Change in log10 viral load was on average 0.0214 (95% HPD: -0.4054 to 0.4326) units higher for hard drug users. This difference was not statistically significant because DIC decreased when hard drug use was removed from the model, indicating that it is not an important predictor for the model. Change in CD4+ T cell count was on average 189.2 units lower (95% HPD: -251.4 to -128.6) for hard drug users. This difference between groups was statistically significant based on DIC increasing when hard drug use was removed from the model, indicating that it is a significant predictor for the model.

Conclusions

Recall the scientific question of interest was whether hard drug use affected HAART treatment response after 2 years in HIV-1 infected bisexual and homosexual men. Treatment response was quantified by two lab-based biological measures, CD4+ T cell count and log10 viral load, and two quality of life meausres, mental quality of life score and physical quality of life score. Based on the results of the analysis we can conlcude that change over 2 years in the CD4+ T cell count and the physical quality of life score is significantly lower for hard drug users. CD4+ T cell count was expected to increase as the treatment began. CD4+ T cell count worsened even more for hard drug users after 2 years of treatment, indicating that their immune systems are not recovering as they should be in response to receiving HAART treatment. Physical quality of life score declined more over the 2 years for hard drug users indicating that hard drug use may amplify the decline in physical quality of life for HIV-1 positive men receiving HAART treatment.

When interpreting the results we must remember to keep in mind the limitations of this study. Although our models adjusted for demographic characteristics, this is still an observational study so the samples of subjects who do and do not report hard drug use at baseline don't exactly match. There were a number of signficant differences between the groups. Among subjects who reported hard drug use there was a higher proportion of current smokers, which we know can also have many negative health impacts. Among hard drug users there were also lower education levels and BMIs. Also, the proportion of subjects that belong to a minority race was higher among hard drug users, although this difference was not statistically significant. Another limitation of this study is that the primary predictor, hard drug use at baseline, is a self-reported measure. It is very possible that subjects lied about their habits. Adherence is another limitation. Although this study did have a metric for adherence that we included in the models, this is another self-reported measure. It is possible that subjects were not entirely honest when reporting their adherence levels. Finally, there was a large amount of missing data in this study. We are never able to rule out that the data is missing not at random. If this is the case, the results of the analysis may be biased.

Reproducible Research Information

The code for this report can be found in the "Code" folder within the "Project1" folder in my github directory. My github directory is as follows: BIOS6623-UCD/bios6623-micpalumbo. A description of each of the files in my "Project1" folder including this report can be found in the readme.md file within the "Project1" folder. The code I used to read in the data is given below and can be found in the file "Proj1DataChecking.R" located in the code folder described above.

```
dat <- read.csv("~/Documents/CU AMC Fall 2017/BIOS6623/Proj1Data/hiv_6623_final.csv")
```

Appendix

Table 1: Summary of descriptive statistics stratified by hard drug use at baseline

	no	yes	p	test
n	382	32		
AGGMENTdiff (mean (sd))	2.39 (11.80)	3.35 (16.30)	0.670	
AGGPHYSdiff (mean (sd))	-1.41 (8.11)	-5.22 (8.79)	0.012	
log10vloaddiff (mean (sd))	-2.73 (1.21)	-2.61 (1.34)	0.612	
CD4PlusTDiff (mean (sd))	181.89 (161.70)	-15.38 (200.63)	<0.001	
age (mean (sd))	43.54 (8.62)	43.44 (9.72)	0.951	
race = other (%)	134 (35.1)	16 (50.0)	0.135	
income (%)			0.014	
< 10000	72 (18.8)	13 (40.6)		
10000-40000	160 (41.9)	10 (31.2)		
>40000	150 (39.3)	9 (28.1)		
BMI (mean (sd))	25.27 (4.27)	23.36 (3.41)	0.015	
education = more than HS (%)	309 (80.9)	19 (59.4)	0.008	
smoker = current (%)	137 (35.9)	23 (71.9)	<0.001	
alcohol_use = >13 /week (%)	30 (7.9)	2 (6.2)	1.000	
marijuana_use = yes (%)	158 (41.4)	12 (37.5)	0.811	
adherence = no (%)	42 (11.0)	1 (3.1)	0.271	

Table 2: Model results for hard drug use*

Outcome	Mean	Std.Dev	X95..HPD.Interval	DIC.full	DIC.reduced	significant
AggMentDiff	-1.1225	1.9128	(-4.9229, 2.5846)	3091.360	3089.638	no
AggPhysDiff	-3.3322	1.4651	(-6.2292, -0.4980)	2881.642	2884.766	yes
Log10VLoadDiff	0.0214	0.2137	(-0.4054, 0.4326)	1284.639	1282.741	no
Leu3NDiff	-189.2000	31.0734	(-251.4, -128.6)	5408.589	5443.617	yes

*Controlling for the following covariates: race, baseline age, baseline smoking status, baseline alcohol use, baseline marijuana use, baseline BMI, baseline education level, baseline income, baseline outcome value, adherence

Figure 1: Treatment response outcomes by reported hard drug use