Project 1 Report

Analyst: Max McGrath

Investigator: Nichole Carlson

Report generated: 10/11/2021

Introduction

We received data of 3935 observations for 715 patients with HIV who began receiving highly active antiretroviral treatment (HAART). The data includes baseline values for each individual prior to starting treatment, then up to 8 years of longitudinal measurements for the patients as they received HAART. Each observation includes laboratory and quality of life measures, as well as demographic and other health information. Most relevant to this analysis, each observation indicates whether patients used illicit drugs via injection or opiates via any method of administration, which will be referred to as "hard drug use" or just "drug use" throughout this analysis.

The goal of this analysis is to identify whether HAART deferentially impacts patients who report using drugs prior to receiving treatment relative to those who did not report using drugs at baseline. The efficacy of HAART will be evaluated using four treatment outcomes: (1) viral load, (2) CD4+ T cell counts, (3) aggregate physical quality of life score, and (4) aggregate mental quality of life score. We will test the hypothesis that patients who report using illicit drugs prior to beginning treatment have worse health outcomes after two years of treatment than those who did not report using hard drugs at baseline.

Methods

Patients were only included in this analysis if they had no missing values for any of the four treatment outcomes, the indicator of whether they used hard drugs at baseline, or any of the baseline covariates described below. Values that were deemed implausible were also removed. Viral load was also log base 10 transformed to account for the long-tailed distribution of those values.

Data was first summarized for relevant covariates and outcomes at both baseline and after two years of treatment. Treatment outcomes were then analyzed using linear models in both a frequentist context (multivariable least squares regression) and using Bayesian linear regression. For each framework four models per outcome were used (a total of sixteen models

per framework), where differences in each outcome (year two values minus baseline) are used as the response for the respective models.

Hard drug use at baseline is the primary variable of interest, so each model was defined to better understand the relationship between hard drug use and each respective outcome. First, models with only hard drug use and baseline outcome values were estimated for each outcome. We will refer to those models as univariable models. Next, confounding variables were added to those models to control for factors that are likely correlated with both drug use and the outcomes. Those confounding variables include age, body mass index, smoking status, education, race/ethnicity. All confounding values were those measured at baseline, except adherence which used second year values. These models will be refered to as multivariable models. Lastly, for each outcome, the univariable and multivariable models were run without hard drug use as a predictor to quantify the impact of removing hard drug use as a predictor. We will refer to these models as univariable without drug use and multivariable without drug use, respectively.

For the Bayesian model, all priors were non-informative to reflect a lack of *a priori* knowledge surrounding the associations between predictors and outcomes. Specifically, coefficient priors were Gaussian with a mean of 0 and a variance of 100,000 and error priors were gamma distributions with shape and rate parameters both equal to 0.001. For each Bayesian model, two Monte Carlo Markov Chains were run for 25,000 iterations of Gibbs sampling to draw independent values from each coefficient's posterior density to estimate those densities. The convergence and mixing of the chains was evaluated using trace plots, autocorrelation function plots, and by visually examining plots of the posterior density estimates, all to ensure that the chains accurately reflect the underlying parameters' distributions. After the chains were examined, each respective pair was combined.

For the Bayesian linear models, parameter estimates, 95% highest posterior density intervals (HPD), penalized deviance information criteria (DIC), and posterior probabilities of coefficients corresponding to a 10% or greater difference from the sample mean outcome values

were all calculated and evaluated to evaluate the impact of hard drug use upon HAART. Those statistics' frequentist counterparts, least-squares coefficient estimates, confidence intervals, and t-test p-values were also used to evaluate the same relationship. Concordance or lack thereof between the two paradigms as well as any limitations to this analysis are also discussed.

Results

Following the removal of incomplete observations, a total sample size of 463 patients remain, 36 who used hard drugs at baseline and 427 who did not. After running the Markov Chains as described above and examining trace plots, autocorrelation function plots, and estimated posterior distribution plots, it is clear that the chains are an accurate estimate of the posterior densities given the observed data and selected priors. Summaries of the data used in this analysis are found in Table 1, and a missingness summary for relevant baseline and year two values is found in Table 2. Box plots for outcomes for those who reported using drugs and those who did not are provided in Figure 1.

Below we describe the results of the Bayesian modeling and contrast those results with those of the frequentist modeling, with full results available in the Appendix in Tables 3 and 4, respectively.

Viral Load

For the univariate model, a coefficient estimate for drug use (the mean of the posterior density) of 0.038 (95% HPD: -0.355, 0.429) was calculated. Since viral load was log base 10 transformed, we back-transform this estimate to arrive at the value 1.091, implying that drug use is, on average, associated with 1.091 times higher difference in HIV copies in a mL of blood after two years of HAART, relative to non-drug users. For the univariable model, the DIC was estimated to be 1438.458, while the same model without drug use had a DIC of 1437.448. The posterior probability of the drug use coefficient corresponding with a 10% change of the sample mean of viral load was calculated to be 0.828.

In both the univariate and multivariable models, HPDs included 0, the DIC changed very little when drug use was removed from the model, and posterior probabilities were insufficient to indicate significance, all indicating the data does not support an association between drug use and change in viral load after two years of HAART. These results are further supported by the frequentist modeling.

CD4+ T Cell Count

For the univariate model with CD4+ T cell count as the outcome, the drug use coefficient for drug use was estimated to be -170.989 (95% HPD: -232.818, -112.506), suggesting that on average drug users have -170.989 lower difference in CD4+ T cells than non-drug users after 2 years of HAART. The DIC for the univariate model was 28.993 lower than the model without drug use as a variable, indicating the inclusion of drug use notably improved the fit of the model. The posterior probability for the CD4+ coefficient indicating a 10% change in CD4+ counts was approximately 1.

Those results remained consistent for the multivariable model, both with and without drug use as a predictor, indicating that the results were robust to the inclusion of potential confounders. The results are also consistent with the frequentist analysis.

Mental Quality of Life Score

For the univariate model with mental quality of life score as the outcome, the coefficient for drug use was estimated to be -0.136, but the 95% HPD (-3.526, 3.250) is practically centered at 0. This, coupled with the only slight decrease in DIC when including drug use as a coefficient, indicate that the data does not support an association between drug use and differing changes in mental quality of life score between drug users and non-users after two years of HAART.

This result is consistent with both the multivariable model and the frequentist analysis.

Physical Quality of Life Score

For the univariate model with physical quality of life score as the outcome, the coefficient for drug use was estimated to be -4.074 (95% HPD: -6.691, -1.312) with a posterior probability of drug use corresponding with an average 10% change in physical quality of life score of 0.9988. The DIC with the inclusion of drug use as a variable is slightly lower than that of the model without it.

These results are somewhat consistent in the multivariable model, although the HPD moves to include 0 (-5.433, 0.050) and the posterior probability of a 10% change drops to 0.985, both indicating some level of confounding between the additional variables and drug uses relationship with physical quality of life.

The results are consistent with those from the frequentist analysis, with similar values for coefficient estimates, AIC, and confidence intervals as their analogous Bayesian statistics. In both cases, after controlling for potential confounders, there remains weak evidence that drug use and change in physical quality of life score after 2 years of HAART are negatively associated.

Conclusions

From the above results, we see strong evidence that differences in CD4+ cell counts after two years of HAART are smaller in individuals that use hard drugs, implying there may be a negative effect of hard drug use on HAART. We also see weak evidence that drug users see reduced changes in physical quality of life score after 2 years of HAART relative to non-drug users. We do not, however, note any significant difference in viral load or mental quality of life score between individuals who use hard drugs and those who do not. As such, we do conclude that drug use is associated with worse health outcomes after two years of HAART.

Limitations and Discussion

The greatest limitation of this study is the small sample of drug users relative to non-users. When considering this difference it must be noted that if it is the case that hard drug use negatively impacts the efficacy of HAART, there are several factors which would likely bias the above results. First, the use of hard drugs was self-reported by individuals being treated. Given the gravity of the social stigma surrounding drug usage (particularly the drugs being studied), it is foreseeable that some users would be disinclined to report their drug use, leading to the presence of drug users in the non-drug using study arm, thereby diminishing the measured effect. Further, it would be foreseeable that dropout rate among drug users would be higher than among non-users due to both diminished health (and death) and to socioeconomic factors, which would both bias our results, as those with the worst health out comes may drop out, and reduce the sample size of drug users in the study.

Given these factors, the above findings should be considered with greater leniency, where even weak effects should be considered indicative that a meaningful correlation may exist. As such, our results are indicative of a meaningful correlation between drug use and worse health outcomes are 2 years of HAART. Further research that aims to expand the quantity of drug using patients should be pursued. This could be accomplished through more active recruiting and provision of HAART treatment to drug using individuals and through refinement of data collection methods to foster a more welcoming environment in which individuals may feel more inclined to self-report their drug use.

Reproducibility

The code used to generate this analysis is available on GitHub at https://github.com/BIOS6624-UCD/bios6624-MaxMcGrath/tree/main/Project1 . The `Background` folder contains information pertinent to understanding the analysis but unnecessary for reproducing it. The `Code` folder contains six files: `1_prepData.R`, `2_EDA.R`, `3_FrequentistAnalysis.R`, `4_BA_DrawChains.R`, `5_BA_AssessChains.R` and `6_BA_AnalyzeChains.R`. These R scripts are dependent upon a data file `Data/hiv_6624.csv` which is not available on GitHub, but may be requested by emailing `max.mcgrath@ucdenver.edu`. To run the complete analysis, each script should be run in the order of the number preceding its filename. The last directory, `Report`, contains the RMarkdown file `report.Rmd` which may be used to generate this report (note that it also depends on the aforementioned data and scripts).

The complete details of the `R` version, package versions, and machine details for the instance which generated this report are provided below.

```
## R version 4.0.4 (2021-02-15)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Catalina 10.15.7
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
```

```
## other attached packages:
##  [1] ggplot2_3.3.5    naniar_0.6.1     mcmcse_1.5-0     rjags_4-10
##  [5] coda_0.19-4      readr_2.0.2      table1_1.4.2     magrittr_2.0.1
##  [9] dplyr_1.0.7      tidyr_1.1.4      kableExtra_1.3.4
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.1.1  xfun_0.22         purrr_0.3.4       lattice_0.20-41
##  [5] colorspace_2.0-2  vctrs_0.3.8       generics_0.1.0    testthat_3.0.2
##  [9] htmltools_0.5.1.1 viridisLite_0.4.0 yaml_2.2.1        utf8_1.2.2
## [13] rlang_0.4.11      pillar_1.6.3      withr_2.4.2       glue_1.4.2
## [17] DBI_1.1.1         lifecycle_1.0.1   stringr_1.4.0     gtable_0.3.0
## [21] munsell_0.5.0     rvest_1.0.0       evaluate_0.14     knitr_1.31
## [25] tzdb_0.1.2        fansi_0.5.0       Rcpp_1.0.7        scales_1.1.1
## [29] webshot_0.5.2     systemfonts_1.0.1 ellipse_0.4.2     hms_1.1.1
## [33] digest_0.6.28     stringi_1.7.4     visdat_0.5.3      grid_4.0.4
## [37] tools_4.0.4       tibble_3.1.4      Formula_1.2-4     crayon_1.4.1
## [41] pkgconfig_2.0.3   ellipsis_0.3.2    xml2_1.3.2        assertthat_0.2.1
## [45] rmarkdown_2.7     svglite_2.0.0     httr_1.4.2        rstudioapi_0.13
## [49] R6_2.5.1          fftwtools_0.9-11  compiler_4.0.4
```

Appendix

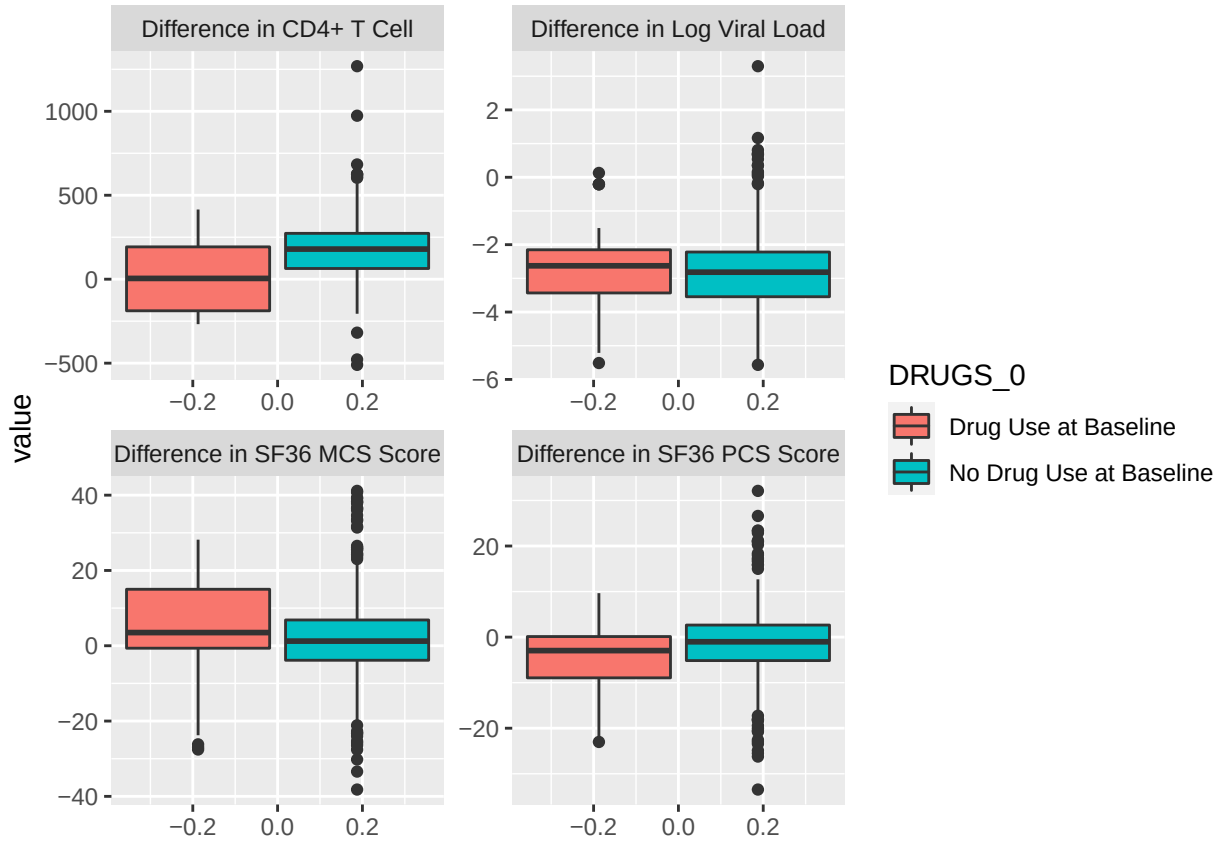

Figure 1: Difference in outcomes for those who reported using drugs at baseline at those who did not.

Table 1: Data Summary

	No Hard Drug Use at Baseline (N=427)	Hard Drug Use at Baseline (N=36)	Overall (N=463)
Change in standardized viral load (log10 copies/ml)			
Mean (SD)	-2.74 (1.22)	-2.72 (1.32)	-2.74 (1.23)
Median [Min, Max]	-2.82 [-5.57, 3.30]	-2.63 [-5.51, 0.128]	-2.80 [-5.57, 3.30]
Change in # of CD4+ cells			
Mean (SD)	182 (175)	11.2 (204)	168 (183)
Median [Min, Max]	179 [-510, 1270]	4.71 [-268, 415]	172 [-510, 1270]
Change in SF36 MCS score			
Mean (SD)	2.42 (11.8)	3.89 (15.7)	2.53 (12.1)
Median [Min, Max]	1.23 [-38.2, 41.1]	3.50 [-27.6, 28.2]	1.44 [-38.2, 41.1]
Change in SF36 PCS score			
Mean (SD)	-1.34 (8.17)	-4.80 (8.38)	-1.61 (8.23)
Median [Min, Max]	-1.02 [-33.5, 32.1]	-2.96 [-23.0, 9.65]	-1.34 [-33.5, 32.1]
Age at baseline (years)			
Mean (SD)	43.1 (8.68)	43.9 (9.52)	43.2 (8.74)
Median [Min, Max]	43.0 [20.0, 73.0]	45.5 [29.0, 61.0]	43.0 [20.0, 73.0]
BMI at baseline			
Mean (SD)	25.3 (4.39)	23.6 (3.45)	25.2 (4.34)
Median [Min, Max]	24.8 [16.5, 45.3]	23.3 [18.0, 31.2]	24.6 [16.5, 45.3]
Race/Ethnicity at baseline			
Black	108 (25.3%)	14 (38.9%)	122 (26.3%)
Hispanic	37 (8.67%)	3 (8.33%)	40 (8.64%)
Other	5 (1.17%)	0 (0%)	5 (1.08%)
White	277 (64.9%)	19 (52.8%)	296 (63.9%)
Education at baseline			
Greater than HS	191 (44.7%)	10 (27.8%)	201 (43.4%)
HS	213 (49.9%)	16 (44.4%)	229 (49.5%)
Less than HS	23 (5.39%)	10 (27.8%)	33 (7.13%)
Smoking status at baseline			
Active Smoker	149 (34.9%)	27 (75.0%)	176 (38.0%)
Non-Smoker	278 (65.1%)	9 (25.0%)	287 (62.0%)
Adherence to medication at 2-years			
Greater than 95% Adherent	382 (89.5%)	35 (97.2%)	417 (90.1%)
Less than 95% Adherent	45 (10.5%)	1 (2.78%)	46 (9.94%)
Baseline standardized viral load (log10 copies/ml)			
Mean (SD)	4.53 (0.930)	4.58 (0.871)	4.53 (0.925)
Median [Min, Max]	4.52 [0.237, 8.28]	4.47 [2.87, 6.40]	4.51 [0.237, 8.28]
Baseline # of CD4+ cells			
Mean (SD)	374 (202)	360 (201)	373 (202)
Median [Min, Max]	360 [10.9, 1220]	397 [10.9, 650]	360 [10.9, 1220]
Baseline SF36 MCS score			
Mean (SD)	45.2 (13.8)	42.0 (11.6)	44.9 (13.6)
Median [Min, Max]	49.5 [7.23, 66.0]	44.4 [22.5, 59.6]	49.1 [7.23, 66.0]
Baseline SF36 PCS score			
Mean (SD)	51.4 (9.00)	49.2 (7.05)	51.2 (8.88)
Median [Min, Max]	53.7 [19.2, 69.0]	47.4 [31.4, 62.9]	53.5 [19.2, 69.0]

Table 2: Missingess Summary

	Baseline		Year Two	
Variable	N Missing	% Missing	N Missing	% Missing
BMI	44	6.15		
Log Viral Load	42	5.87	19	3.75
CD4+ Cell Count	24	3.36	19	3.75
SF36 MCS Score	2	0.28	6	1.19
SF36 PCS Score	2	0.28	6	1.19
Hard Drug Use	0	0.00		
Educational Attainment	0	0.00		
Age	0	0.00		
Race/ethnicity	0	0.00		
Smoking Status	0	0.00		
Adherence			0	0.00

[a] Note that this table does not account for dropout. 209 patients were present in year one that were not present in year two

Table 3: Bayesian Analysis Summary

		Drug Use Coefficient			
Model	Penalized DIC	Estimate	HPD Lower	HPD Upper	Posterior Probability
Standardized Viral Load (log10 copies/ml)					
Univariable w/o Drug Use	1437.448				
Univariable	1438.458	0.0376	-0.3553	0.4294	0.8282
Multivariable w/o Drug Use	1410.525				
Multivariable	1410.898	-0.1656	-0.5657	0.2267	0.8753
CD4+ cells					
Univariable w/o Drug Use	6139.357				
Univariable	6110.364	-170.9887	-232.8183	-112.5057	1.0000
Multivariable w/o Drug Use	6137.161				
Multivariable	6112.035	-166.6399	-230.3931	-103.0180	1.0000
SF36 MCS score					
Univariable w/o Drug Use	3438.930				
Univariable	3439.971	-0.1355	-3.5264	3.2499	0.8830
Multivariable w/o Drug Use	3431.338				
Multivariable	3432.072	-1.0141	-4.5029	2.5527	0.9061
SF36 PCS score					
Univariable w/o Drug Use	3226.686				
Univariable	3218.785	-4.0736	-6.6910	-1.3116	0.9988
Multivariable w/o Drug Use	3205.014				
Multivariable	3202.249	-2.7213	-5.4334	0.0500	0.9845

Table 4: Frequentist Analysis Summary

		Drug Use Coefficient			
Model	AIC	Estimate	Lower CI	Upper CI	p-value
Standardized Viral Load (log10 copies/ml)					
Univariable w/o Drug Use	1440.457				
Univariable	1442.421	0.0379	-0.3519	0.4276	0.8487
Multivariable w/o Drug Use	1422.373				
Multivariable	1423.689	-0.1660	-0.5655	0.2336	0.4147
CD4+ cells					
Univariable w/o Drug Use	6142.366				
Univariable	6114.327	-171.0260	-231.5520	-110.5000	0.0000
Multivariable w/o Drug Use	6149.026				
Multivariable	6124.843	-166.9054	-230.9396	-102.8712	0.0000
SF36 MCS score					
Univariable w/o Drug Use	3441.939				
Univariable	3443.933	-0.1297	-3.5203	3.2609	0.9401
Multivariable w/o Drug Use	3443.186				
Multivariable	3444.863	-1.0130	-4.5608	2.5349	0.5750
SF36 PCS score					
Univariable w/o Drug Use	3229.695				
Univariable	3222.747	-4.0723	-6.7432	-1.4014	0.0029
Multivariable w/o Drug Use	3216.862				
Multivariable	3215.040	-2.7210	-5.4864	0.0445	0.0538