Project 1 Write-Up

Brenton Graham
10/10/2022

Introduction

An ongoing prospective cohort study (the Multicenter AIDS Cohort Study) has collected up to eight years of longitudinal data measuring the health of HIV-1-positive subjects who have been treated with highly active antiretroviral treatment (HAART). While HAART is a standard treatment for individuals infected with HIV-1, the efficacy of the treatment is uncertain for individuals who use hard drugs, such as heroin and cocaine. The purpose of this secondary study is to assess if treatment responses differ between HIV-1-positive subjects who do and do not report hard drug use at baseline. Primary treatment response outcomes of interest in this study include viral load, CD4+ T cell count, aggregate physical quality of life score, and aggregate mental quality of life score. This secondary study focuses on treatment response measurements at baseline and two years, with respect to treatment initiation. More aggressive treatment options ought to be explored for hard drug users if health response disparities are detected. Additionally, alternative treatments should be explored if treatments are unexpectedly associated with worsened quality of life scores. The primary hypotheses that will be tested in this report include the following.

- *Hypothesis 1*: Subjects who report hard-drug use at baseline are associated with **increased** viral load at two years, as compared to subjects who do not report hard-drug use at baseline.
- *Hypothesis 2*: Subjects who report hard-drug use at baseline are associated with **decreased** CD4+ T cell counts at two years, as compared to subjects who do not report hard-drug use at baseline.
- *Hypothesis 3*: Subjects who report hard-drug use at baseline are associated with **decreased** aggregate physical quality of life scores at two years, as compared to subjects who do not report hard-drug use at baseline.
- *Hypothesis 4*: Subjects who report hard-drug use at baseline are associated with **decreased** aggregate mental quality of life scores at two years, as compared to subjects who do not report hard-drug use at baseline.

Investigators have requested that each of the stated hypotheses be statistically tested with both a Frequentist and Bayesian framework in this report. While the two approaches are expected to yield similar conclusions, estimate and interval comparisons should provide a thorough basis for inference.

Methods

Exploratory Data Analysis

Summary statistics and data visualizations were used to explore the relationship between hard-drug use and each of the treatment response outcomes, as well as check for the presence of outliers. Outcome distributions for hard-drug users and non-users are shown in Figure 1. Plots in Figure 1 include data from subjects that were included in the complete-case, transformed and filtered data set that was used for modeling. Methods for data transformation are described in the subsequent sections.

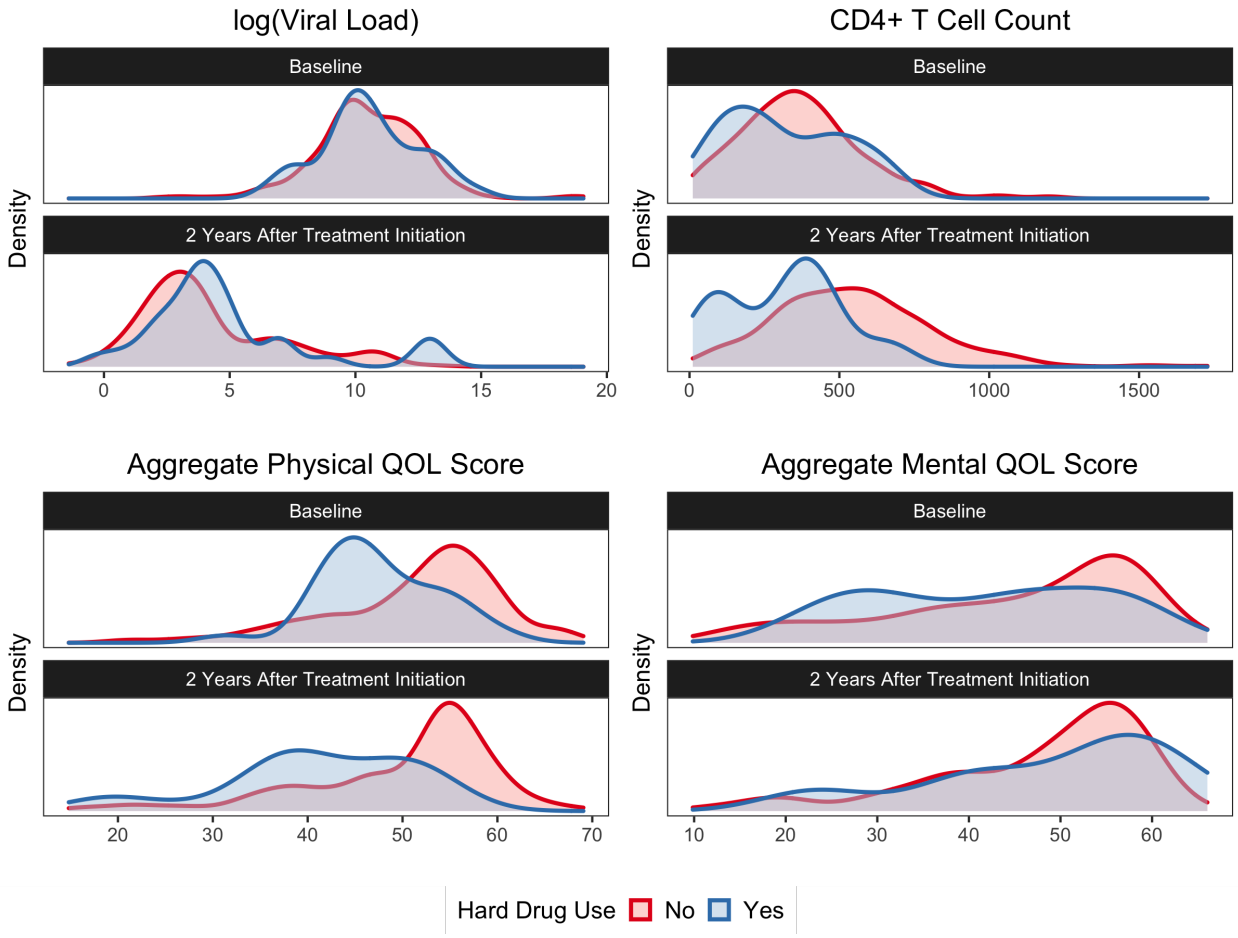

Figure 1: Treatment response outcome distributions at baseline and two years for hard-drug users and non-users.

Data Transformation

Prior to analysis and model fitting, the raw data set was transformed and filtered to best test the hypotheses of interest. Categorical data were converted from encoded values to informative grouping labels. The data were then filtered to only include observations from the baseline and year two time points. Observations were further filtered to only include subjects who had data at both time points. Next, the data set was transformed for compatibility with a baseline-as-covariate modeling structure. The transformation included

the following steps. First, the data set was converted from long to wide format with respect to the two time points. Second, year two data were dropped for the primary variable (hard-drug status) and potential covariates (ie. only baseline measurements included), with the exception of the medicine adherence covariate. Year two data were used for this covariate since adherence measurements were only collected after baseline. The rationale behind using baseline covariates (instead of year two covariates) for analysis is that we are interested in comparing subjects who do and do not report hard-drug use *at baseline*. Finally, the viral load outcome was log-transformed at each time point to get the distributions closer to Gaussian.

Missing Data

Missing data were visualized and assessed following data transformation and filtering. A hypothesis test (Little's missing completely at random (MCAR) test) was implemented to determine the nature of the missing data. Due to time constraints, only complete-case observations were used for analysis. Limitations to this approach will be discussed further in the limitations section of this report.

Variable Selection

The raw data set contains an extensive set of potential covariates. The following variable selection process was implemented to minimize issues with generalizability and multicollinearity when modeling our treatment response outcomes. Removing sources of multicollinearity should help ensure convergence in our Bayesian regression models. First, variables were assessed for missingness. Variables with more than 20% of observations missing were dropped. Second, variables that were found to have the same value for each observation were dropped since these variables essentially contain no information with respect to the study population (eg. all subjects are HIV-1 positive). Third, variables that were used to derive the primary variable (hard drug use) were dropped since the information from these variables are captured within the primary variable. Fourth, multicollinearity was assessed using a variance inflation factor (VIF) approach. A recursive process was implemented to remove variables with the largest VIF values until all variables had VIF values less than 5. Fifth, and finally, hash/marijuana use variables were removed due to observed inconsistencies (eg. 24 of 160 subjects who selected "No" for hash/marijuana use in the past year selected "Daily" for use frequency).

Statistical Modeling

Investigators requested that each treatment response outcome be modeled using both a Frequentist and Bayesian framework. As such, Frequentist (multiple) linear regression and Bayesian linear regression models were fit for each treatment response. A baseline-as-covariate approach was implemented for each model. That is, two year outcomes were modeled using baseline outcome measurements as a covariate to adjust for baseline differences. Hard drug use was used as the primary explanatory variable in all models. Covariates and precision variables included all independent variables that were not dropped during the variable selection process. Point estimates, 95% confidence intervals and p-values were used for Frequentist inference while point estimates and 95% highest posterior density (HPD) intervals were used for Bayesian inference. A significance level of $\alpha = 0.05$ was used for testing statistical hypotheses in the Frequentist framework. Bayesian regression parameters and credible intervals were estimated using a Markov Chain Monte Carlo implementation in R with 2 chains and 5,000 iterations (rjags). Vague prior distributions were specified for each parameter in our Bayesian regression models. $Normal(\mu = 0, \sigma^2 = 1000)$ prior distributions were used for all covariates while an $InverseGamma(\alpha = 0.01, \beta = 0.01)$ prior distribution was used for the error term. Diagnostic plots, including Gelman-Rubin-Brooks plots, were used to assess MCMC outputs and visually inspect for convergence.

Results

Table 1 describes the study population that remained following data transformation, filtering, and complete-case selection. 98 of 506 subjects who had been seen at baseline and year two were removed due to missing data in either one of the outcomes or one of the independent variables.

Table 1: Demographics for study population by hard-drug use

	Hard-Drug Non-Users	Hard-Drug Users
n	378	30
Age (median [range])	43.00 [20.00, 73.00]	45.50 [29.00, 61.00]
BMI (median [range])	24.71 [16.50, 45.28]	22.33 [18.00, 30.44]
HighBloodPressure = Yes (%)	102 (27.0)	2 (6.7)
Income (%)		
<$10K	74 (19.6)	14 (46.7)
$10-20K	38 (10.1)	3 (10.0)
$20-30K	74 (19.6)	3 (10.0)
$30-40K	53 (14.0)	4 (13.3)
$40-50K	43 (11.4)	0 (0.0)
$50-60K	62 (16.4)	6 (20.0)
$60K+	34 (9.0)	0 (0.0)
FrailtyRelatedPhenotype = Yes (%)	15 (4.0)	0 (0.0)
Depressed = Yes (%)	134 (35.4)	17 (56.7)
SmokingStatus (%)		
Never	122 (32.3)	0 (0.0)
Former	122 (32.3)	6 (20.0)
Current	134 (35.4)	24 (80.0)
DrinksPerWeek (%)		
None	87 (23.0)	8 (26.7)
1-3	178 (47.1)	15 (50.0)
4-13	86 (22.8)	6 (20.0)
14+	27 (7.1)	1 (3.3)
Race (%)		
White, non-Hisp	246 (65.1)	13 (43.3)
White, Hisp	13 (3.4)	0 (0.0)
Black, non-Hisp	99 (26.2)	14 (46.7)
Black, Hispanic	2 (0.5)	0 (0.0)
Other, non-Hisp	3 (0.8)	0 (0.0)
Other, Hisp	15 (4.0)	3 (10.0)
TreatmentAdherence (%)		
100%	156 (41.3)	15 (50.0)
95-99%	183 (48.4)	14 (46.7)
75-94%	30 (7.9)	1 (3.3)
<75%	9 (2.4)	0 (0.0)

As shown in Table 1, the study population isn't well-balanced in terms of some key variables, including hard-drug use, income, smoking status, and race. These imbalances might introduce some bias in our model estimates that should be considered when drawing inferences.

Frequentist and Bayesian regression models were fit for each treatment response outcome at 2 years, adjusting for baseline measurements. Results from all regression models (both Frequentist and Bayesian) are shown in Table 2. Estimates reported in Table 2 represent the measured effect of hard-drug use on two-year treatment response as compared to no hard-drug use. Positive β-coefficient estimates indicate that the corresponding treatment response had a larger increase in hard-drug users, as compared to non-users, over the two-year treatment period. Negative β-coefficient values indicate that the corresponding treatment response had a larger decrease in hard-drug users, as compared to non-users, over the two-year treatment period. Insignificant associations are inferred by 95% HPD intervals that contain zero and p-values that are less than the specified significance level ($\alpha = 0.05$), for the Bayesian and Frequentist results respectively.

Overall, there is a general agreement between the Bayesian and Frequentist results in terms of inference. Hard-drug use is associated with a significant decrease in CD4+ T cell count and Aggregate Physical Quality of Life Score over the first two years of treatment. According to the Bayesian and Frequentist models, respectively, CD4+ T cell counts were 81.31 units (95% HPDI: (-127.46, -36.50)) and 191.45 units (95% CI: (-259.88, -123.02), $p < 0.001$) lower for hard-drug users, as compared to non-users, on average. Aggregate Physical Quality of Life Scores were 4.79 units (95% HPDI: (-7.85, -1.59)) and 4.84 units (95% CI: (-7.95, -1.73), $p = 0.002$) lower for hard-drug users, as compared to non-users, on average. Surprisingly, hard-drug use is associated with a significant increase in Aggregate Mental Quality of Life Score over the first two years of treatment (95% HPDI: (0.08, 7.97); 95% CI: (0.12, 8.00); $p = 0.044$). No significant association was found between hard-drug use and viral load (95% HPDI: (-0.65, 1.39); 95% CI: (-0.68, 1.36), $p = 0.509$).

While the two frameworks yielded similar inference, Bayesian regression estimates were found to be more conservative (or less extreme), which is expected. For example, CD4+ T cell count estimates for hard-drug users were notably different between the approaches, whereby the Frequentist estimate was much more extreme.

Table 2: Effect of hard drug use on 2-year treatment responses in Bayesian and Non-Bayesian regression models

	Bayesian Framework		Frequentist Framework		
Treatment Response	Estimate	95% HPDI	Estimate	95% CI	p-value
log(Viral Load)	0.34	(-0.65, 1.39)	0.34	(-0.68, 1.36)	0.5093
CD4+ T Cell Count	-81.31	(-127.46, -36.50)	-191.45	(-259.88, -123.02)	<0.001**
Physical QOL Score	-4.79	(-7.85, -1.59)	-4.84	(-7.95, -1.73)	0.002**
Mental QOL Score	4.04	(0.08, 7.97)	4.06	(0.12, 8.00)	0.0435*

Conclusions

Statistical evidence from both Frequentist and Bayesian regression methods suggest the following.

1. Hard-drug use is not associated with viral load over the first two years of treatment.

2. Hard-drug use is significantly associated with decreased CD4+ T cell counts over the first two years of treatment.

3. Hard-drug use is significantly associated with decreased Aggregate Physical Quality of Life Scores over the first two years of treatment.

4. Hard-drug use is significantly associated with increased Aggregate Mental Quality of Life Scores over the first two years of treatment.

Despite statistical evidence, limitations in this study are plentiful and should be thoroughly considered before actions are taken. First, a complete-case analysis was used to analyze this data, removing data from

98 subjects. There is reason to believe that the missing data are not MCAR, which means complete-case analysis is likely an inappropriate method for missing data handling. Missing data ought to be imputed and a sensitivity analysis should be carried out in future work to test for the biases that may be present in these results. Second, the imbalance of our study population likely introduces bias into our results. The imbalance that exists makes it difficult to conclude whether or not an effect is due to hard-drug use or one of the independent variables. Third, and similarly, a more thorough investigation of confounding should be implemented to improve our ability to confidently draw causal conclusions. Fourth, and finally, the study population is quite narrow in demographics. All subjects non-HIV-1 positive men which likely diminishes the generalizability of the results.