Project1

Joe Nagel

Table 1: Descriptive Statistics of Baseline Covariates and Year 2 Adherence

Characteristic	No Hard Drug Use N = 437	Hard Drug Use N = 39	Overall N = 476
SF-36 MCS, mean (SD)	45.1 (13.7)	42.3 (11.2)	44.9 (13.5)
SF-36 PCS, mean (SD)	51.3 (9.1)	47.7 (8.5)	51.0 (9.1)
CD4+ T Cell Count, mean (SD)	375.4 (201.1)	352.2 (194.7)	373.5 (200.5)
Log Viral Load, mean (SD)	4.5 (0.9)	4.5 (0.9)	4.5 (0.9)
Age (Years), mean (SD)	43.1 (8.7)	44.6 (9.5)	43.3 (8.7)
BMI, mean (SD)	25.3 (4.4)	23.6 (3.4)	25.2 (4.3)
Missing/Improbable, n (%)	10 (2.3%)	3 (7.7%)	13 (2.7%)
Adherence, n (%)			
>=95%	388 (89%)	38 (97%)	426 (89%)
<95%	49 (11%)	1 (2.6%)	50 (11%)
Smoking Status, n (%)			
Never/Former Smoker	282 (65%)	9 (23%)	291 (61%)
Current Smoker	155 (35%)	30 (77%)	185 (39%)
Education, n (%)			
No College Degree	242 (55%)	29 (74%)	271 (57%)
College Degree or Greater	195 (45%)	10 (26%)	205 (43%)
Race/Ethnicity, n (%)			
Non-Hispanic White	279 (64%)	19 (49%)	298 (63%)
Other	158 (36%)	20 (51%)	178 (37%)

Table 2: Frequentist and Bayesian Model Results for Hard Drug Use Effect

	Frequentist			**Bayesian**			
Outcome	Estimate	95% CI	p-value	Estimate	95% HDI	$P(\lvert\beta\rvert >$ Threshold)	Δ LOO-IC
SF-36 MCS	-0.29	[-3.75, 3.17]	0.870	-0.30	[-3.57, 3.32]	0.272	-3.51
SF-36 PCS	-3.34	[-6.06, -0.61]	0.017	-3.33	[-5.99, -0.66]	0.834	3.19
CD4+ T Cell Count	-163.88	[-226.35, -101.41]	<0.001	-148.99	[-206.12, -85.85]	0.999	24.31
Log Viral Load	-0.06	[-0.46, 0.34]	0.770	-0.05	[-0.44, 0.33]	0.018	-1.94

Introduction

HIV (human immunodeficiency virus) is a virus that attacks the immune system; it destroys T-cells and inhibits your body's ability to fight infection. This analysis evaluates the effectiveness of highly active antiretroviral treatment (HAART), the standard treatment for HIV patients. The data come from the Multicenter AIDS Cohort Study, an ongoing study investigating HIV infection in homosexual and bisexual men in the United States. The dataset includes longitudinal demographic information, laboratory measurements, and 36-Item Short Form Health Survey (SF-36) quality of life summaries on men infected with HIV. Data was collected on subjects at the beginning of HAART treatment (baseline) and every year thereafter for up to 8 years. The primary goal of this analysis is to characterize how treatment response after 2 years of treatment differs based on a subjects hard drug use status at baseline. Specifically, we will investigate the difference in SF-36 Mental Component Summary (MCS) score, SF-36 Physical Component Summary (PCS) score, HIV viral load, and CD4+ T cell count between baseline and two years across hard drug use status at baseline while adjusting for baseline outcomes, age, bmi, smoking status, education, and race. Additionally, we will investigate if any significant differences across drug use are explained by study protocol adherence.

Preliminary Methods

We will begin by checking the analysis variables for missingness and outliers; we will specifically check if missingness differs between hard drug users and non-hard drug uses. We will calculate the change in outcomes as the value at Year 2 minus the value at Year 0. We will also visualize the distributions of the four outcome difference scores using histograms and boxplots to identify potential outliers and assess the need for any transformations prior to analysis. For the primary analysis, we will generate descriptive statistics to compare baseline demographic and clinical characteristics between those who reported hard drug use and those who did not. To evaluate the effect of baseline hard drug use on treatment response, we will fit four separate multivariable linear regression models—one for each outcome (change in PCS, MCS, CD4+ count, and viral load). The primary predictor of interest will be baseline hard drug use status. All models will adjust for the specified demographic variables: age, BMI, smoking status, education, and race. We will approach modeling from both a Bayesian and non-Bayesian approach. For the non-Bayesian, we will use standard Ordinary Least Squares (OLS) regression, reporting point estimates, 95% confidence intervals, and p-values. For the Bayesian, we will fit Bayesian linear regression models. We will specify non-informative or weakly informative prior distributions for our parameters and estimate posterior distributions using Markov Chain Monte Carlo (MCMC) methods. We will report posterior means, 95% Highest Posterior Density (HPD) credible intervals, and posterior probabilities to evaluate the primary research question. We will compare the results from both frameworks and determine if there are any differences in the interpretations. Finally, we will investigate if any relationship between hard drug use and the outcomes is explained by adherence to the treatment protocol.

Methods

Data Management

This analysis only considered subjects with the four primary measures, PCS, MCS, CD4+ count, and viral load, measured at both baseline and two years. Consistent with standard practice in the literature, viral load was analyzed on the $\log_{10}$ scale, where a one-unit increase corresponds to a 10-fold (order of magnitude) increase in viral load. We investigated if missingness in the primary measures differed between hard drug users and non hard drug users. Due to low sample sizes within specific strata, the categories of several demographic variables were combined: race/ethnicity was combined into non-Hispanic White and other, education was combined into no college degree and college degree or greater, and smoking was combined into current smokers and never/former smokers. Study protocol adherence at year 2 was dichotomized as $\geq 95\%$ and $< 95\%$. Additionally, biologically improbable BMI measurements (BMI < 0 or > 400) were excluded from the analysis. Finally, for each outcome, the difference between measurements at baseline and year 2 was calculated and assess for normality and outliers. A table of descriptive statistics was created for the demographic variables used in the analysis as well as the baseline outcome measurements.

Analysis Approach

Prior to analysis, the difference between baseline and year 2 outcomes were visualized for each outcome using histograms.

Bayesian Modeling and Diagnositics.

For all Bayesian models, the No-U-Turn Sampler (NUTS) was used with 4 chains. Parameters were initialized using random values sampled from a Unif$(-2, 2)$ distribution on the unconstrained parameter space. Each chain ran for 1000 warmup iterations and 2000 total iterations per chain. The Bayesian models employed normally distributed likelihoods and non-informative priors: intercepts and fixed effects had Normal$(0, 100)$ priors, and the residual standard errors had Half-Normal$(0, 100)$ priors.

The same diagnostic procedure was followed for all of the Bayesian models. First, we ensured the effective sample size (ESS) was sufficient for all parameters (> 1000) and that the R-hat (Gelman-Rubin statistic) was close to 1, indicating convergence.

Convergence was further assessed by plotting each parameter's density for each chain. Additionally, trace and autocorrelation plots were visually inspected to evaluate chain mixing and sample dependency.

Results

Table 1 shows summary statistics for the demographic variables and primary measures at baseline. Missingness in the primary measures was higher for the hard drug use group (41%) than the non-hard drug use group (33%), indicating that missingness may depend on hard drug use status.

all models had effective sample sizes well over 1000 and no evidence of issues with autocorrelation. convergence was assessed with trace plots and chain density overlays for each parameter.

there was no evidence of issues with convergence in any of the bayesian models with all rhats near 1.

Discussion

Potential Limitation of this work: it could be that hard drug users have better adherence because the ones that dont have good adherence did not have measurements at year 2