Data Analysis Plan

Introduction

The data for this project come from the Multicenter AIDS Cohort Study (MACS), a longitudinal prospective cohort study of HIV-infected men followed annually after initiation of highly active antiretroviral therapy (HAART). Our study population consists of HIV-positive men enrolled in MACS who initiated HAART and had complete baseline (year 0) and year 2 follow-up data. The primary exposure of interest is baseline hard drug use status (yes/no). We examine four treatment response outcomes measured at baseline and year 2: (1) Viral load (log10-transformed copies/mL), (2) CD4+ T-cell Count (LEU3N), (3) Physical Quality of Life (AGG_PHYS) and (4) Mental Quality of Life (AGG_MENT). The primary hypothesis is that compared to non-drug users, individuals with baseline hard drug use will show worse treatment response at year 2, indicated by higher viral loads, lower CD4+ counts, and lower quality-of-life scores. Potential confounders include age, BMI, race/ethnicity, education, and baseline smoking status.

Preliminary Methods

Data Cleaning and Management: Special missing or improbable values were recoded to missing according to the codebook. Viral load was log-transformed to address extreme right skewness. CD4 counts were examined both on the raw and square-root scale. The analytic dataset includes subjects meeting all of the following criteria: (1) complete baseline data; (2) complete follow-up data at year 2; and (3) initiated ART (everART = 1). This selection ensures that we analyze only individuals with confirmed HAART initiation and complete outcome data at both timepoints. The exposure and all baseline covariates were derived

from year 0 data. Race and education were collapsed into broader categories to improve model stability.

Statistical Analysis: Baseline characteristics will be summarized overall and stratified by hard drug use status. For each of the four outcomes, an ANCOVA framework will be used: $Y_2 = \beta_0 + \beta_1(\text{Hard Drug Use}_{baseline}) + \beta_2 Y_0 + \mathbf{X}\beta + \varepsilon$, where Y_2 = Year 2 outcome, Y_0 = Baseline value of the outcome, X = baseline covariates. β_1 represents the adjusted difference in Year 2 outcome between baseline hard drug users and non-users, and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. This approach improves efficiency and controls for baseline outcome differences. We will also fit the same models using Bayesian methods with Stan. For each outcome, we will specify weakly informative priors on regression coefficients (N(0, sd = 100)) and a half-normal prior on residual standard deviations. Posterior estimates will be obtained via MCMC sampling. Posterior inference will include 95% HPDI and effective sample size diagnostics. Model fit will be evaluated using information criteria for Bayesian models. The key parameters of interest will be compared in terms of point estimates and interval estimates.

Preliminary Descriptive Analysis - Table 1

The analytic cohort includes 506 participants at baseline. Of these, 39 (7.7%) reported hard drug use, and 467 (92.3%) did not. Baseline age and viral load were similar between groups. Hard drug users were more likely to be current smokers and to have lower educational attainment. $CD4^+$ T-cell counts and quality-of-life scores were modestly lower among hard drug users. These differences support adjustment for baseline covariates in subsequent regression analyses.

Table 1: Baseline characteristics of participants by baseline hard drug use status.

	No	Yes	Overall
	(N=467)	(N=39)	(N=506)
Age (years)			
Mean (SD)	43.2 (8.72)	44.6 (9.49)	43.3 (8.78)
Median [Min, Max]	43.0 [20.0, 73.0]	47.0 [29.0, 61.0]	43.0 [20.0, 73.0]
BMI (kg/m2)			
Mean (SD)	26.5 (23.4)	23.6 (3.45)	26.3 (22.5)
Median [Min, Max]	24.7 [16.5, 514]	23.3 [18.0, 31.2]	24.6 [16.5, 514]
Missing	14 (3.0%)	3 (7.7%)	17 (3.4%)
Race			
Black	123 (26.3%)	17 (43.6%)	140 (27.7%)
Other	24 (5.1%)	3 (7.7%)	27 (5.3%)
White	320 (68.5%)	19 (48.7%)	339 (67.0%)
Education			
College+	205 (43.9%)	10 (25.6%)	215 (42.5%)
High school or less	95 (20.3%)	16 (41.0%)	111 (21.9%)
Some college	167 (35.8%)	13 (33.3%)	180 (35.6%)
Smoking status			
Current	166 (35.5%)	30 (76.9%)	196 (38.7%)
Former	159 (34.0%)	9 (23.1%)	168 (33.2%)
Never	142 (30.4%)	0 (0%)	142 (28.1%)
Viral load (log10 copies/mL)			
Mean (SD)	4.52 (0.920)	4.52 (0.855)	4.52 (0.914)
Median [Min, Max]	4.51 [0.237, 8.28]	4.44 [2.87, 6.40]	4.48 [0.237, 8.28]
Missing	9 (1.9%)	0 (0%)	9 (1.8%)
CD4+ T-cell count			
Mean (SD)	378 (202)	352 (195)	376 (202)
Median [Min, Max]	361 [10.9, 1220]	271 [10.9, 650]	361 [10.9, 1220]
Missing	8 (1.7%)	0 (0%)	8 (1.6%)
Physical QOL (SF-36, 0-100)			
Mean (SD)	51.3 (9.21)	47.7 (8.50)	51.0 (9.20)
Median [Min, Max]	53.7 [19.2, 69.0]	46.6 [28.8, 62.9]	53.5 [19.2, 69.0]
Missing	1 (0.2%)	0 (0%)	1 (0.2%)
Mental QOL (SF-36, 0-100)			
Mean (SD)	45.5 (13.7)	42.3 (11.2)	45.3 (13.5)
Median [Min, Max]	49.7 [7.23, 69.8]	45.3 [22.5, 59.6]	49.1 [7.23, 69.8]
Missing	1 (0.2%)	0 (0%)	1 (0.2%)