INTERIM DATA ANALYSIS PLAN

Analysts: Rebecca Du

Investigators: Dr. Camille Moore

Date: February 2026

1 Introduction

This project is a secondary data analysis of the Multicenter AIDS Cohort Study (MACS), a prospective cohort study of HIV-1 infection in homosexual and bisexual men in four U.S. cities. The dataset includes up to 8 years of annual laboratory and quality of life measures on 715 HIV-infected men after beginning highly active antiretroviral treatment (HAART), with year 0 representing each subject's last untreated visit. Limited in vitro and animal evidence suggests that hard drug use (heroin, cocaine, injection drugs) inhibits the immune system and increases HIV replication, but human evidence is inconclusive. We hypothesize that baseline hard drug users will have worse treatment response at 2 years post-HAART across four outcomes: higher viral load (VLOAD), lower CD4+ T cell count (LEU3N), and lower aggregate physical and mental quality of life scores (AGG_PHYS, AGG_MENT) from the SF-36, after adjusting for baseline values and confounders. Each hypothesis will be tested in both a frequentist and Bayesian framework, and results compared.

2 Preliminary Methods

2.1 Data cleaning and analytic sample.

Sentinel values in the codebook (9 = insufficient data, -1 = improbable, -9 = not specified) were set to missing across all affected variables. Categorical variables with sparse cells were collapsed (race into 4 groups, education into 3, income into 3 brackets). Viral load was $\log_{10}$-

transformed as $\log_{10}$(VLOAD + 1) to address extreme right skew. The analytic sample consists of subjects with both a baseline and year 2 visit ($n = 506$ of 715; 29.2% lost to follow-up). At baseline, 66 subjects (9.2%) reported hard drug use and 649 did not. Logistic regression indicated that attrition is predicted by observed baseline characteristics (race, income, CD4 count), suggesting data are not missing completely at random. The primary analysis will use complete cases, with multiple imputation as a sensitivity analysis.

2.2 Statistical analysis.

For each outcome, an unadjusted comparison between exposure groups will be followed by a multivariable linear regression using an ANCOVA framework, regressing the year 2 outcome on baseline hard drug use and adjusting for the corresponding baseline outcome value. ANCOVA is favored over a change-score approach based on moderate-to-strong baseline–year 2 correlations observed in exploratory scatterplots. Confounders were selected by combining baseline imbalances between exposure groups (standardized mean differences > 0.2 in descriptive tables) with biological plausibility: age, race/ethnicity, education, income, BMI, smoking status, CESD depression score, and kidney disease. ART adherence is excluded from the primary model as a potential mediator; a sensitivity analysis including adherence will assess the direct effect. QQ plots indicate that log-transformed viral load, CD4 count, and both quality of life scores are approximately normal with minor tail departures; residual diagnostics will be checked, and robust standard errors or a Student-t likelihood considered if violations are substantial. P-values < 0.05 will be considered statistically significant.

2.3 Bayesian analysis.

The same models will be fit in a Bayesian framework with Gaussian likelihoods (log scale for viral load) and weakly informative Normal$(0, \sigma)$ priors on regression coefficients, where σ is scaled to each outcome's standard deviation to allow the data to drive inference. Posterior distributions will be summarized via posterior means, 95% credible intervals, and the posterior probability that the hard drug use effect is in the hypothesized direction. Bayesian and frequentist results will be compared in terms of point estimates, interval widths, and inferential conclusions.