Project 1 Final report

Shuai Zhu

2024-10-05

1 Introduction

The data used in this analysis come from the ongoing Multicenter AIDS Cohort Study (MACS), a prospective cohort study designed to understand the natural and treated histories of HIV-1 infection in homosexual and bisexual men across four major cities in the United States. This dataset includes eight years of longitudinal data from 715 HIV-infected men, capturing laboratory measurements, quality of life scores, demographic information, and other health-related data collected after the initiation of highly active antiretroviral therapy (HAART), which is the standard treatment for patients with HIV.

The primary research question is to examine how treatment response, two years after initiating HAART, differs between individuals who reported hard drug use at baseline and those who did not. Four key measures of treatment response are considered: viral load, CD4+ T cell counts, and physical and mental quality of life scores.

2 Methods

2.1 Data cleaning

Baseline and two year measurement was filtered for further analysis since the purpose of this project is find out how treatment response differ two years after treatment. The data with BMI greater than 200 or less than 0 was removed since it is impossible. The records with complete case was used for further analysis and the number of observations was reduced to 425. Furthermore, BMI was categorized to four levels underweight (BMI $<$ 18.5 kg/m2), healthy (BMI 18.5 - 24.9 kg/m2), overweight(BMI 24.9 - 30 kg/m2) and obese (BMI $>$

30 kg/m2). Adherence was dichotomized into >=95% and <95%. The education levels was collapsed into three levels (High school or before, Some college, and Graduate or Post-graduate).

2.2 Data analysis

Both frequentist and Bayesian approaches were employed to assess differences in treatment response by baseline hard drug use. The four key outcomes used to assess treatment response were viral load, CD4+ T cell counts, and physical and mental quality of life scores (AGG_PHYS and AGG_MENT). The objective was to model the impact of hard drug use, adjusting for several covariates, including baseline treatment response, BMI, age, education level, and adherence. Viral load, due to its skewed distribution, was log-transformed to meet the assumption of normality. For the frequentist approach, four multivariable linear regression models were fitted, each predicting one of the treatment response outcomes. The model assumptions—independence, linearity, homoscedasticity, and normality—were carefully evaluated using standard diagnostic tools. Independence was verified by inspecting the structure of the data and residuals. Linearity was checked by assessing the relationship between predictors and outcome variables using residual plots. Homoscedasticity was evaluated using residuals vs. fitted values plots. Normality of residuals was tested through QQ-plots.

For Bayesian regression models, both non-informative and vague priors were used. The non-informative priors of beta are distributed with mean 0 and standard deviation 10^7. The vague priors of beta are distributed with mean 0 and standard deviation 10^6. The prior distribution for the model error was set as a half-Cauchy distribution with a scale parameter of 2.5. Bayesian inference was carried out using Markov Chain Monte Carlo (MCMC) sampling. Each model was run with 4 MCMC chains, with each chain consisting of 2,000 iterations, including a 1,000 iteration burn-in period to ensure convergence. The posterior distributions for all model parameters were summarized, and credible intervals were used to quantify uncertainty in the estimates.

For the frequentist models, standard metrics such as p-values and confidence intervals were used to assess the significance and effect sizes of the predictors. For the Bayesian models, the convergence of MCMC chains was assessed through trace plots. Posterior means and 95% credible intervals were reported for each parameter to provide a full picture of the uncertainty around the estimates.

R version 4.4.4 was used for all models.

3 Result

The percentage of missing data was visualized in figure one. It is easily to see that four outcome measurement of treament response at year 2 and education level had about 30% missing. Viral load and CD4+ T Cell Count at baseline and BMI level had missing range from 3.4% to 5.9%. Physical quality of life and mental quality of life at Baseline, age and adherence had almost no missing. 425 participants was kept after removing subject with any missing value. Only 58% subjects was used for analysis. Among those subjects with complete cases, 35(8.2%) participants reported those as hard drugs user. This was shown in tableone.

3.1 Frequentist

After adjusting for baseline log viral load, age, BMI, education levels, and adherence, the frequentist multivariable linear regression model suggests that patients who identified as hard drug users at baseline had 0.96 times the viral load compared to those who did not identify as hard drug users at baseline (95% CI: (0.39, 2.39), p value = 0.9367). This indicates that, on average, hard drug users had a slightly lower viral load compared to non-drug users.

After adjusting for baseline CD4+ T cell counts, age, BMI, education levels, and adherence, the frequentist multivariable linear regression model suggests that patients who identified as hard drug users at baseline had 186.284 lower CD4+ T cell counts compared to those who did not identify as hard drug users at baseline with the confidence interval suggesting that the true difference could range from a reduction of 124 to 249 CD4+ T cell(p value < 0.05).

After adjusting for baseline CD4+ T cell counts, age, BMI, education levels, and adherence, the frequentist multivariable linear regression model suggests that patients who identified as hard drug users at baseline had 186.284 lower CD4+ T cell counts compared to those who did not identify as hard drug users at baseline with the confidence interval suggesting that the true difference could range from a reduction of 124 to 249 CD4+ T cell(p value < 0.05).

After adjusting for baseline physical quality of life score, age, BMI, education levels, and adherence, the frequentist multivariable linear regression model suggests that patients who identified as hard drug users at baseline had an average 3.32-point lower physical quality of life score compared to non-drug users (95% CI: -6.02, -0.62). The confidence interval indicates that the true difference in physical quality of life score could range from a reduction of 6.02 to 0.62 points, and since the p-value is less than 0.05, this result is statistically significant,

indicating a meaningful negative association between hard drug use and physical quality of life.

After adjusting for baseline physical quality of life score, age, BMI, education levels, and adherence, the frequentist multivariable linear regression model suggests that patients who identified as hard drug users at baseline had an average 1.17-point lower physical quality of life score compared to non-drug users (95% CI: -4.60, -2.25). The confidence interval indicates that the true difference in physical quality of life score could range from a reduction of 4.60 to 2.25 points, and since the p-value is less than 0.05, this result is statistically significant. This suggests a meaningful negative association between hard drug use and physical quality of life.

3.2 Bayesian

4 Conclusion

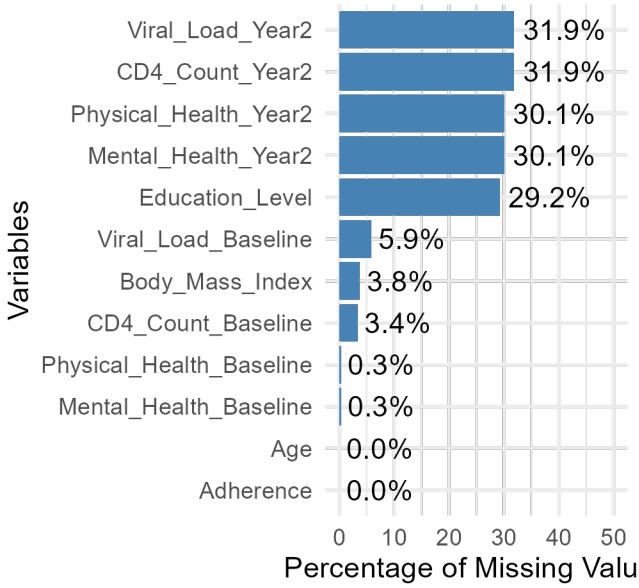

Figure 1: Percentage of Missing Data by Variable

Table 1: Summary of outcomes and predictors stratified by hard drugs

	Hard drugs use	No hard drugs use	Overall
	(N=66)	(N=649)	(N=715)
Viral Load at Baseline			
Mean (SD)	265000 (609000)	823000 (10200000)	771000 (9730000)
Median [Min, Max]	45900 [739, 2520000]	33700 [1.73, 191000000]	33800 [1.73, 191000000]
Missing	3 (4.5%)	39 (6.0%)	42 (5.9%)
Viral Load at Year 2			
Mean (SD)	32900 (114000)	6220 (40800)	8360 (51100)
Median [Min, Max]	42.0 [0.740, 424000]	31.0 [0.247, 711000]	31.0 [0.247, 711000]
Missing	27 (40.9%)	201 (31.0%)	228 (31.9%)
CD4+ T Cell Count at Baseline			
Mean (SD)	335 (187)	385 (211)	380 (210)
Median [Min, Max]	301 [10.9, 778]	363 [10.9, 1220]	361 [10.9, 1220]
Missing	0 (0%)	24 (3.7%)	24 (3.4%)
CD4+ T Cell Count at Year 2			
Mean (SD)	366 (240)	560 (265)	544 (268)
Median [Min, Max]	348 [60.0, 971]	537 [39.5, 1730]	516 [39.5, 1730]
Missing	27 (40.9%)	201 (31.0%)	228 (31.9%)
Physical Quality of Life at Baseline			
Mean (SD)	44.8 (9.58)	50.7 (9.37)	50.1 (9.54)
Median [Min, Max]	45.3 [27.0, 62.9]	53.3 [19.2, 69.0]	52.8 [19.2, 69.0]
Missing	0 (0%)	2 (0.3%)	2 (0.3%)
Physical Quality of Life at Year 2			
Mean (SD)	43.8 (11.6)	49.9 (10.0)	49.4 (10.2)
Median [Min, Max]	43.2 [18.2, 63.9]	53.3 [14.8, 69.1]	53.0 [14.8, 69.1]
Missing	27 (40.9%)	188 (29.0%)	215 (30.1%)
Mental Quality of Life at Baseline			
Mean (SD)	43.4 (12.4)	45.6 (13.4)	45.4 (13.3)
Median [Min, Max]	46.6 [20.8, 59.6]	49.7 [7.23, 69.8]	49.3 [7.23, 69.8]
Missing	0 (0%)	2 (0.3%)	2 (0.3%)
Mental Quality of Life at Year 2			
Mean (SD)	45.9 (13.5)	47.6 (11.8)	47.5 (11.9)
Median [Min, Max]	47.6 [21.3, 65.3]	51.5 [10.5, 66.7]	51.2 [10.5, 66.7]
Missing	27 (40.9%)	188 (29.0%)	215 (30.1%)
Age (years)			
Mean (SD)	43.7 (9.85)	42.4 (9.38)	42.6 (9.42)
Median [Min, Max]	45.5 [26.0, 61.0]	42.0 [19.0, 73.0]	42.0 [19.0, 73.0]
Body Mass Index (kg/m²)			
Healthy	41 (62.1%)	309 (47.6%)	350 (49.0%)
Obsese	5 (7.6%)	81 (12.5%)	86 (12.0%)
Overweight	13 (19.7%)	210 (32.4%)	223 (31.2%)
Underweight	7 (10.6%)	22 (3.4%)	29 (4.1%)
Missing	0 (0%)	27 (4.2%)	27 (3.8%)
Adherence Level			
Mean (SD)	0.576 (0.498)	0.641 (0.480)	0.635 (0.482)
Median [Min, Max]	1.00 [0, 1.00]	1.00 [0, 1.00]	1.00 [0, 1.00]
Education Level			
High school	16 (24.2%)	95 (14.6%)	111 (15.5%)
some college	14 (21.2%)	272 (41.9%)	286 (40.0%)
Graduate, Post Graduate	9 (13.6%)	100 (15.4%)	109 (15.2%)
Missing	27 (40.9%)	182 (28.0%)	209 (29.2%)

Table 2: Comparions of estimate and 95% confident intervals and credit intervals between Frequentist and Bayesian approach

	Frequentist	2.5%	97.5%	Vague	2.5%	97.5%
Log Viral Load	-0.037	-0.946	0.873	-0.035	-0.916	0.852
CD4+ Count	-186.284	-248.508	-124.060	-186.052	-247.837	-123.013
Physical score	-3.319	-6.015	-0.623	-3.307	-6.069	-0.610
Mental score	-1.172	-4.598	2.253	-1.183	-4.526	2.244

Table 3: Estimatae and 95% Credit interval of Bayesian approach with noninformative priors

	Estimate	2.5%	97.5%
Log Viral Load	-0.045358	-0.9430382	0.8793678
CD4+ Cell Count	-185.667304	-248.4468299	-124.5246236
Physical score	-3.321138	-6.0445899	-0.6430269
Mental score	-1.199944	-4.6906949	2.3279289