Project 1 final report

Shuai Zhu

2024-10-04

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.

3 Result

4 Conclusion

Table 1: Summary of outcomes and predictors by hard drugs

	0	1	Overall
	(N=390)	(N=35)	(N=425)
Log-transformed Viral Load at Baseline			
Mean (SD)	10.5 (2.05)	10.6 (2.01)	10.5 (2.05)
Median [Min, Max]	10.4 [2.20, 19.1]	10.3 [6.61, 14.7]	10.4 [2.20, 19.1]
Log-transformed Viral Load at Year 2			
Mean (SD)	4.07 (2.73)	4.13 (3.30)	4.07 (2.78)
Median [Min, Max]	3.43 [-1.40, 13.5]	3.74 [-0.301, 13.0]	3.43 [-1.40, 13.5]
CD4+ T Cell Count at Baseline			
Mean (SD)	377 (197)	361 (204)	376 (197)
Median [Min, Max]	361 [12.4, 1220]	453 [10.9, 650]	361 [10.9, 1220]
CD4+ T Cell Count at Year 2			
Mean (SD)	565 (258)	372 (252)	549 (263)
Median [Min, Max]	544 [39.5, 1730]	357 [60.0, 971]	529 [39.5, 1730]
Physical Quality of Life at Baseline			
Mean (SD)	51.5 (8.67)	48.8 (6.86)	51.3 (8.56)
Median [Min, Max]	53.7 [22.4, 69.0]	46.7 [31.4, 62.9]	53.5 [22.4, 69.0]
Physical Quality of Life at Year 2			
Mean (SD)	50.0 (9.94)	44.4 (12.1)	49.5 (10.2)
Median [Min, Max]	53.3 [14.8, 68.9]	45.5 [18.2, 63.9]	53.2 [14.8, 68.9]
Mental Quality of Life at Baseline			
Mean (SD)	44.9 (13.9)	42.6 (11.3)	44.7 (13.7)
Median [Min, Max]	49.3 [7.23, 66.0]	45.1 [22.9, 59.6]	48.9 [7.23, 66.0]
Mental Quality of Life at Year 2			
Mean (SD)	47.7 (11.6)	46.2 (14.2)	47.6 (11.8)
Median [Min, Max]	51.2 [10.5, 66.7]	49.6 [21.3, 65.3]	51.2 [10.5, 66.7]
Age (years)			
Mean (SD)	43.0 (8.82)	44.2 (9.43)	43.1 (8.86)
Median [Min, Max]	43.0 [20.0, 73.0]	47.0 [29.0, 61.0]	43.0 [20.0, 73.0]
Body Mass Index (kg/m²)			
Healthy	192 (49.2%)	25 (71.4%)	217 (51.1%)
Obsese	47 (12.1%)	2 (5.7%)	49 (11.5%)
Overweight	138 (35.4%)	7 (20.0%)	145 (34.1%)
Underweight	13 (3.3%)	1 (2.9%)	14 (3.3%)
Adherence Level			
Mean (SD)	0.897 (0.304)	1.00 (0)	0.906 (0.292)
Median [Min, Max]	1.00 [0, 1.00]	1.00 [1.00, 1.00]	1.00 [0, 1.00]
Education Level			
Graduate, Post Graduate	81 (20.8%)	9 (25.7%)	90 (21.2%)
High school	80 (20.5%)	12 (34.3%)	92 (21.6%)
some college	229 (58.7%)	14 (40.0%)	243 (57.2%)