BIOST 540 Midterm

Zichen Liu & Ben Stan

May 7, 2021

Introduction

The HIV virus decreases a patient's CD4 cell counts over time, leaving the body vulnerable to opportunistic infections. Although no vaccine exists, patients who follow drug regimens as part of antiretroviral therapy can mitigate the devastating health outcomes of AIDS. The AIDS Clinical Trial Group (ACTG) study 193A was a double-blind study where patients with CD4 counts at or below 50 cells per mm^3 were randomized to one of four drug regimens with 600 mg zidovudine: 1) alternating monthly with 400 mg didanosine, 2) plus 2.25 mg of zalcitabine, 3) plus 400 mg of didanosine, and 4) plus 400 mg of didanosine plus 400 mg of nevirapine. We are interested in assessing the treatment effectiveness of the four regimens by evaluating log CD4 counts over the first 48 weeks of follow-up and comparing changes over time between regimens. Additionally, we are interested in any differences in log CD4 counts and treatment effectiveness by gender and baseline age.

Methods

Descriptive Analysis: To accurately investigate the relationship between log CD4 counts and treatment group, we compare the gender and age distributions of the subjects in each group; this is shown in **Table 1**. We then compare log CD4 over the course of the study for each group, shown in **Table 2** and **Figure 1**. *Confirmatory Analysis:* The unbalanced data in this study requires special consideration, as subjects vary in the timing of their measurements, and missing data is highly prevalent. In considering the correlations between measurements, shown in **Figure 2**, we believe that an exponential covariance structure is justified, as the correlation between measurements decreases as the time between them increases, and the data is unbalanced. To examine if changes in log CD4 over time differ between treatment groups, we will use generalized least squares with a linear spline and a knot at 16 weeks. This time point has been cited by previous studies as a change-point in the growth of log CD4. Treating week as a

continuous variable captures the variation in measurement times. **Model 1** will include main effects for week, treatment group, weeks since week 16, and interaction terms between treatment group and the two time variables:

$$E[Y_{ij} | t_{ij}] = \beta_0 + \beta_1 t_{ij} + \beta_2 (t_{ij} - 16)^+ + \beta_3 I_{(group=2)} + \beta_4 I_{(group=3)} + \beta_5 I_{(group=4)} + \beta_6 t_{ij} \times I_{(group=2)} + \beta_7 t_{ij} \times I_{(group=3)} + \beta_8 t_{ij} \times I_{(group=4)} + \beta_9 (t_{ij} - 16)^+ \times I_{(group=2)} + \beta_{10} (t_{ij} - 16)^+ \times I_{(group=3)} + \beta_{11} (t_{ij} - 16)^+ \times I_{(group=4)}$$

Model 2 additionally adjusts for baseline age and gender. For both models, we are interested in the same null hypothesis, H_0: $\beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = 0$. We will use REML to obtain unbiased estimates and standard errors for the coefficients and use a Wald test for significance testing.

Results

Descriptive Analysis: Based on **Table 1**, there are no significant differences between treatment groups in gender or age composition. The total subjects, subjects by gender, mean age, and median age are all comparable across groups. *Missing Data*: Of the 1309 subjects in the study, only 159 of them have observations for all six weeks (12%), and the mean measurements is 3.76. To compare those with missing data to those with complete data, we will consider baseline log CD4. There are 10 subjects without baseline data, so these individuals will be ignored. We consider groups defined by the number of unique weeks with data collected. These are shown in **Table 3**. A noteworthy trend is that the mean baseline log CD4 counts are lower for the subjects with only one or two weeks of data (2.77 log cells/mm^3 for both groups) than subjects with more data collected. Also, subjects with only one measurement taken have the lowest percentage assigned to treatment 4 (17%). *Confirmatory Analysis:* The results of fitting **Model 1** are in **Table 4**, and the results of fitting **Model 2** are in **Table 5**. The results of fitting **Model 1** suggest that the estimated mean difference in log CD4 count for two populations in treatment group 1 that differ in time of measurement by one week is 0.0142 cells/mm^3 when considering times before week 16 and 0.0184 cells/mm^3 when considering times after week 16, with the later week having a lower log CD4 count. When comparing two populations in treatment group 2, the estimated mean difference in log CD4 corresponding to a one week difference in time is 0.00767 log cells/mm^3 before week 16

and 0.0194 after, with the later week having a lower log CD4 count. The estimated trend by week for treatment groups 3 and 4 changes sign at week 16. Before week 16, the estimated mean difference in log CD4 counts for two groups differing by one week is 0.0011 cells/mm^3 in group 3 and 0.0213 cells/mm^3 in group 4, with the later week having higher counts. Following week 16, the estimated mean difference corresponding to a one week difference is 0.0184 cells/mm^3 for group 3 and 0.0197 for group 4, with the later week having lower counts. Based on the magnitude of these effects, it appears that treatment 4 is the most effective. There is strong evidence to suggest that the rates of change over time differ among the treatment groups over the whole study period (p-value < 0.0001) and over the period following week 16 (p-value < 0.0001). Adjusting for the covariates of age and gender did not change these conclusions (p-values remained < 0.0001 for both comparisons). This is consistent with the exploratory analyses, which showed that the four groups had similar distributions of gender and baseline age. However, trends such as patients with fewer measurements having different baseline log CD4 values and different likelihoods of being assigned to treatment groups suggest unaddressed biases in missing data. Unfortunately, the inconsistency of the data makes this issue difficult to explore further.

Tables and Figures

Treatment group	N	Gender	Min age	P25 age	Mean age	Median age	P75 age	Max age
1	325	M: 290 F: 35	15.8	32.0	37.9	36.6	42.8	70.6
2	324	M: 282 F: 42	19.4	31.6	37.7	36.9	42.4	66.4
3	330	M: 286 F: 44	14.9	31.9	37.5	37.1	41.6	74.2
4	330	M: 289 F: 41	15.8	31.7	37.9	36.9	42.6	62.8

Table 1: Descriptive statistics for patient characteristics by treatment group.

Treatment group	Week[a]	N[b]	Min log CD4	P25 log CD4	Mean log CD4	Median log CD4	P75 log CD4	Max log CD4
1	0	320	0	2.44	2.98	3.04	3.56	4.98
	8	225	0	2.40	2.83	2.89	3.50	5.31
	16	241	0	2.30	2.80	2.89	3.43	5.40
	24	170	0	2.20	2.58	2.64	3.12	4.22
	32	200	0	2.08	2.56	2.64	3.09	4.54
	40	57	0	1.61	2.31	2.48	3.04	3.93
2	0	322	0	2.40	2.93	3.07	3.61	5.16
	8	219	0	2.40	2.95	3.04	3.66	5.27
	16	247	0	2.30	2.84	2.89	3.50	5.11
	24	177	0	2.20	2.62	2.71	3.40	4.90
	32	181	0	2.08	2.66	2.71	3.40	4.95
	40	71	0	1.50	2.35	2.40	3.22	4.86
3	0	327	0	2.40	2.91	3.02	3.58	5.20

	8	224	0	2.40	3.09	3.04	3.76	5.75
	16	254	0	2.30	2.97	3.04	3.71	5.51
	24	173	0	2.20	2.78	2.83	3.50	5.37
	32	192	0	2.30	2.82	2.89	3.47	5.66
	40	59	0	2.01	2.78	2.56	3.70	5.36
4	0	330	0	2.30	2.84	3.01	3.48	5.06
	8	233	0	2.40	3.16	3.22	4.11	5.78
	16	254	0	2.40	3.24	3.33	3.97	6.30
	24	173	0	2.40	2.99	3.00	3.76	5.77
	32	211	0	2.25	2.96	3.00	3.73	6.10
	40	71	0	2.14	2.90	3.00	3.47	5.37

Table 2: Descriptive statistics for outcomes by treatment group.
[a] Week is rounded to the nearest multiple of 8 weeks
[b] Patients with multiple measurements rounded to the same week bin are deduplicated

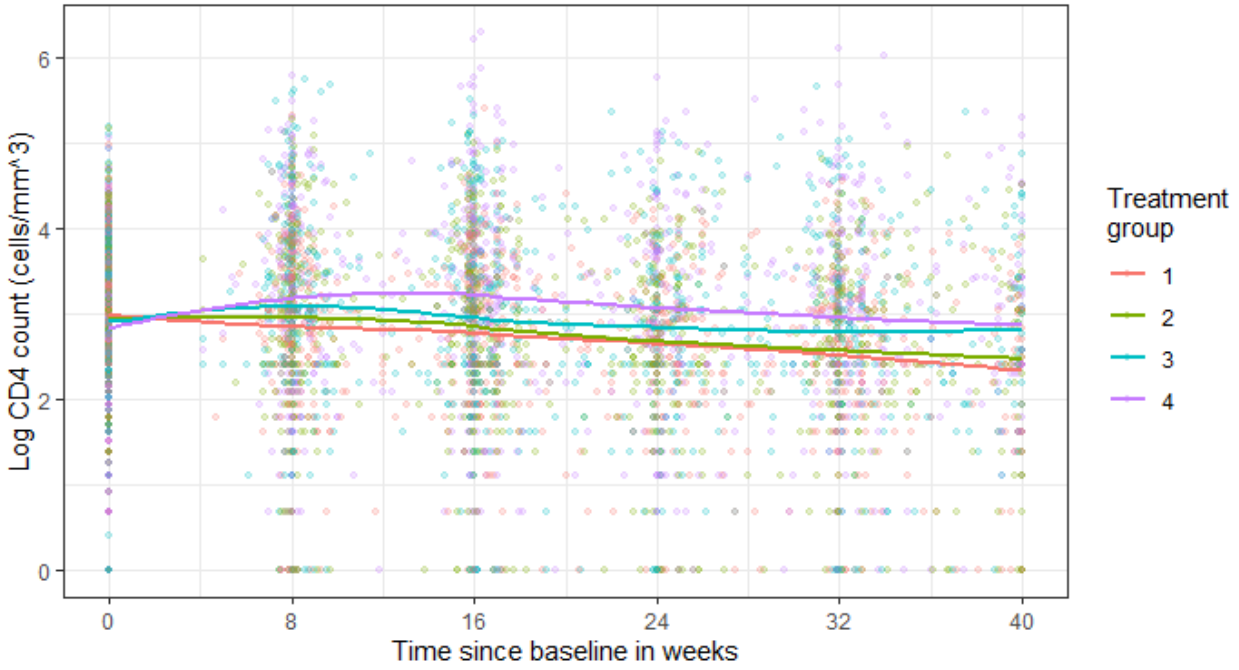

Figure 1: Plot of log CD4 counts over time for all subjects with LOESS smooths for each treatment group.

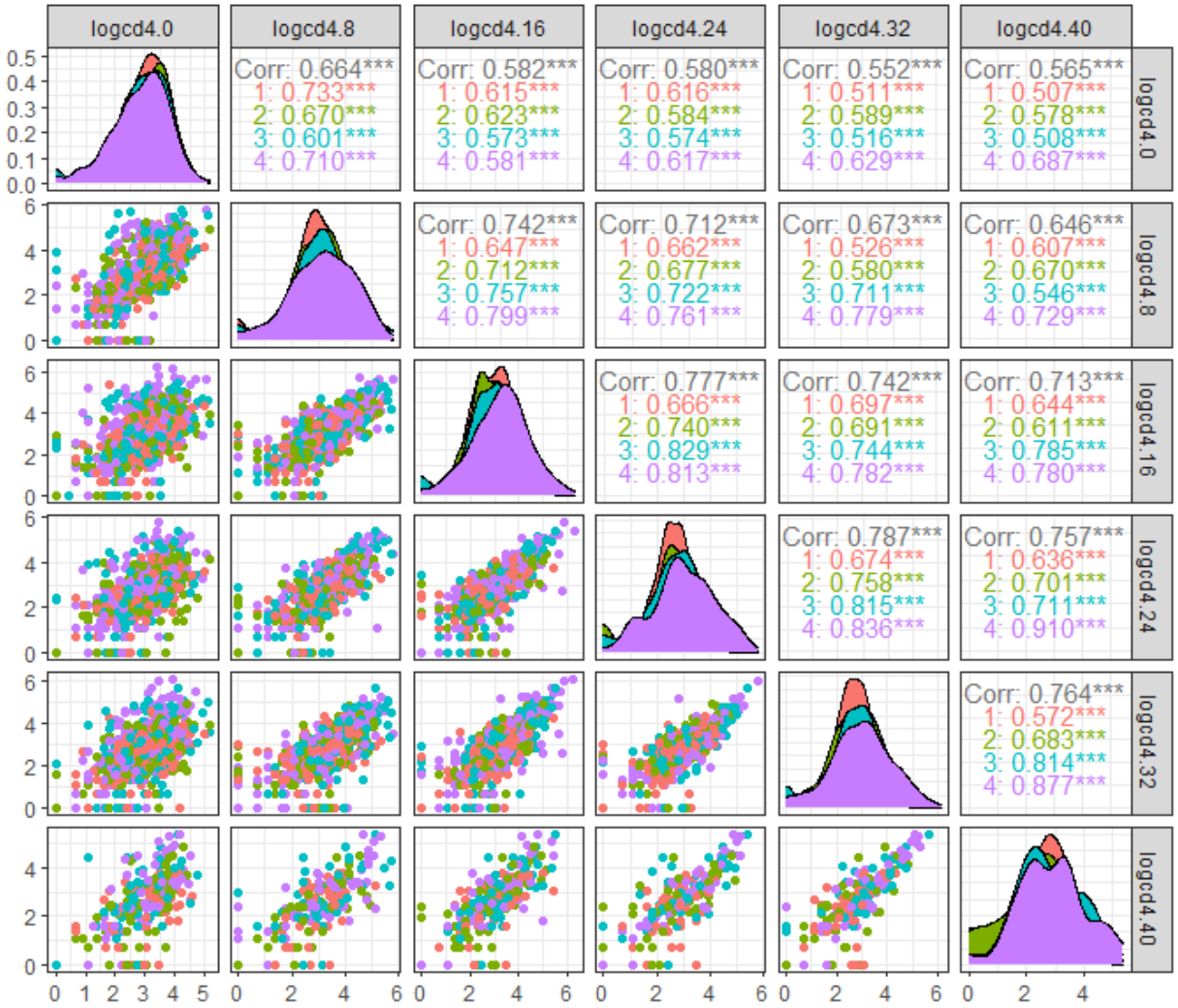

Figure 2: Correlations between log CD4 values by week and treatment group.

Num weeks[a]	N	Gender	% in group 1/2/3/4	Mean age	Min log CD4	P25 log CD4	Mean log CD4	Med log CD4	P75 log CD4	Max log CD4
1	128	M: 108 F: 20	1: 26% 2: 30% 3: 27% 4: 17%	38.5	0	2.08	2.61	2.77	3.47	4.34
2	153	M: 124 W: 29	1: 24% 2: 22% 3: 21% 4: 33%	36.9	0	2.20	2.72	2.77	3.47	5.04
3	257	M: 223 W: 34	1: 27% 2: 22%	37.7	0	2.56	3.02	3.11	3.54	5.19

			3: 29% 4: 22%							
4	263	M: 231 W: 32	1: 21% 2: 28% 3: 24% 4: 27%	38.3	0	2.40	3.00	3.07	3.60	5.11
5	339	M: 305 W: 34	1: 27% 2: 23% 3: 25% 4: 25%	37.3	0	2.40	2.91	3.02	3.60	5.16
6	159	M: 147 W: 12	1: 21% 2: 26% 3: 23% 4: 30%	38.0	0	2.48	3.00	3.04	3.60	4.76

Table 3: Descriptive statistics for baseline values for groups defined by the number of measurements taken; group 6 corresponds to subjects that had measurements taken in all time periods of the study.
[a] Patients with multiple measurements rounded to the same week bin are deduplicated

	Estimate	95%CI Lower	95%CI Upper	p-value
Intercept	2.975	2.861	3.089	$< 1*10^{-5}$
Group 2	-0.0160	-0.177	0.145	0.845
Group 3	-0.0482	-0.209	0.112	0.556
Group 4	-0.0952	-0.255	0.0651	0.244
Week	-0.0142	-0.0207	-0.00763	$< 1*10^{-5}$
Week Spline	-0.00414	-0.0150	0.00672	0.455
Group 2 * Week	0.00653	-0.00263	0.0157	0.163
Group 3 * Week	0.0153	0.00620	0.0244	0.001
Group 4 * Week	0.0355	0.0265	0.0446	$< 1*10^{-5}$
Group 2 * Week Spline	-0.00754	-0.0227	0.00762	0.330

Group 3 * Week Spline	-0.0154	-0.0306	-0.000183	0.047
Group 4 * Week Spline	-0.0369	-0.0520	-0.0219	$< 1*10^{-5}$

Table 4: Estimates, 95% confidence intervals, and p-values for the regression coefficients of **Model 1**.

	Estimate	95%CI Lower	95%CI Upper	p-value
Intercept	2.659	2.379	2.940	$< 1*10^{-5}$
Baseline Age	0.0111	0.00494	0.0172	0.0004
Gender	-0.115	-0.269	0.0385	0.142
Group 2	-0.0174	-0.178	0.143	0.831
Group 3	-0.0467	-0.207	0.113	0.567
Group 4	-0.0976	-0.257	0.0621	0.231
Week	-0.0141	-0.0206	-0.00756	$< 1*10^{-5}$
Week Spline	-0.00417	-0.0150	0.00669	0.452
Group 2 * Week	0.00651	-0.00266	0.0157	0.164
Group 3 * Week	0.0152	0.00613	0.0243	0.001
Group 4 * Week	0.0355	0.0264	0.0446	$< 1*10^{-5}$
Group 2 * Week Spline	-0.00754	-0.0227	0.00762	0.330
Group 3 * Week Spline	-0.0153	-0.0305	-0.000120	0.048
Group 4 * Week Spline	-0.0369	-0.0520	-0.0219	$< 1*10^{-5}$

Table 5: Estimates, 95% confidence intervals, and p-values for the regression coefficients of **Model 2**.

Code

```
## Load necessary libraries
rm(list = ls())
library(ggplot2)
library(GGally)
library(reshape2)
library(nlme)
library(tidyverse)

## Load data and create index, week_rounded, and spline columns
cd4 <- read.csv('datasets/cd4.csv')[,-1] # already in long format
cd4 <- cd4 %>%
  group_by(id) %>%
  mutate(index = row_number()) %>%
  ungroup()
cd4 <- cd4 %>%
  mutate(week_rounded = round(week/8)*8) %>%
  mutate(week_spline = (week-16)*(week>16))

## Summary by index
cd4 %>%
  group_by(index) %>%
  summarise(n = n(),
            n_men = sum(sex),
            n_women = n()-sum(sex),
            min_week = min(week, na.rm=T),
            max_week = max(week, na.rm=T),
            mean_week = mean(week, na.rm=T),
            p25_week = quantile(week, 0.25, na.rm=T),
            median_week = quantile(week, 0.5, na.rm=T),
            p75_week = quantile(week, 0.75, na.rm=T))
```

```
# Based on this, rounding week to nearest 8-week interval seems reasonable

## Table 1: Summary by age and gender
cd4_dedupe <- cd4[!duplicated(cd4$id),]
table1 <- cd4_dedupe %>%
  group_by(group) %>%
  summarise(n = n(),
            n_males = sum(sex),
            n_females = n() - sum(sex),
            min_age = min(age),
            p25_age = quantile(age, 0.25),
            mean_age = mean(age),
            median_age = quantile(age, 0.5),
            p75_age = quantile(age, 0.75),
            max_age = max(age))

## Table 2: Summary by group, week_rounded
cd4_dedupe <- cd4[!duplicated(cd4[,c('id', 'week_rounded')]),]
table2 <- cd4_dedupe %>%
  group_by(group, week_rounded) %>%
  summarise(n = n(),
            min_cd4 = min(logcd4, na.rm=T),
            p25_cd4 = quantile(logcd4, 0.25, na.rm=T),
            mean_cd4 = mean(logcd4, na.rm=T),
            median_cd4 = quantile(logcd4, 0.5, na.rm=T),
            p75_cd4 = quantile(logcd4, 0.75, na.rm=T),
            max_cd4 = max(logcd4, na.rm=T))

## Figure 1: Plot all subjects' data with LOESS line
figure1 <- ggplot(data=cd4, aes(x=week, y=logcd4, color=as.factor(group))) +
```

```
  geom_point(alpha=0.2, size=1) +
  geom_smooth(method='loess', se=F) +
  scale_x_continuous(breaks=c(0,8,16,24,32,40)) +
  labs(x='Time since baseline in weeks', y='Log CD4 count (cells/mm^3)',
       color='Treatment\ngroup', title='Log CD4 over time with LOESS smooths by
treatment group') +
  theme_bw()
print(figure1)

## Figure 2: Plot correlation between time periods
cd4_dedupe <- cd4[!duplicated(cd4[,c('id', 'week_rounded')]),]
# Create wide table to plot and calculate correlations
cd4_wide <- cd4_dedupe[ ,c("id","group","logcd4","week_rounded")] %>%
  pivot_wider(names_from = week_rounded,
              values_from = logcd4)
# Create plot with correlations
figure2 <- ggpairs(cd4_wide, columns=3:8, aes(color=as.factor(group))) +
  labs(title='Correlation between log CD4 measurements by treatment group') +
  theme_bw()
print(figure2)

## Calculate stats to determine number of missing obs per individual
subj_summary <- cd4 %>%
                group_by(id,group) %>%
                summarise(n = n(),
                          unique = n_distinct(week_rounded),
                          first_week = min(week_rounded),
                          last_week = max(week_rounded))
subj_summary_no_bl <- subj_summary[subj_summary$first_week != 0,]
subj_summary_dupes <- subj_summary[subj_summary$n != subj_summary$unique,]
subj_summary_complete <- subj_summary[subj_summary$unique == 6,]
```

```
baseline_vals <- subset(cd4,week_rounded==0)
baseline_vals <-
baseline_vals[!duplicated(baseline_vals[,c('id','week_rounded')],fromLast = TRUE),]
baseline_vals <- merge(baseline_vals,subj_summary[,c("id","unique")],by="id")

## Table 3: Summary of baseline metrics by number of observations for each subject
table3 <- baseline_vals %>%
                group_by(as.factor(unique)) %>%
                summarise(n = n(),
                        n_men = sum(sex),
                        n_women = n()-sum(sex),
                        pct_1 = sum(ifelse(group==1,1,0))/n(),
                        pct_2 = sum(ifelse(group==2,1,0))/n(),
                        pct_3 = sum(ifelse(group==3,1,0))/n(),
                        pct_4 = sum(ifelse(group==4,1,0))/n(),
                        mean_age = mean(age, na.rm=T),
                        mean_cd4 = mean(logcd4, na.rm=T),
                        min_cd4 = min(logcd4, na.rm=T),
                        p25_cd4 = quantile(logcd4, 0.25, na.rm=T),
                        median_cd4 = quantile(logcd4, 0.5, na.rm=T),
                        p75_cd4 = quantile(logcd4, 0.75, na.rm=T),
                        max_cd4 = max(logcd4, na.rm=T))

## Perform gls - Model 1
mod1 <- gls(logcd4 ~ as.factor(group)*week+as.factor(group)*week_spline,
        method="REML",
        corr=corExp(form = ~ week | id, nugget=T),
        data=cd4, na.action = "na.omit")
summary(mod1)
intervals(mod1)
anova(mod1)
```

```
## Perform gls - Model 2
mod2 <- gls(logcd4 ~ age+sex+as.factor(group)*week+as.factor(group)*week_spline,
        method="REML",
        corr=corExp(form = ~ week | id, nugget=T),
        data=cd4, na.action = "na.omit")
summary(mod2)
intervals(mod2)
anova(mod2)
```