**Biost 540 Midterm**

**Yutong Wu, Kaiyue Yu, Zihan Zheng**

**Introduction:** A CD4 cell count below 200 is one of the diagnostic criteria for AIDS. The AIDS Clinical Trial Group (ACTG) study 193A was a randomized, double-blind study of 1309 AIDS patients with advanced immune suppression. Patients were randomized to one of the four daily regimens, i.e. zidovudine alternating monthly with group 1: 400 mg didanosine, group 2: zidovudine plus 2.25mg of zalcitabine, group 3: zidovudine plus 400mg of didanosine, and group 4: zidovudine plus 400mg of didanosine plus 400mg of nevirapine. The CD4 counts were scheduled to be measured at baseline and every 8 weeks in the 40-week follow-up period. However, the CD4 count data is unbalanced due to mistimed measurements, skipped measurements and loss to follow-up.

In this project, we will investigate the relative clinical efficacy of the four treatment regimens for HIV-infected patients with advanced disease via longitudinal measurements of subjects’ CD4 cell counts throughout the 40-week study period.

**Methods:** *Descriptive Analysis*: To examine the association between treatment regimen and log-CD4 cell counts over time, we first compare the baseline characteristics of subjects by treatment group, ruling out baseline age and gender as potential confounders **(Table 1)**. We also compared baseline age or gender composition between subjects with complete observations (visit times>=6) and subjects with missing data **(Table 5)**. After splitting the study period into five 8-week intervals, we generate summary statistics to compare the mean log-CD4 cell counts between the four treatment groups within each interval (**Table 2**). **Figure 1** displays the longitudinal change of log-CD4 cell counts throughout the 40 weeks for the four treatment groups side by side. Regarding gender and baseline age as possible precision variables, we compare the mean log-CD4 cell count in each of the 8-week time intervals by gender (1 for male and 0 for female) and baseline age group, respectively (**Table 3 and Table 4**). Besides, **Figure 2 and Figure 3** displays how the longitudinal change of log-CD4 cell counts differ by gender or baseline age category.

*Confirmatory Analysis:* To determine if treatment regimens affect changes in log-CD4 counts over time, we perform 3 potential linear mixed models (LMM) including group, week and their interaction with linear spline model for time. Using sensitivity analysis and AIC, We choose model 2: LMM with random intercepts and slopes, correlated, with linear spline (based on lowest value of AIC) (**Table 6**). Model 2 allows us to distinguish between between-treatment groups and within-treatment group sources of variability. Since our dataset is unbalanced with a large amount of missing values, model 2 is particularly helpful in modeling the covariance of the log-CD4 counts. We use ML estimation to obtain point estimates and standard errors for the regression parameters and simultaneously test if the interaction terms are all 0 using a likelihood ratio test. We use ML because ML and REML are similar when the number of independent samples is much larger than the dimension of β. In model 2, the estimated random effects of residual are and. The estimated random effects of residual suggests that there are small variability in log-CD4 cell counts across individuals (= 0.648), small variability in change of log-CD4 cell counts over time (= 0.0025), and the association between log-CD4 cell counts and rate of changes is negligible (= 0.0003). To access the normality of model 2, we plot residuals and only observe normality in random intercepts (**Figure 4**).

To determine if there is a difference in treatment regimens in terms of changes in log-CD4 counts over time controlling for gender and baseline age, we perform 3 potential LMM including age, sex, group, week and interaction between group and week with linear spline model for time. We perform similar analysis and choose mod5: LMM with random intercepts and slopes, correlated, with linear spline (**Table 8**). In model 5,the estimated random effects of residual are, which can be interpreted similarly as above**,** and . To access the normality of model 5, we plot residuals and only observe normality in random intercepts (**Figure 5**).

**Results:** *Descriptive Analysis*: Based on the descriptive statistics and graphics, we observe that the baseline log-CD4 cell counts are similar across all treatment groups, genders and age categories, respectively, at baseline (**Table 2, 3 and 4**). Patients in treatment group 4 tend to have moderately higher mean log-CD4 counts compared to other treatment groups in each time interval after the eighth week **(Table 2)**; patients in treatment group 1 (zidovudine alterna[‘;;;./lan log-CD4 counts than females, and this gender difference tends to become wider over time **(Table 3)**. Older patients tend to have slightly higher lower mean log-CD4 counts compared to younger patients across all the time intervals after the eighth week **(Table 4)**.

*Confirmatory Analysis:* There is strong evidence at the 0.05 level that the treatment groups differ in terms of changes in mean log-CD4 counts over time (LRT: p< 0.001). Point estimates and 95% CIs for the differences in mean log-CD4 counts between each of one week differences for each treatment regimens group are presented in **Table 7**.

In week 0-16, compared to group 1, we estimated the difference in difference in mean increase of log-CD4 counts is 0.003 (95% CI: -0.002, -0.007) higher in group 2, 0.007 (0.002, 0.011) higher in group 3, 0.016 (0.011, 0.021) higher in group 4. In the window of 16 - 40 week, compared to group 1, we estimate that the difference in difference in average decrease of log-CD4 counts is 0.018 (-0.028, -0.008) lower in group 2, 0.014 (-0.024, -0.004) lower in group 3, 0.005 (-0.015, 0.006) lower in group 4.

With controlling age and sex, there is statistically significant evidence at the 0.05 level

that the treatment regimens groups differ in terms of changes in mean log-CD4 counts

over time (LRT: p<0.001). Additionally, the estimated coefficients of weekSpline16 and

interaction terms stay the same after adjusted age and sex. However, the corresponding

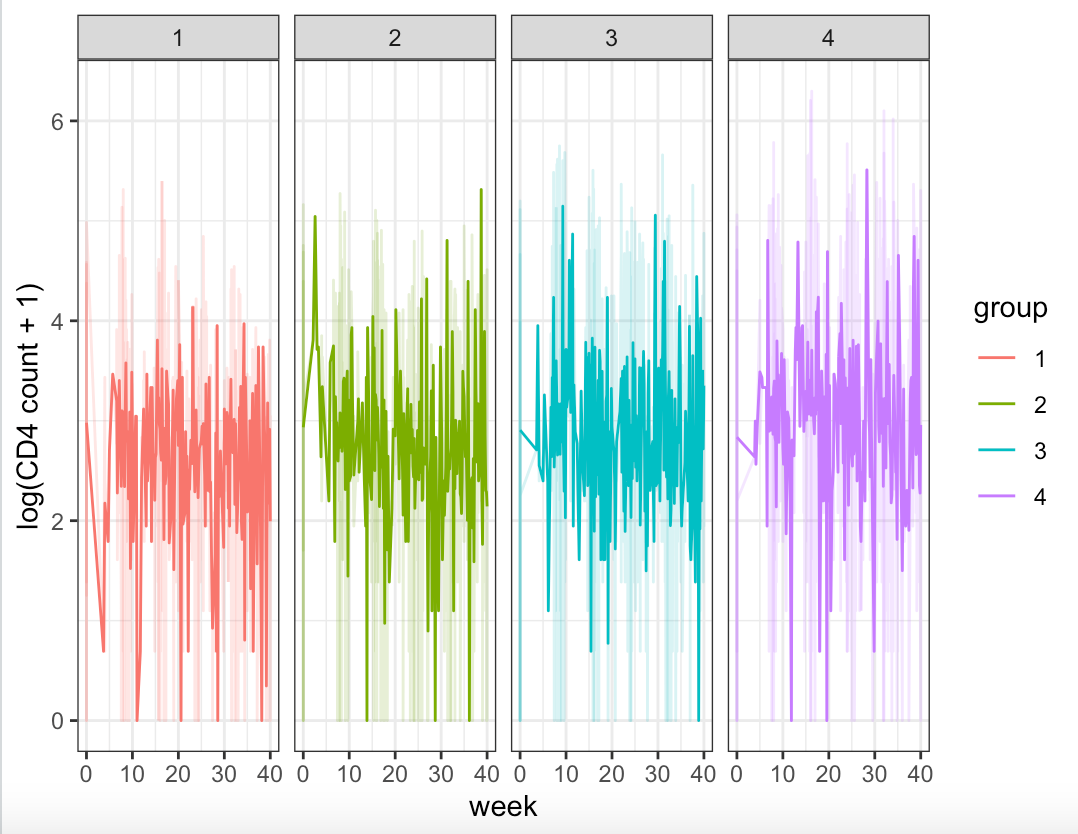
confidence intervals are slightly smaller after adjustment (**Table 9**).

In conclusion, Group 4 (AZT/ddI plus nevirapine) has relatively the highest effectiveness because it increases CD4 counts the most before week 16 and decreases CD4 counts the least for HIV-infected patients with advanced disease. On the contrary, Group 1 has relatively the lowest effectiveness. The results will remain the same regardless of adjusting age and sex.

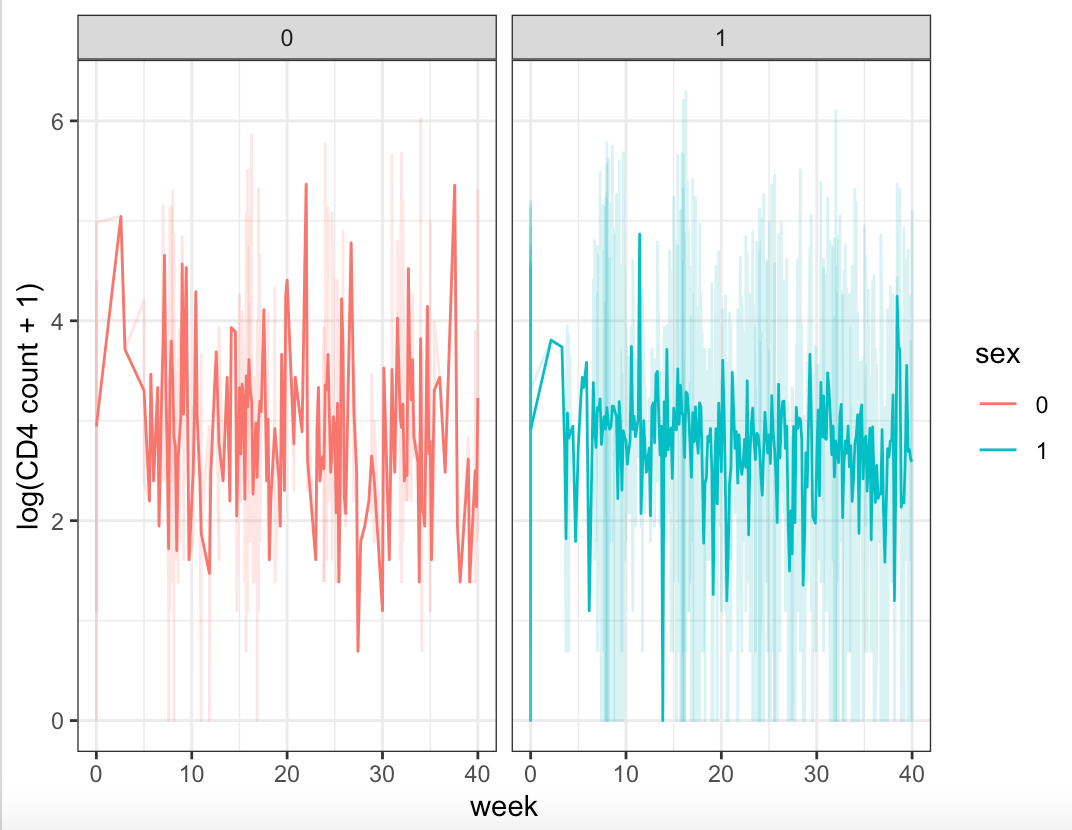
# Appendix A: Tables and Figures

|  | **% Male** | **% Female** | **Mean Age** | **Median Age** | **SD of Age** |
| --- | --- | --- | --- | --- | --- |
| **Group 1** | **89.7** | **10.3** | **37.7** | **36.6** | **8.4** |
| **Group 2** | **88.1** | **11.9** | **37.7** | **36.9** | **7.9** |
| **Group 3** | **89.1** | **10.9** | **37.4** | **37.1** | **7.9** |
| **Group 4** | **88.5** | **11.5** | **38.0** | **36.9** | **8.4** |

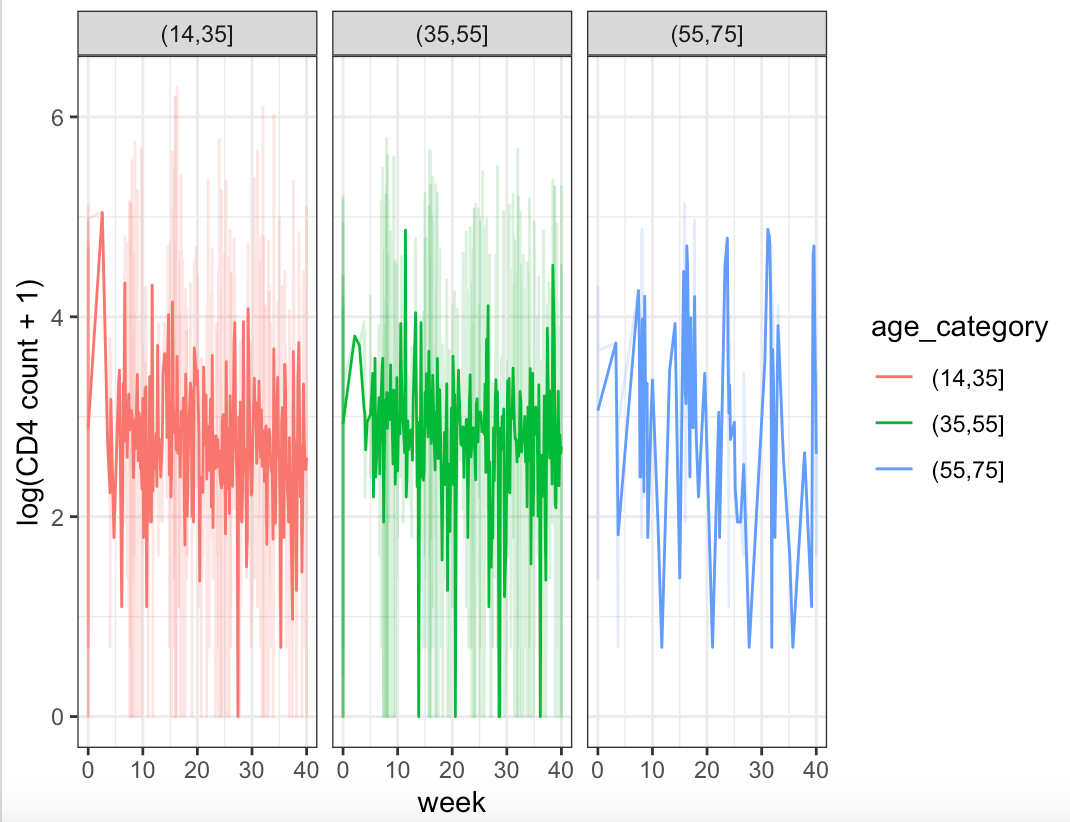
**Table 1 Summary of age and gender distribution by treatment group**

****

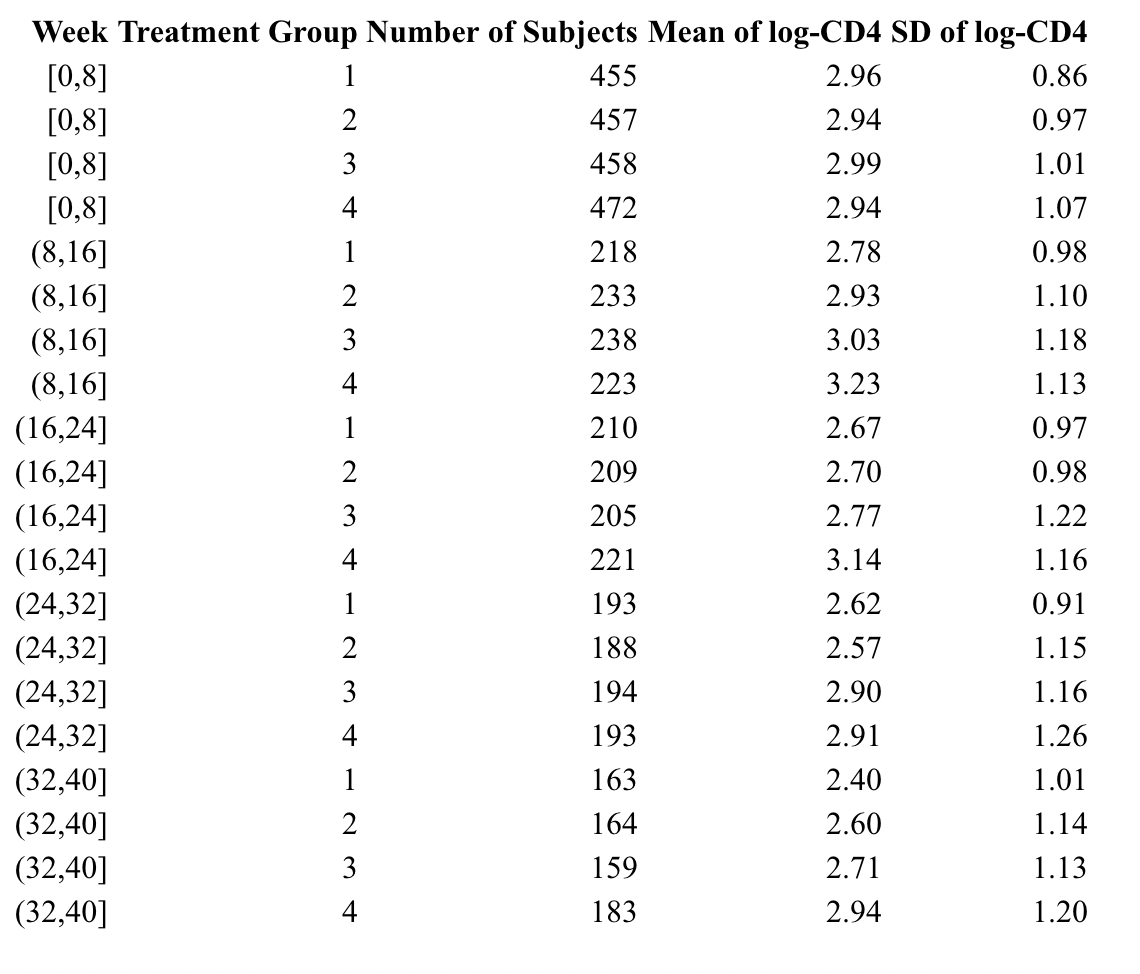
**Figure 1 Change of log-CD4 counts by treatment group**

****

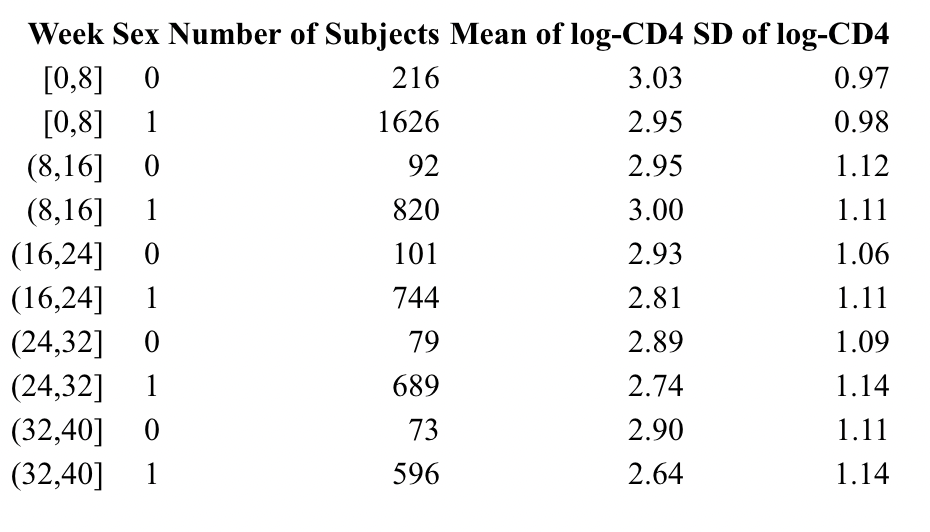
**Figure 2 Change of log-CD4 counts by Sex**

****

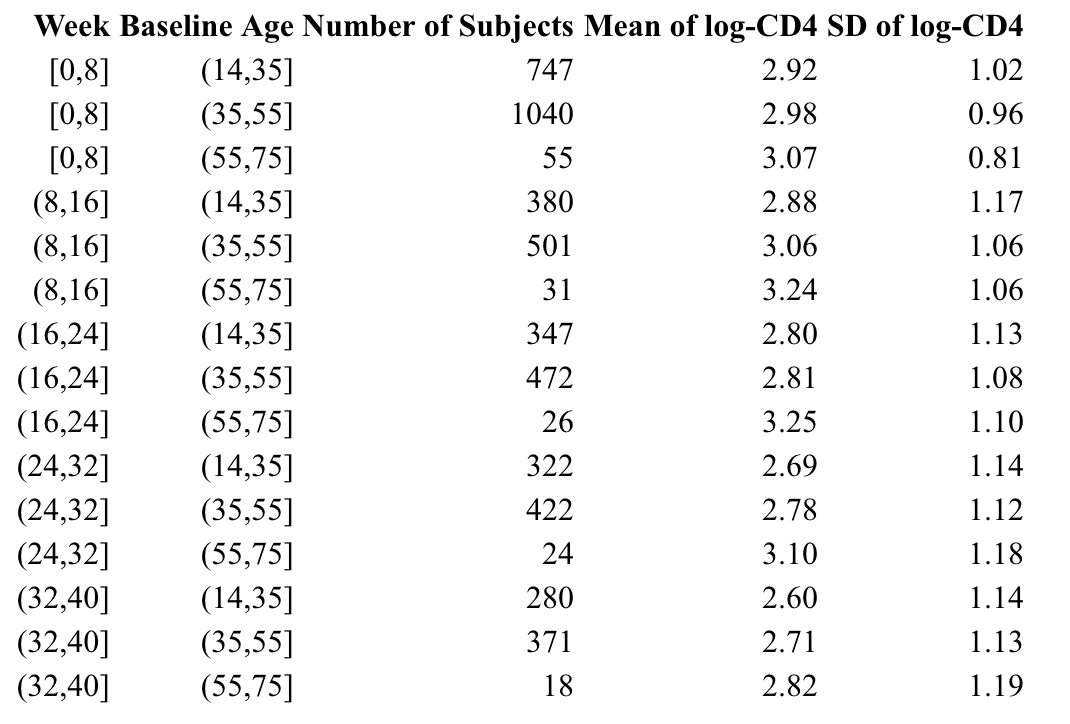
**Figure 3 Change of log-CD4 counts by age group**

****

**Table 2 Summary statistics by treatment group and time period**

****

**Table 3 Summary statistics by sex and time period**

****

**Table 4 Summary statistics by baseline and time period**

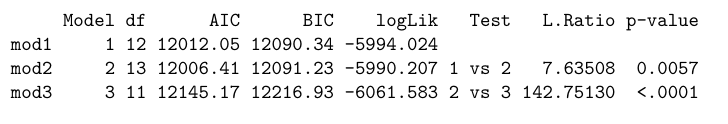
|  | **% Male** | **% Female** | **Mean Age** | **Median Age** | **SD of Age** |
| --- | --- | --- | --- | --- | --- |
| **Subject with complete data** | **91.8** | **8.2** | **38.0** | **37.5** | **8.1** |
| **Subject with missing data** | **86.9** | **13.1** | **37.7** | **36.8** | **8.2** |

**Table 5 Summary statistics of baseline characteristics comparing subjects with and without missing data**

Mod1: LMM with random intercepts and slopes, uncorrelated, with linear spline;

Mod2: LMM with random intercepts and slopes, correlated, with linear spline;

Mod3: LMM with random intercepts and slopes, uncorrelated, with cubic spline.

****

**Table 6 Anova for model 1 - 3 comparison**

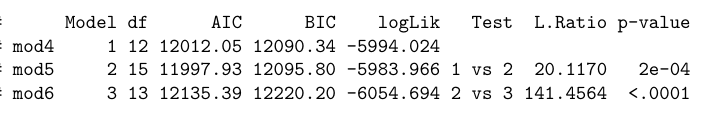
|  | **coefficients** | **lower CI** | **upper CI** | **p value** |
| --- | --- | --- | --- | --- |
| **(Intercept)** | **2.932** | **2.830** | **3.035** | **<0.0001** |
| **group2** | **0.008** | **-0.136** | **0.152** | **0.908** |
| **group3** | **0.005** | **-0.139** | **0.148** | **0.948** |
| **group4** | **0.017** | **-0.126** | **0.160** | **0.815** |
| **week** | **-0.006** | **-0.011** | **-0.002** | **0.004** |
| **weekSpline16** | **-0.021** | **-0.026** | **-0.015** | **<0.0001** |
| **group2:week** | **0.003** | **-0.002** | **0.007** | **0.304** |
| **group3:week** | **0.007** | **0.002** | **0.011** | **0.008** |
| **group4:week** | **0.016** | **0.011** | **0.021** | **<0.0001** |

**Table 7 Results from model 2**

Mod4: LMM with random intercepts and slopes, uncorrelated, with linear spline

Mod5: LMM with random intercepts and slopes, correlated, with linear spline

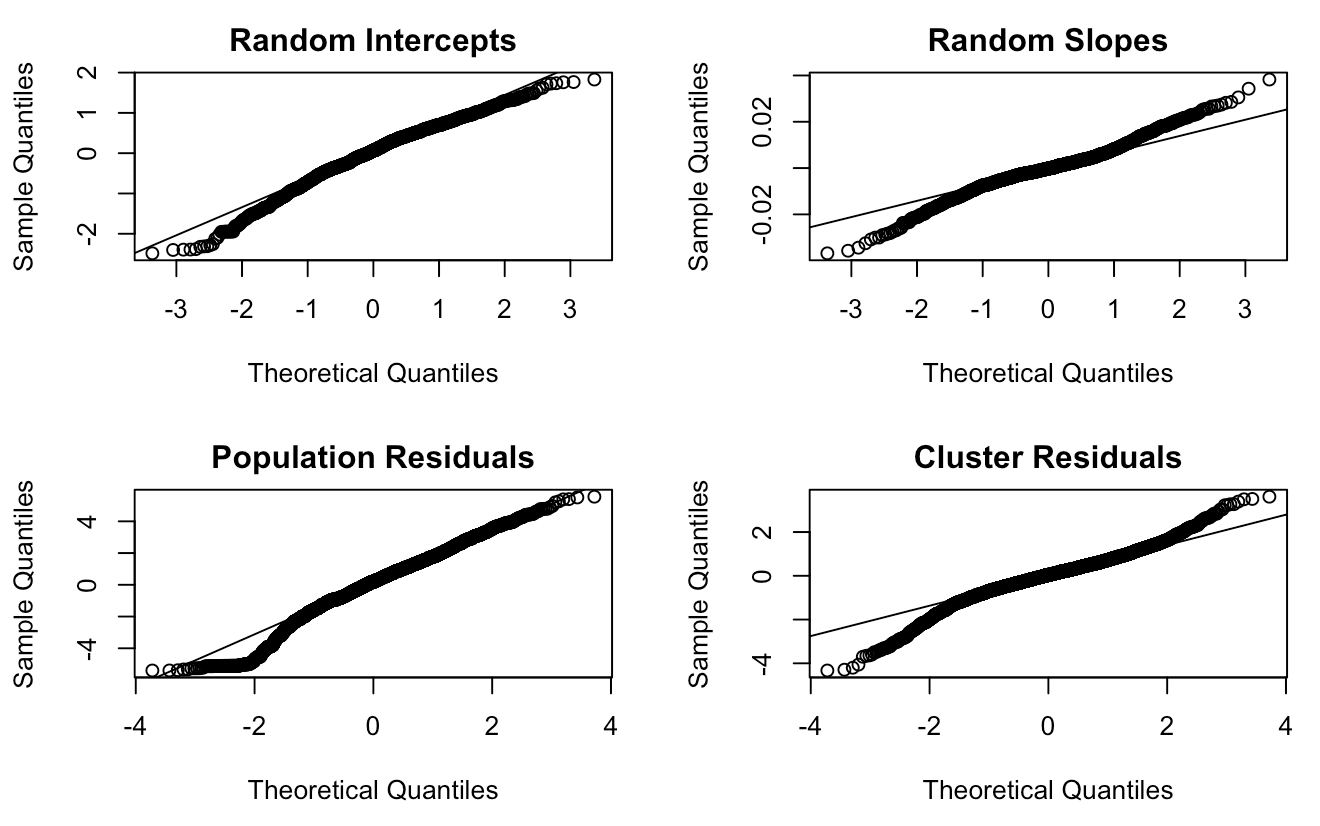
Mod6: LMM with random intercepts and linear spline

****

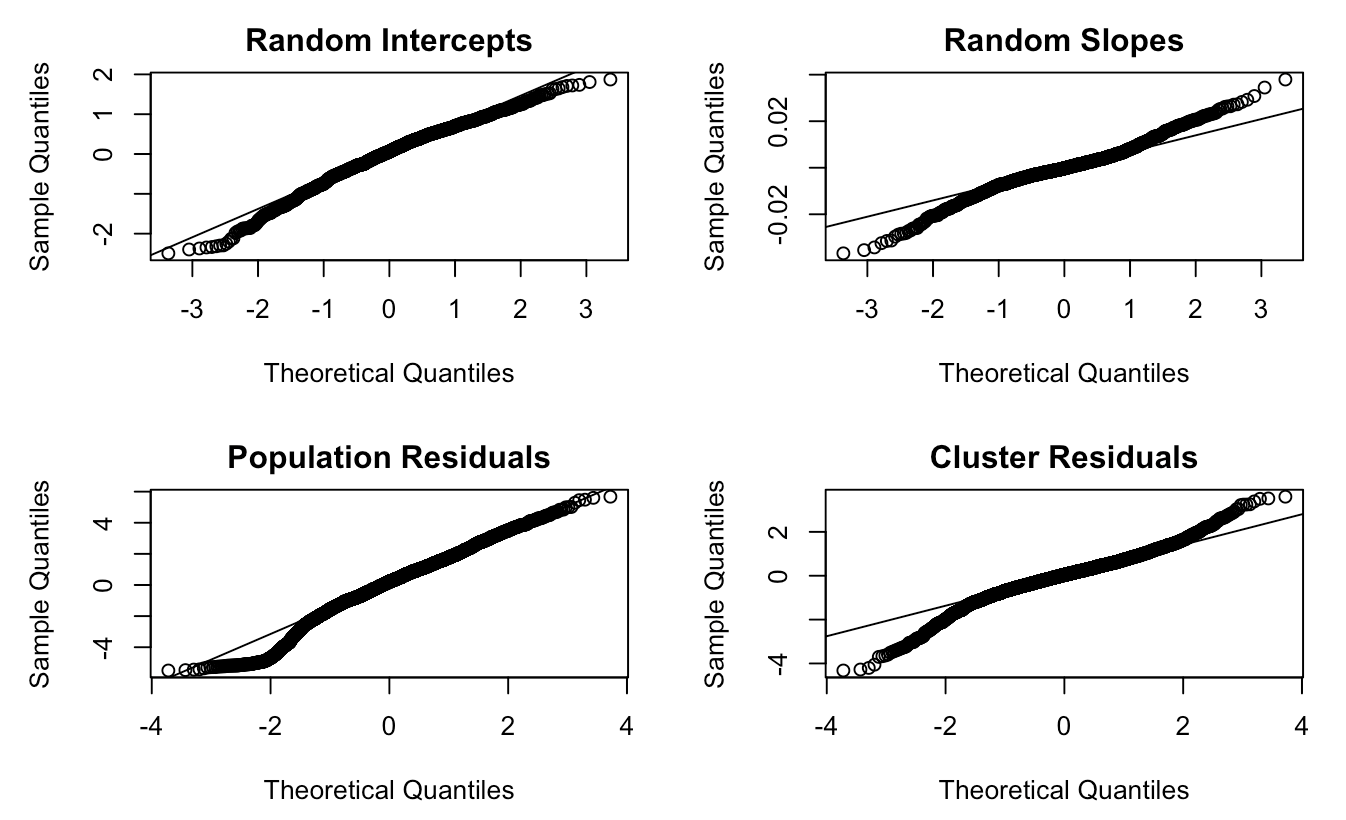
**Table 8 Anova for model 4 - 6 comparison**

|  | **coefficients** | **lower CI** | **upper CI** | **pvalue** |
| --- | --- | --- | --- | --- |
| **(Intercept)** | **2.598** | **2.325** | **2.870** | **<0.0001** |
| **age** | **0.011** | **0.005** | **0.017** | **0.001** |
| **sex1(male)** | **-0.079** | **-0.230** | **0.072** | **0.306** |
| **group2** | **0.008** | **-0.136** | **0.151** | **0.916** |
| **group3** | **0.007** | **-0.136** | **0.150** | **0.923** |
| **group4** | **0.015** | **-0.128** | **0.158** | **0.835** |
| **week** | **-0.006** | **-0.011** | **-0.002** | **0.004** |
| **weekSpline16** | **-0.021** | **-0.026** | **-0.016** | **<0.0001** |
| **group2:week** | **0.003** | **-0.002** | **0.007** | **0.308** |
| **group3:week** | **0.007** | **0.002** | **0.011** | **0.008** |
| **group4:week** | **0.016** | **0.011** | **0.021** | **<0.0001** |

**Table 9 Results from model 5**

****

**Figure 4 Residuals of model 2**

****

**Figure 5 Residuals of model 5**

# Appendex B: Code

library(ggplot2)

library(tidyverse)

library(nlme)

library(joineR)

library(MASS)

library(lattice)

library(dplyr)

cd4<-cd4[,-1]

cd4$age\_category<-cut(cd4$age, c(14,35,55,75))

cd4$period<-cut(cd4$week, c(0,8,16,24,32,40), include.lowest=TRUE)

#Compare baseline characteristics between treatment groups

with(cd4, table(group, sex))

group1<-subset(cd4, group==1)

mean(group1$age)

sd(group1$age)

median(group1$age)

group2<-subset(cd4, group==2)

mean(group2$age)

sd(group2$age)

median(group2$age)

group3<-subset(cd4, group==3)

mean(group3$age)

sd(group3$age)

median(group3$age)

group4<-subset(cd4, group==4)

mean(group4$age)

sd(group4$age)

median(group4$age)

#Compare by treatment groups

#Plot

cd4$group<-factor(cd4$group)

p<-ggplot(data=cd4, aes(x=week, y=logcd4, group=group, col=group))

p+geom\_line(alpha=0.2)+facet\_grid(.~group)+geom\_line(data=cd4%>%

group\_by(group, week)%>%summarise(logcd4=mean(logcd4)),

aes(x=week, y=logcd4, group=group))+theme\_bw()+ylab("log(CD4 count + 1)")

#Descriptive Statistics

stat<-cd4%>%group\_by(period, group)%>%summarise(number\_of\_subjects=sum(!is.na(logcd4))%>%round(2),

logcd4\_mean=mean(logcd4, na.rm=TRUE)%>%round(2), logcd4\_sd=sd(logcd4, na.rm=TRUE)%>%round(2))

knitr::kable(stat, col.names = c("Week","Treatment Group","Number of Subjects",

"Mean of log-CD4","SD of log-CD4"), align="rrrrr")

#Compare by baseline age

#Plot

p<-ggplot(data=cd4, aes(x=week, y=logcd4, group=age\_category, col=age\_category))

p+geom\_line(alpha=0.2)+facet\_grid(.~age\_category)+geom\_line(data=cd4%>%

group\_by(age\_category, week)%>%summarise(logcd4=mean(logcd4)),

aes(x=week, y=logcd4, group=age\_category))+theme\_bw()+ylab("log(CD4 count +1)")

#Descriptive Statistics

stat1<-cd4%>%group\_by(period, age\_category)%>%summarise(number\_of\_subjects=sum(!is.na(logcd4)),

logcd4\_mean=mean(logcd4, na.rm=TRUE)%>%round(2), logcd4\_sd=sd(logcd4, na.rm=TRUE)%>%round(2))

knitr::kable(stat1, col.names = c("Week","Baseline Age","Number of Subjects",

"Mean of log-CD4","SD of log-CD4"), align="rrrrr")

#Compare by sex

#Plot

cd4$sex<-factor(cd4$sex)

p<-ggplot(data=cd4, aes(x=week, y=logcd4, group=sex, col=sex))

p+geom\_line(alpha=0.2)+facet\_grid(.~sex)+geom\_line(data=cd4%>%

group\_by(sex, week)%>%summarise(logcd4=mean(logcd4)),

aes(x=week, y=logcd4, group=sex))+theme\_bw()+ylab("log(CD4 count + 1)")

#Descriptive Statistics

stat2<-cd4%>%group\_by(period, sex)%>%summarise(number\_of\_subjects=sum(!is.na(logcd4)),

logcd4\_mean=mean(logcd4, na.rm=TRUE)%>%round(2), logcd4\_sd=sd(logcd4, na.rm=TRUE)%>%round(2))

knitr::kable(stat2, col.names = c("Week","Sex","Number of Subjects",

"Mean of log-CD4","SD of log-CD4"), align="rrrrr")

#Missing Data

visit\_counts<-cd4%>%group\_by(id)%>%summarise(n=n(), nobs=sum(!is.na(logcd4)))

cd4$visit\_times<-visit\_counts$nobs[match(cd4$id, visit\_counts$id)]

cd4\_baseline<-cd4[,c("id", "sex", "age", "visit\_times")]

cd4\_baseline<-cd4\_baseline[!duplicated(cd4\_baseline),]

complete\_case<-cd4\_baseline[cd4\_baseline$visit\_times>=6,]

missing\_case<-cd4\_baseline[cd4\_baseline$visit\_times<6,]

table(complete\_case$sex)

table(missing\_case$sex)

summary(complete\_case$age)

summary(missing\_case$age)

**Part 2**

cd4.long <- read.csv("~/Downloads/cd4.csv")

cd4.long$sex <- relevel(factor(cd4.long$sex), ref=1) # making males the reference

cd4.long$group <- relevel(factor(cd4.long$group), ref=1) # make 1 = zidovudine alternating monthly with 400 mg didanosine the reference group

cd4.long$time <- as.numeric(cut(cd4.long$week, c(0,8,16,24,32,40), include.lowest=TRUE))

head(cd4.long)

### unadjusted models

# add knot at week 16

cd4.long$weekSpline16 <- (cd4.long$week - 16)\*(cd4.long$week > 16)

##### LMM Random intercepts + slopes uncorrelated with linear spline

mod1 <- lme( logcd4 ~ group\*week + weekSpline16,

method = "ML", data = cd4.long,

random = reStruct( ~ 1 + week | id, pdClass="pdDiag", REML=F))

summary(mod1)

##### LMM Random intercepts + slopes correlated with linear spline

mod2 <- lme( logcd4 ~ group\*week + weekSpline16,

method = "ML", data = cd4.long,

random = reStruct( ~ 1 + week | id, pdClass="pdSymm", REML=F))

mod2\_result<-summary(mod2)

mod2\_result

####### LMM Random intercepts with linear spline

mod3 <- lme( logcd4 ~ group\*week + weekSpline16,

method = "ML", data = cd4.long,

random = reStruct( ~ 1 | id, REML=F))

summary(mod3)

anova(mod1,mod2,mod3) # pick model2

interval1 <- intervals(mod2,which = "fixed")

#table of result

table1 <- data.frame(coefficients = round(mod2\_result$coefficients$fixed,3),

lower = round(interval1$fixed[,1],3),

upper = round(interval1$fixed[,3],3),

pvalue = round(mod2\_result$tTable[,5],3))

knitr::kable(table1, col.names = gsub("[.]"," ", names(table1)), align = "lccrr",caption = "Table4 Results from model2")

### Hypothesis Testing

reduced.mod2 <- lme( logcd4 ~ group + week + weekSpline16,

method = "ML", data = cd4.long,

random = reStruct( ~ 1 + week | id, pdClass="pdSymm", REML=F))

anova(reduced.mod2,mod2)

anova(mod2)

### adjusted models

# add knot at week 16

cd4.long$weekSpline16 <- (cd4.long$week - 16)\*(cd4.long$week > 16)

##### LMM Random intercepts + slopes uncorrelated with linear spline ######

mod4 <- lme( logcd4 ~ group\*week + weekSpline16,

method = "ML", data = cd4.long,

random = reStruct( ~ 1 + week | id, pdClass="pdDiag", REML=F))

summary(mod4)

##### LMM Random intercepts + slopes correlated with linear spline ######

mod5 <- lme( logcd4 ~ age + sex + group\*week + weekSpline16,

method = "ML", data = cd4.long,

random = reStruct( ~ 1 + week | id, pdClass="pdSymm", REML=F))

mod5\_result <- summary(mod5)

mod5\_result

####### LMM Random intercepts with linear spline

mod6 <- lme( logcd4 ~ age + sex + group\*week + weekSpline16,

method = "ML", data = cd4.long,

random = reStruct( ~ 1 | id, REML=F))

summary(mod6)

anova(mod4,mod5,mod6) # pick model5

interval2 <- intervals(mod5,which = "fixed")

#table of results

table2 <- data.frame(coefficients = round(mod5\_result$coefficients$fixed,3),

lower = round(interval2$fixed[,1],3),

upper = round(interval2$fixed[,3],3),

pvalue = round(mod5\_result$tTable[,5],3))

knitr::kable(table2, col.names = gsub("[.]", " ", names(table2)), align = "lccrr",caption = "Table5 Results from model5")

### Hypothesis Testing

reduced.mod5 <- lme(logcd4 ~ age + sex + group+week + weekSpline16,

method = "ML", data = cd4.long,

random = reStruct( ~ 1 + week | id, pdClass="pdSymm", REML=F))

anova(reduced.mod5,mod5)

anova(mod5)

#mod2

cluster.res <- resid(mod2, level=1, type="normalized")

pop.res <- resid(mod2, level=0, type="normalized")

bi <- mod2$coefficients$random$id

plot(cd4.long$week, cluster.res, xlab="week")

par(mfrow=c(2,2))

qqnorm(bi[,1], main="Random Intercepts")

qqline(bi[,1])

qqnorm(bi[,2], main="Random Slopes")

qqline(bi[,2])

qqnorm(pop.res, main="Population Residuals")

qqline(pop.res)

qqnorm(cluster.res, main="Cluster Residuals")

qqline(cluster.res)

#mod5

cluster.res <- resid(mod5, level=1, type="normalized")

pop.res <- resid(mod5, level=0, type="normalized")

bi <- mod5$coefficients$random$id

plot(cd4.long$week, cluster.res, xlab="week")

par(mfrow=c(2,2))

qqnorm(bi[,1], main="Random Intercepts")

qqline(bi[,1])

qqnorm(bi[,2], main="Random Slopes")

qqline(bi[,2])

qqnorm(pop.res, main="Population Residuals")

qqline(pop.res)

qqnorm(cluster.res, main="Cluster Residuals")

qqline(cluster.res)