Introduction: The ACL study is a longitudinal study about differences in life between African Americans and White Americans covering aspects such as sociological, psychological, and physical health items. Five waves of data were collected in 1986, 1989, 1994, 2002, and 2011 respectively, to explore how these aspects varied in participants. In this project, our goal is to assess the trends in the odds of functional impairment over time and determine how this effect varies across racial groups. We selected a total of 1586 participants with baseline age ranging from 25 to 50 and chose baseline age, sex, socioeconomic status, and functional impairment indicator for each wave as our variables of interest.

Methods: *Descriptive Analysis:* Since this is not a randomized study, we first need an overview of the distribution of each variable. We are interested in the differences between racial groups, so demographic features, and the proportion of those with functional impairment of participants are displayed by racial groups in Table 1. The number of missing values will also be shown in Table 1 to understand whether participants dropped out or came back after some missed follow-ups. We also want to know if a certain group of people are more likely to miss follow-ups, so Figure 1 will summarize the total count of participants with or without functional impairment categorized by those with and without missing data at each wave for both racial groups.

Confirmatory Analysis: For the primary analysis, we want to answer three questions. The first one is using all available data to analyze if there are changes in odds of functional impairment over time vary by racial group, adjusting for baseline age, sex, socioeconomic status, and past functional impairment status. Since we are more interested in the difference between racial groups rather than for any individuals in any group, we will use a GEE model with year (numerical) as the time variable and past functional impairment status defined as the functional impairment status recorded in the previous wave (missing if the participant missed the previous follow-up). We will select the independence working matrix for this model. The final model for this question is presented as:

$$\text{logit}\left(\text{E}\left[Y_{ij}\middle|X_{ij}, Y_{ij-1}\right]\right) = \beta_0 + \beta_1 1_{Race=W} + \beta_2 year + \beta_3 1_{Race=W} \times year + \beta_4 age + \beta_5 1_{Sex=Male} + \beta_6 1_{SES=Middle} + \beta_7 1_{SES=Upper} + \beta_8 1_{pastFI=1}$$

Wald test is used to assess whether the coefficient of interaction term $\beta_3 = 0$ to answer the first question. Using the same model, we want to answer the second question that if past functional impairment status is a significant predictor of current functional impairment status. We again will use a Wald test to assess whether the coefficient of past functional impairment status $\beta_8 = 0$

to answer this question. The third question is to examine differences in trends in functional impairment over time by racial group, adjusting for baseline age, sex, and socioeconomic status. We assume the data are missing completely at random, so we will fit a GEE model with an independence working matrix, but instead, use waves (categorical) as the time variable. The model for this question is presented as:

$$\begin{aligned}\text{logit}\left(\mathrm{E}\left[Y_{ij}\middle|X_{ij}\right]\right) = {} & \beta_0 + \beta_1 1_{Race=W} + \beta_2 1_{Wave=2} + \beta_3 1_{Wave=3} + \beta_4 1_{Wave=4} + \beta_5 1_{Wave=5} \\ & + \beta_6 1_{Race=W} \times 1_{Wave=2} + \beta_7 1_{Race=W} \times 1_{Wave=3} + \beta_8 1_{Race=W} \times 1_{Wave=4} \\ & + \beta_9 1_{Race=W} \times 1_{Wave=5} + \beta_{10} age + \beta_{11} 1_{Sex=Male} + \beta_{12} 1_{SES=Middle} + \beta_{13} 1_{SES=Upper}\end{aligned}$$

A Wald test will be conducted to test if the coefficient $\beta_1 = 0$ to answer whether there is a difference in functional impairment between two races at baseline, and another Wald test to test if the coefficients $\beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$ to answer whether there is a difference in the rate of change in functional impairment over each wave. Since GEE does not work very well with missing at random data, so there might be some bias with the results for the last model. Thus, we will conduct a sensitivity analysis that utilizes multiple imputations using method "2l.bin" to predict missing data and fit the same GEE model on the imputed data to re-assess the third question.

Results: *Descriptive Analysis:* The results shown in Table 1 indicates that the proportion of functional impairment increases at every wave for both groups, and African American have overall higher proportion. However, we cannot simply conclude the difference in functional impairment results from racial difference because the demographic features of these two groups are very different especially in the structure of socioeconomic status and sex distribution. There are more White Americans that have upper socioeconomic, and there are more female African American participants. Also, according to Table 1 and Figure 1, it seems like those with no functional impairment are more likely to come to every follow-up, but the difference is not obvious, and we know some participants came back after missing some follow-ups.

Confirmatory Analysis: The coefficients for the first proposed GEE model, with year as time variable and including past functional impairment, is displayed in Table 2. The odds of functional impairment over time are not statistically different between two racial groups (P = 0.62). The estimated odds of functional impairment is 8% higher (95% CI: 6% higher to 11% higher) for African Americans and 8% higher (95% CI: 4% higher to 12% higher) for White Americans for each additional year. From the same table, we can see that the odds of functional

impairment is 14.88 times higher (95% CI: 10.49, 21.16) in those with past functional impairment, meaning that past functional impairment, or the functional impairment status of the latest previous follow-up, is a statistically significant predictor ($P < 0.001$) of current functional impairment status. Table 3 shows the coefficients for the second proposed GEE model with wave as the time variable. The estimated coefficient for the race variable indicates that the odds of functional impairment for White Americans is 13% higher (95% CI: 38% lower to 107% higher) compared to African American at baseline (wave 1), and this difference is not statistically significant between two racial groups ($P = 0.56$). The rest estimated coefficients demonstrate the rate of change in functional impairment for each wave comparing the two racial groups, and by conducting a Wald test, we have strong evidence ($P = 0.012$) against the null hypothesis that there is no difference in the rate of change over time. For the sensitivity test, missing data were predicted using multiple imputations, and results were yielded as in Table 4. The pooled estimated coefficient for the race variable indicates that the odds of functional impairment for White Americans is 14.6% higher (95% CI: 24.7% lower to 74.3% higher) compared to African Americans at baseline (wave 1), and this difference is not statistically significant between two racial groups ($P = 0.53$), which agrees with the primary analysis. Since this is a pooled result rather than a single model, we cannot conduct the same Wald test as in the primary analysis, but by looking at the P values for the estimated coefficients of the interaction terms, there is obviously strong evidence ($P < 0.01$) against the null hypothesis that there is no difference in the rate of change over time.

Discussion: Overall, the odds of functional impairment is about 8% higher for both racial group for each additional year, and past functional impairment status is a very important predictor of current status. Furthermore, though there are no difference in functional impairment at baseline, there are significant differences in the rate of change in functional impairment over time comparing the two groups, and this holds for both using only available data and multiple imputation data. Some limitations are that we cannot determine the missing pattern as shown in Figure 1 directly, meaning that the data might be neither missing completely at random nor missing at random, thus rendering biased results. Second is we did not conduct a rigorous Wald test in the sensitivity analysis, so we do not know the exact p-value for testing the null hypothesis. Also, though the working matrix did not affect the validity of the GEE models, we did not check the true correlation to select the most efficient matrix.

Tables and Figures

Table 1: Demographic features of participants and proportion of functional impairment at each wave by racial group and overall.

	AA (N=546)	W (N=1040)	Overall (N=1586)
Baseline age			
Mean (SD)	36.2 (7.27)	35.8 (7.12)	35.9 (7.17)
Sex			
Female	354 (64.8%)	562 (54.0%)	916 (57.8%)
Male	192 (35.2%)	478 (46.0%)	670 (42.2%)
Socioeconomic status			
Low	125 (22.9%)	126 (12.1%)	251 (15.8%)
Middle	222 (40.7%)	272 (26.2%)	494 (31.1%)
Upper	199 (36.4%)	642 (61.7%)	841 (53.0%)
Wave 1 (1986)			
Mean (SD)	0.0751 (0.264)	0.0615 (0.240)	0.0662 (0.249)
Wave 2 (1989)			
Mean (SD)	0.124 (0.330)	0.0750 (0.264)	0.0909 (0.288)
Missing	128 (23.4%)	160 (15.4%)	288 (18.2%)
Wave 3 (1994)			
Mean (SD)	0.225 (0.418)	0.101 (0.302)	0.140 (0.347)
Missing	146 (26.7%)	152 (14.6%)	298 (18.8%)
Wave 4 (2002)			
Mean (SD)	0.345 (0.476)	0.168 (0.375)	0.220 (0.414)
Missing	207 (37.9%)	215 (20.7%)	422 (26.6%)
Wave 5 (2011)			
Mean (SD)	0.529 (0.500)	0.318 (0.466)	0.389 (0.488)
Missing	102 (18.7%)	151 (14.5%)	253 (16.0%)

Figure 1: Total count of participants with or without functional impairment categorized by those with and without missing data at each wave for both racial groups.

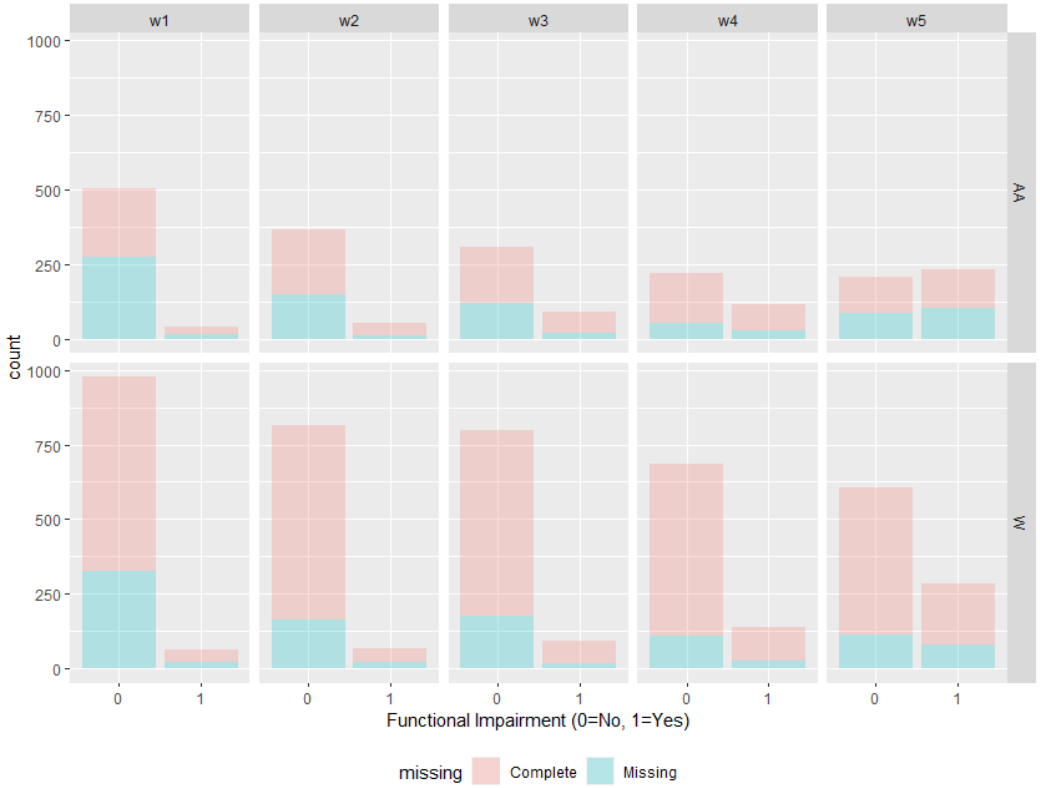

Table 2: Exponentiated point estimates (95% CI) for the GEE model (year as numeric time variable and includes past functional impairment) of race, year, their interaction term, and past functional impairment variables.

Variable	Exponentiated Estimates (95% CI)	P-Value
$1_{Race=W}$	2.68e+04 (4.23e-21, 1.70e+29)	0.64
$year$	1.08 (1.06, 1.11)	< 0.001
$1_{Race=W} \times year$	0.995 (0.967, 1.02)	0.62
$1_{pastFI=1}$	14.88 (10.49, 21.16)	< 0.001

Table 3: Exponentiated point estimates (95% CI) for the GEE model (wave as categorical time variable) of race and race-wave interaction term variables.

Variable	Exponentiated Estimates (95% CI)	P-Value
$1_{Race=W}$	1.13 (0.62, 2.07)	0.56
$1_{Race=W} \times 1_{Wave=2}$	0.759 (0.380, 1.515)	0.26
$1_{Race=W} \times 1_{Wave=3}$	0.490 (0.250, 0.959)	0.0029
$1_{Race=W} \times 1_{Wave=4}$	0.486 (0.250, 0.947)	0.0024
$1_{Race=W} \times 1_{Wave=5}$	0.491 (0.255, 0.946)	0.0024

Table 4: Pooled exponentiated point estimates (95% CI) for the GEE model after multiple imputations (wave as categorical time variable) of race and race-wave interaction term variables.

Variable	Exponentiated Estimates (95% CI)	P-Value
$1_{Race=W}$	1.146 (0.753, 1.743)	0.53
$1_{Race=W} \times 1_{Wave=2}$	0.730 (0.446, 1.195)	0.21
$1_{Race=W} \times 1_{Wave=3}$	0.519 (0.322, 0.839)	0.0075
$1_{Race=W} \times 1_{Wave=4}$	0.526 (0.326, 0.846)	0.0082
$1_{Race=W} \times 1_{Wave=5}$	0.494 (0.312, 0.784)	0.0027

Code Appendix

```
setwd("C:/Users/second/Desktop/BIOST 540/Final")

library(reshape2)
library(dplyr)
library(VIM)
library(ggplot2)
library(wgeesel)
library(geepack)
library(multcomp)
library(doBy)
library(tidyverse)

### Read data
acl <- read.csv("acl_subset.csv")

### Turn into long format
dat <- acl[,-1]
dat.long <- melt(dat, id=c("id", "sex", "race", "ses", "age"))
dat.long$year <- 1986
dat.long$year[dat.long$variable=="w2"] <- 1989
dat.long$year[dat.long$variable=="w3"] <- 1994
dat.long$year[dat.long$variable=="w4"] <- 2002
dat.long$year[dat.long$variable=="w5"] <- 2011

dat.long <- dat.long %>% arrange(id,year)

### Exploratory analysis
# Demographic features
library(table1)
label(dat$age) <- "Baseline age"
label(dat$sex) <- "Sex"
label(dat$ses) <- "Socioeconomic status"
label(dat$w1) <- "Wave 1 (1986)"
label(dat$w2) <- "Wave 2 (1989)"
label(dat$w3) <- "Wave 3 (1994)"
label(dat$w4) <- "Wave 4 (2002)"
label(dat$w5) <- "Wave 5 (2011)"
dat$race <- factor(dat$race, levels = c("AA","W"),
                           labels = c("African American",
                                      "White American"))

table1(~ age + sex + ses + w1 + w2 + w3 + w4 + w5| race, data=dat,
       render.continuous=c(.="Mean (SD)"))


# Missing pattern
```

```
ids.miss <- unique(dat.long$id[is.na(dat.long$value)])
dat.long$missing <- "Complete"
dat.long$missing[dat.long$id %in% ids.miss] <- "Missing"

ggplot(dat.long %>% filter(!is.na(value)),
       aes(x = as.factor(value), fill = missing)) +
  geom_bar(alpha =0.25) +
  facet_grid(race ~ variable) +
  xlab("Functional Impairment (0=No, 1=Yes)") +
  theme(legend.position="bottom")

### Q2
dat.long$lag1y <- ylag(dat.long$id,dat.long$value,1)

# Remove ids with no ylag value
dat.long.avail <- dat.long %>%
  filter(!is.na(lag1y))

# Model
mod_avail <- geeglm(value ~ race*year + age + sex + ses + lag1y, id = id,
                      data = dat.long.avail,
                      family=binomial(link="logit"))
summary(mod_avail)
inf <- glht(mod_avail)
mod_avail_ci <- confint(inf)[[9]] # 95% CI for all covariates
print(exp(mod_avail_ci))

# 95% CI for white american
lambda2 <- c(0, 0, 1, 0, 0, 0, 0, 0, 1)
exp(lambda2 %*% mod_avail$coefficients +
      qnorm(c(0.025, 0.5, 0.975)) *
c(summary(mod_avail)$coefficients$Std.err %*% lambda2))

### Q3
# Primary
mod_time <- geeglm(value ~ race*variable + age + sex + ses, id = id,
                    data = dat.long,
                    family=binomial(link="logit"))
summary(mod_time)
inf2 <- glht(mod_time)
mod_time_ci <- confint(inf2)[[9]]
print(exp(mod_time_ci))


l1 <- c(0,0,0,0,0,0,0,0,0,0,1,0,0,0)
l2 <- c(0,0,0,0,0,0,0,0,0,0,0,1,0,0)
```

```
l3 <- c(0,0,0,0,0,0,0,0,0,0,0,1,0)
l4 <- c(0,0,0,0,0,0,0,0,0,0,0,0,1)

esticon(mod_time, rbind(l1,l2,l3,l4), joint.test = T)

# Sensitivity
library(mice)
library(lme4)
dat.sens <- dat.long[,c(1:7)]

pred_long <- make.predictorMatrix(dat.sens)
pred_long["value","id"] <- -2
imp_long <- mice(dat.sens, method = "2l.bin",
                 pred = pred_long, seed = 540,
                 maxit = 1, m = 20, print=F)
densityplot(imp_long)

fit.imp_long <- with(imp_long,
                     geeglm(value ~ race*variable + age + sex + ses, id =
id,
                            family=binomial(link="logit")))

imp_mod <- summary(pool(fit.imp_long))

l_race <- c(0,1,0,0,0,0,0,0,0,0,0,0,0,0)
exp(l_race %*% imp_mod$estimate +
      qnorm(c(0.025, 0.5, 0.975)) * c(imp_mod$std.err %*% l_race))

exp(l1 %*% imp_mod$estimate +
      qnorm(c(0.025, 0.5, 0.975)) * c(imp_mod$std.err %*% l1))
exp(l2 %*% imp_mod$estimate +
      qnorm(c(0.025, 0.5, 0.975)) * c(imp_mod$std.err %*% l2))
exp(l3 %*% imp_mod$estimate +
      qnorm(c(0.025, 0.5, 0.975)) * c(imp_mod$std.err %*% l3))
exp(l4 %*% imp_mod$estimate +
      qnorm(c(0.025, 0.5, 0.975)) * c(imp_mod$std.err %*% l4))
```