BIOST 540 Homework 4

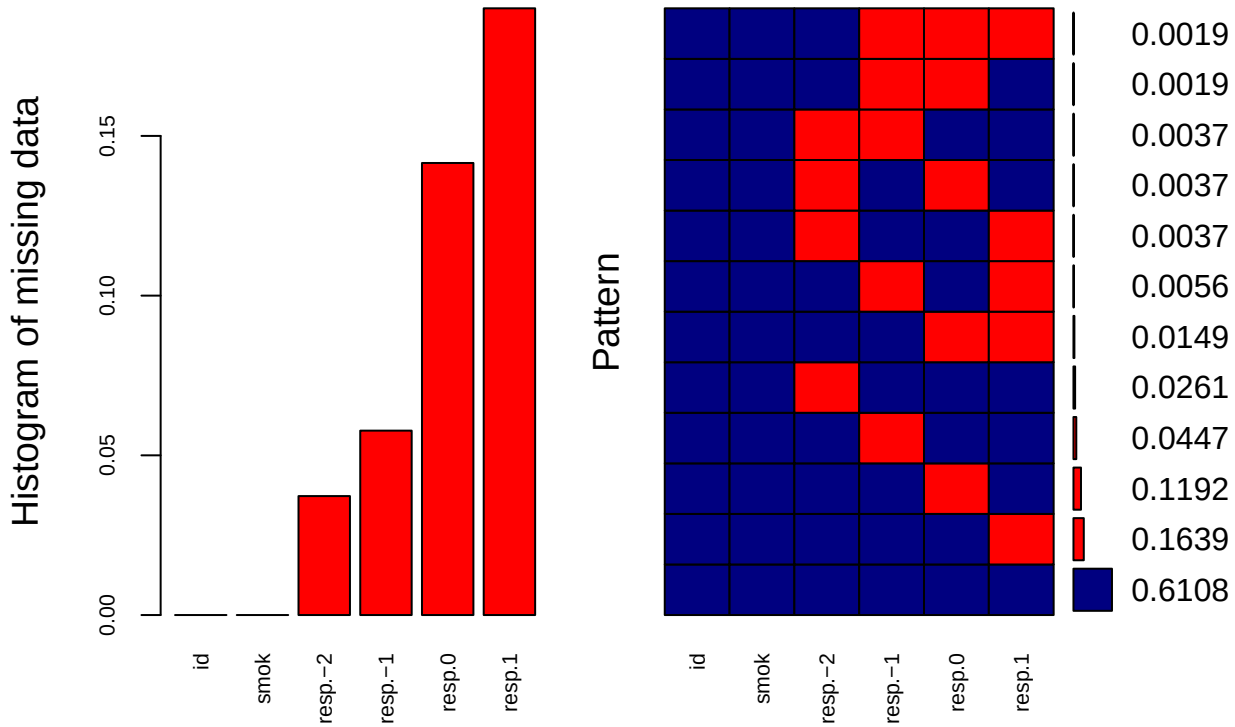
Ivy Zhang, Hantong Hu

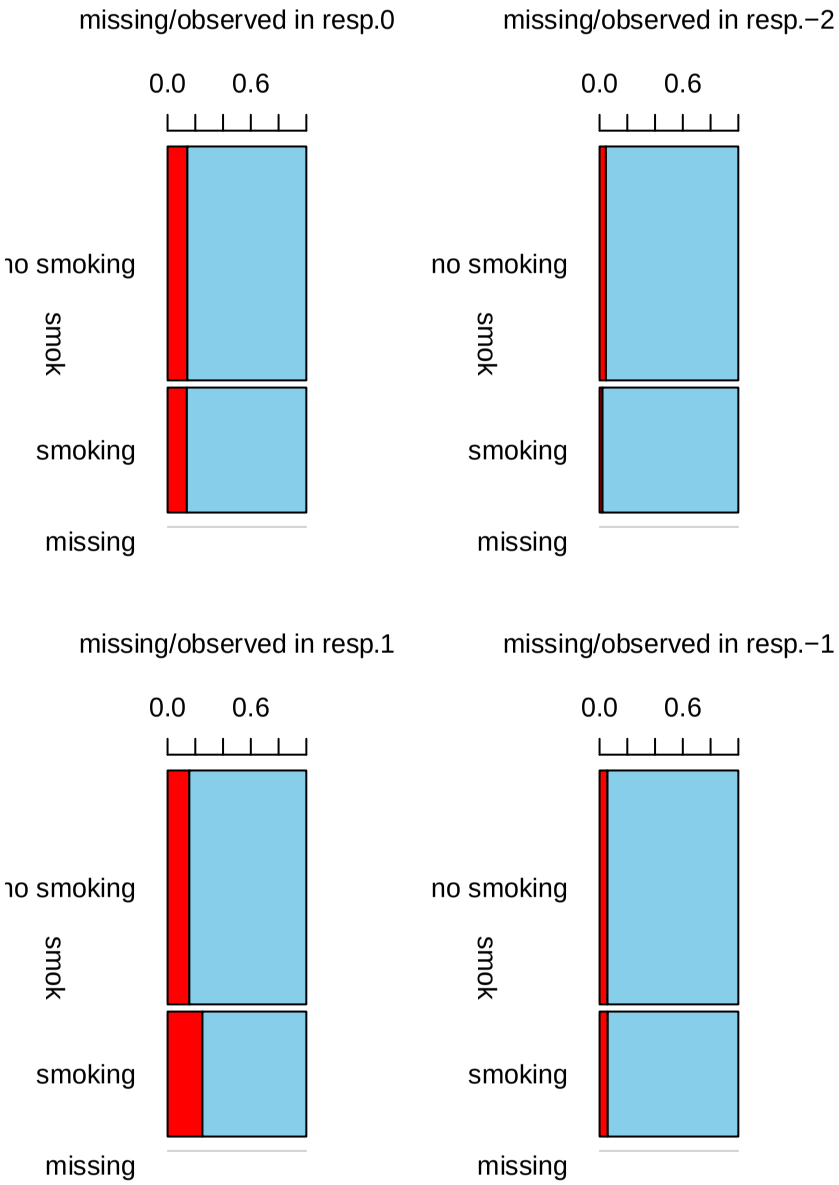
05/31/2021

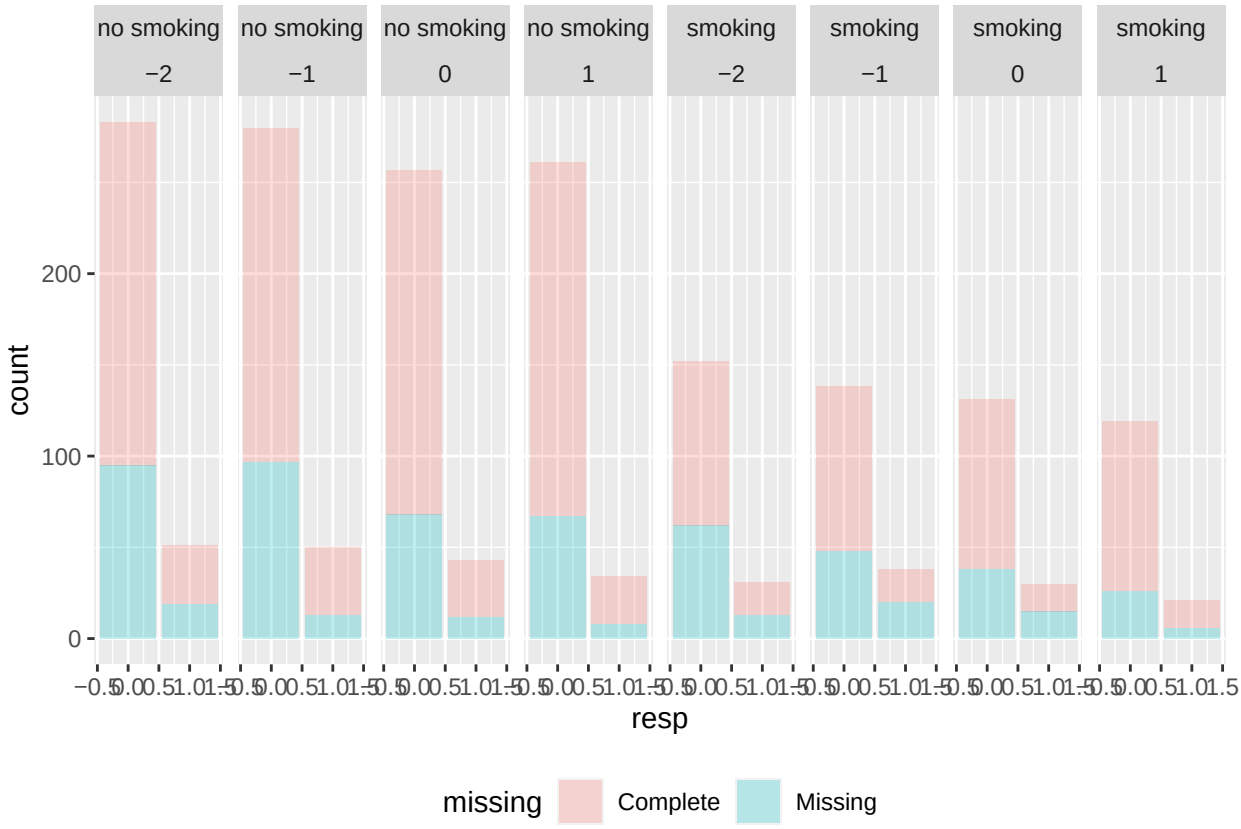
Responses

Part A

1. There is a total of 229 missing data in the datasets. Based on the following plots, we can see the smoking group tends to have a larger missing proportion compared to the non-smoking group except in the baseline. The later the time is, the higher proportion of the missing data in that time point. Most of people are missing only one time, and the most usual situation of missing data is the participants are only missing the data at age-9 time point. Based on the last graph, we can see the respiratory disease group seems to have a higher proportion of missing data than the no-disease group.




missing/observed in resp.0
0.0
0.6
no smoking
smok
smoking
missing
missing/observed in resp.−2
0.0
0.6
no smoking
smok
smoking
missing
missing/observed in resp.1
0.0
0.6
no smoking
smok
smoking
missing
missing/observed in resp.−1
0.0
0.6
no smoking
smok
smoking
missing



2. For 2), 3), 4), we are fitting the following model:

$$Logit(E(\text{resp}_{ij}|\text{smok}_{ij}, \text{age}_{ij}, b_{0i})) = \beta_0 + \beta_1 \text{smok}_{ij} + \beta_2 \text{age}_{ij} + b_{0i}$$

The tables show the exponentiated estimated coefficients and their 95% CI.

Table 1: Estimated Coefficients of Conditional Model in Complete Data

	Estimated Coefficients	Lower Bound of 95% CI	Upper Bound of 95% CI
(Intercept)	0.045	0.028	0.067
smoksmoking	1.490	0.870	2.563
age	0.839	0.734	0.957

3.

Table 2: Estimated Coefficients of Conditional Model in Missing Data

	Estimated Coefficients	Lower Bound of 95%CI	Upper Bound of 95%CI
(Intercept)	0.047	0.029	0.072
smoksmoking	1.604	0.930	2.788
age	0.891	0.773	1.026

4. For the imputations using the data in wide format and the method "logreg", the convergence assessment is as following. To assess the convergence, at all time points, the mean of the imputed datasets are very similar. The standard deviation is also very similar in the 25 imputed datasets at 7, 8, 9 years old time point, but there is an outlier in the 6-years old time point at the iteration of 6.

Table 3: Estimated Coefficients of Imputations Using the Data in Long Format

	Estimated Coefficients	Lower Bound of 95%CI	Upper Bound of 95%CI
Intercept	0.054	0.033	0.086
smoking	1.521	0.924	2.504
age	0.851	0.724	0.999

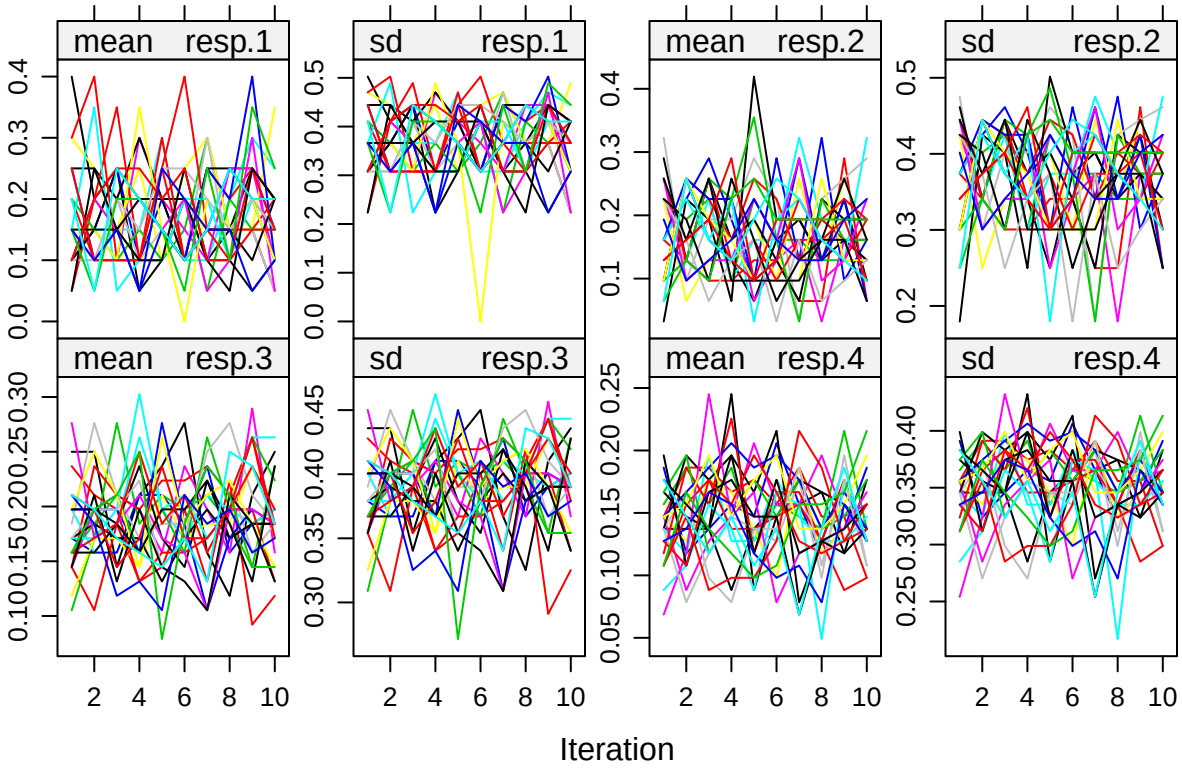


Table 4: Estimated Coefficients of Imputations Using the Data in Wide Format

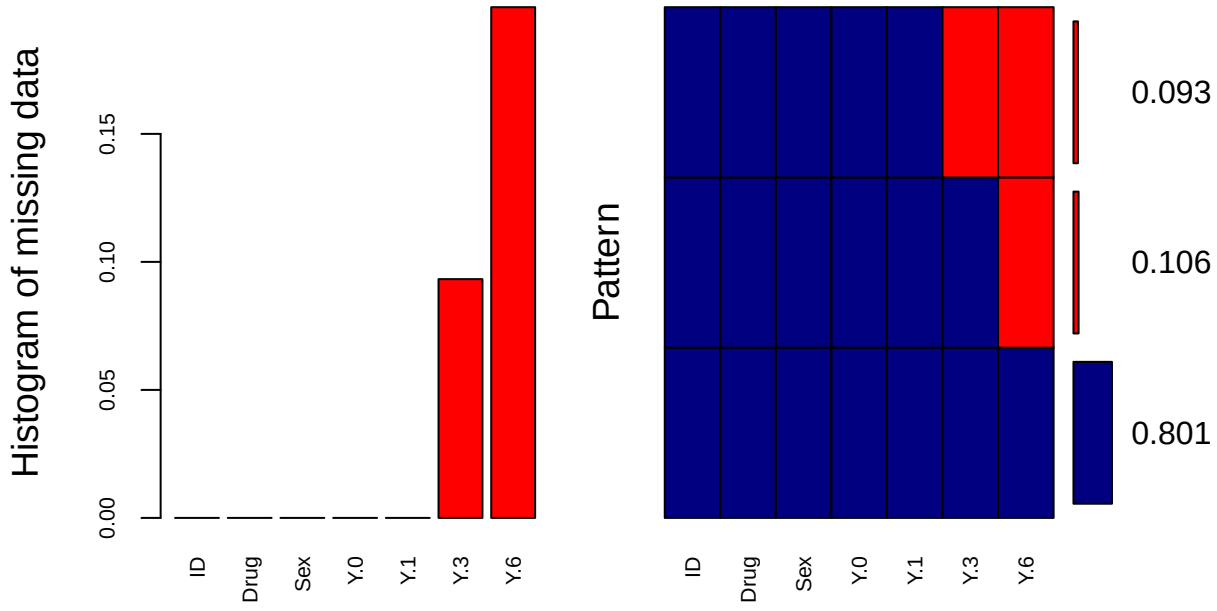
	Estimated Coefficients	Lower Bound of 95%CI	Upper Bound of 95%CI
Intercept	0.048	0.030	0.074
smoking	1.610	0.938	2.762
age	0.892	0.777	1.024

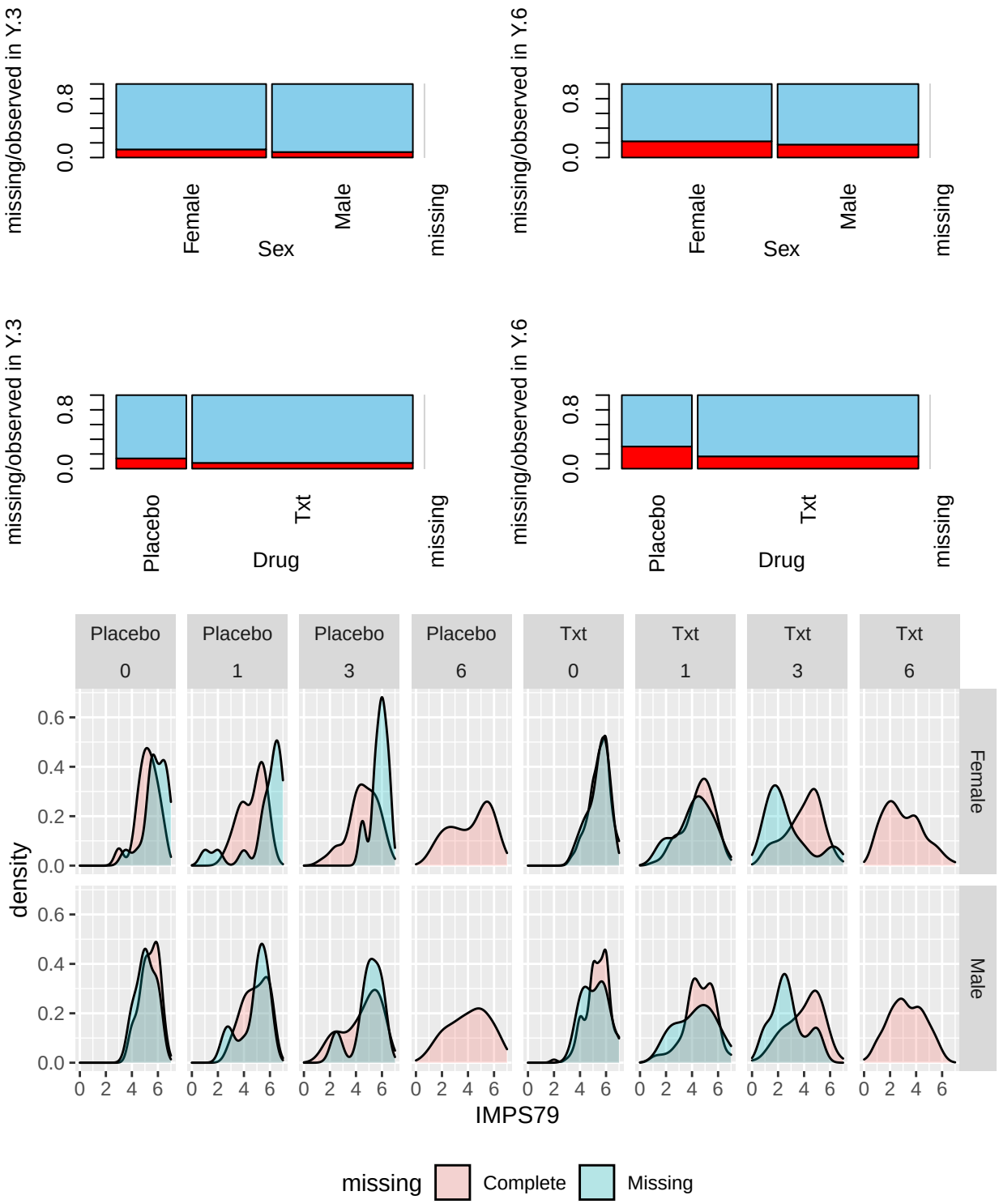
5. We can see from the previous results that there are differences between conditional models in complete dataset and dataset with missingness. The estimated coefficients are similar between the imputed

datasets and unimputed datasets. The variance of the estimated coeffcients are smallest in the conditional model over the imputed datasets of imputations using the data in long format among all conditional models. This conditional model also has a more similar result to the conditional model that fitted in the complete dataset compared to the data with missingness. The conditional model over the imputed datasets of imputations using the data in wide format has almost the same estimated coefficients as the conditional model over the data with missingness and no imputations, but with smaller variance.

Part B

1. In this exploratory analysis, we first find there's only missing values in the last two follow-ups (Week 3 and 6). At Week 3, about 10% participants dropped out and at Week 6, another ~10% participants dropped out. Thus we did barplots by treatment group and by sex for the last two follow-ups and find more female participants dropped out compared to male, and more participants receiving placebo dropped out compared to those receiving treatment. We finally created density plots by treatment group and sex over time and find that the placebo group with higher IMPS score tend to drop out and treatment group with lower IMPS score tend to drop out. This pattern exists in both genders.





2. We fit the following marginal model on complete observations,

$$logit(E[Y_{ij}|X_{ij}]) = \beta_0 + \beta_1 Week_{ij} + \beta_2 Drug_{ij} + \beta_3 Sex_{ij}$$

where $E[Y_{ij}|X_{ij}]$ is the probability of having high severity of schizophrenia condition on covariates X_{ij} for patient i at Week j. Week here is treated as numerical data. The results (not exponentiated)

are presented as the following table. Since this data contains missing data, we will use independence working matrix. The independence matrix will be used in question 3 also.

Table 5: Marginal Model For All Available Data

	Estimate	Std.err	Wald	Pr(>|W|)
(Intercept)	2.735	0.245	124.874	0.000
Week	-0.449	0.027	284.319	0.000
DrugTxt	-0.799	0.235	11.566	0.001
SexMale	0.113	0.176	0.410	0.522

3. For the missing model that missingness depends on time and treatment, we have:

$$logit(Pr[R_{ij} = 1|X_{ij}]) = \beta_0 + \beta_1 Week_{ij} + \beta_2 Drug_{ij}$$

, and for the missing model that missingness depends on time, treatment, sex, and most recent previous outcome, we have:

$$logit(Pr[R_{ij} = 1|Y_{ij-1}, X_{ij}]) = \beta_0 + \beta_1 Week_{ij} + \beta_2 Drug_{ij} + \beta_3 Sex_{ij} + \beta_4 Y_{ij-1}$$

The results (not exponentiated) are presented as the following tables.

Table 6: Marginal Model For Missing depending on Time and Treatment

	Estimates	Robust SE	z value	Pr(>|z|)
(Intercept)	2.735	0.244	11.193	0.000
Week	-0.441	0.026	-16.646	0.000
DrugTxt	-0.821	0.234	-3.504	0.000
SexMale	0.107	0.176	0.607	0.544

Table 7: Marginal Model For Missing depending on Time, Treatment, Sex, and Most Recent Previous Outcome

	Estimates	Robust SE	z value	Pr(>|z|)
(Intercept)	2.726	0.241	11.309	0.000
Week	-0.452	0.026	-17.271	0.000
DrugTxt	-0.803	0.233	-3.447	0.001
SexMale	0.109	0.178	0.615	0.539

4. In the three models, the effects of treatment is smaller in the model with only available data. This is reasonable because in the exploratory analysis, we find that in the placebo group, patients with higher scores tend to drop out and in the treatment group, patients with lower scores tend to drop out, so modeling on only available data would underestimate the effect of treatment. However, the differences, not only of the treatment but also other predictors, are very small among all the three models. This is probably due to a relatively low drop-out rate.

Code Appendix

```
#--------------------Load Packages------------------
library(ggplot2)
library(dplyr)
library(lattice)
library(lme4)
library(geepack)
library(VIM)
library(mice)
library(broom.mixed)
library(wgeesel)
library(knitr)
#--------------------Read Data-----------------------
miss.sixcity <- read.csv("C:/Users/second/Desktop/BIOST 540/Data Set/miss-sixcity.csv")
miss.sixcity = miss.sixcity[,-1]
miss.sixcity$smok = as.factor(miss.sixcity$smok)
sixcity <- read.csv("C:/Users/second/Desktop/BIOST 540/Data Set/sixcity.csv")
sixcity = sixcity[,-1]
sixcity$smok = as.factor(sixcity$smok)
levels(sixcity$smok) = c("no smoking","smoking")
levels(miss.sixcity$smok) = c("no smoking","smoking")
#--------------------Exploratory Data Analysis-------
miss.sixcity.temp = miss.sixcity[,-5]
miss.sixcity.temp.wide = reshape(miss.sixcity.temp, timevar = "age",idvar = "id",
                                 v.names = "resp", direction = "wide")
#-----------Histogram and missing pattern by cluster plots
aggr_plot = aggr(miss.sixcity.temp.wide, col = c("navyblue","red"),
                 numbers = TRUE, sortVars = F,
                 labels = names(miss.sixcity.temp.wide), cex.axis = .7, gap = 3,
                 ylab = c("Histogram of missing data","Pattern"))
#-----------Create barplots of proportions of missing data over time by smoking status group
par(mfrow=c(2,2))
spineMiss(miss.sixcity.temp.wide[, c("smok", "resp.-2")])
spineMiss(miss.sixcity.temp.wide[, c("smok", "resp.-1")])
spineMiss(miss.sixcity.temp.wide[, c("smok", "resp.0")])
spineMiss(miss.sixcity.temp.wide[, c("smok", "resp.1")])
#------------Bar plot by smoking status group and time
ids.miss = unique(miss.sixcity.temp$id[is.na(miss.sixcity.temp$resp)])
miss.sixcity.temp$missing = "Complete"
miss.sixcity.temp$missing[miss.sixcity.temp$id %in% ids.miss] = "Missing"
ggplot(miss.sixcity.temp, aes(x = resp, fill = missing)) +
  geom_bar(alpha =0.25) +
  facet_grid(. ~ smok + age) +
  theme(legend.position="bottom")
#------------------------Fiting Conditional Model over the complete Dataset---------
cond_complete = glmer(resp~smok+age+(1|id), data = sixcity,
                      family = binomial(link = "logit"), nAGQ = 20)
cond_complete.cc = confint(cond_complete)
cond_complete.cc = exp(cond_complete.cc)[-1,]
colnames(cond_complete.cc) = c("Lower Bound of 95% CI","Upper Bound of 95% CI")
complete_table =as.matrix(exp(summary(cond_complete)$coeff[,1]))
colnames(complete_table) ="Estimated Coefficients"
```

```r
knitr::kable(round(cbind(complete_table,cond_complete.cc),3),
      caption = "Estimated Coefficients of Conditional Model in Complete Data")
#-----------------------Fitting conditional model over the data with missingess----------
cond_miss = glmer(resp~smok+age+(1|id), data = miss.sixcity,
                  family = binomial(link = "logit"), nAGQ = 20)
cond_miss.cc = confint(cond_miss)
cond_miss.cc = exp(cond_miss.cc)[-1,]
colnames(cond_miss.cc) = c("Lower Bound of 95%CI","Upper Bound of 95%CI")
miss_table =as.matrix(exp(summary(cond_miss)$coeff[,1]))
colnames(miss_table) ="Estimated Coefficients"

kable(round(cbind(miss_table,cond_miss.cc),3),
      caption = "Estimated Coefficients of Conditional Model in Missing Data")
#--------------Fitting conditional model over data with imputations using the data in long format-----
miss.sixcity.long = miss.sixcity.temp
pred = make.predictorMatrix(miss.sixcity.long)*2
pred["resp","id"] = -2
imp_long = mice(miss.sixcity.long, method = "2l.bin", pred =pred,
                seed = 22, maxit = 1, m = 25, print = F)
fit.imp_long = with(imp_long, glmer(resp~smok+age+(1|id),
                                    family = binomial(link = "logit"), nAGQ = 20))
pool.imp_long = summary(pool(fit.imp_long),conf.int= T, exponentiate = T)
pool.imp_long[,-1] = round(pool.imp_long[,-1],3)
pool.imp_long = pool.imp_long[,c(2,7,8)]
rownames(pool.imp_long) = c("Intercept","smoking","age")
colnames(pool.imp_long) = c("Estimated Coefficients","Lower Bound of 95%CI","Upper Bound of 95%CI")


kable(pool.imp_long,
      caption = "Estimated Coefficients of Imputations Using the Data in Long Format")
#-----------------Fitting conditional model over data with imputations using the data in wide format---
miss.sixcity.temp = miss.sixcity[,-3]
miss.sixcity.temp.wide = reshape(miss.sixcity.temp, timevar = "time",idvar = "id",
                                 v.names = "resp", direction = "wide")
pred = make.predictorMatrix(miss.sixcity.temp.wide)
pred[,"id"]=0
imp = mice(miss.sixcity.temp.wide, m = 25, maxit = 10, method="logreg",
           seed = 540, pred = pred, print = F)
plot(imp, layout=c(4,2))

fitmodel = function(i, tempData){
  aux = complete(tempData,i)
  aux.long = reshape(aux, idvar = "id", varying = list(seq(3,6)),
                     v.names = "resp", direction = "long")
  aux.long = arrange(aux.long, id, time)
  aux.long$time = aux.long$time - 3
  fit = glmer(resp~smok+time+(1|id), data = aux.long,
              family = binomial(link = "logit"), nAGQ = 20)
  return(c('coef'=summary(fit)$coeff[,1],
           'var'=(summary(fit)$coefficients[,2])^2, df = df.residual(fit)))
}
v = sapply(1:25, FUN = fitmodel, imp)
```

```r
source("C:/Users/second/Desktop/BIOST 540/Data Set/poolmi.R")
pooled.estimates <- apply(v[1:3,],1,mean)
pooled.se <- multipleImputationStandardErrors(v[1:3,], v[4:6,])
pooled.df <- multipleImputationDegreesOfFreedom(v[1:3,], v[4:6,], v[7,1])
pooled.pv <- pt(-abs(pooled.estimates/pooled.se), df=pooled.df)*2
tb <- cbind(pooled.estimates, pooled.se, pooled.df, pooled.pv)
pool_table = matrix(exp(tb[,1]), ncol = 1)
rownames(pool_table) = c("Intercept","smoking","age")
colnames(pool_table) = "Estimated Coefficients"
confident_interval = matrix(NA, nrow = 3 , ncol = 2)
confident_interval[1,] = exp(tb[1,1] + qnorm(c(0.025,0.975))*tb[1,2])
confident_interval[2,] = exp(tb[2,1] + qnorm(c(0.025,0.975))*tb[2,2])
confident_interval[3,] = exp(tb[3,1] + qnorm(c(0.025,0.975))*tb[3,2])
colnames(confident_interval) = c("Lower Bound of 95%CI","Upper Bound of 95%CI")

kable(round(cbind(pool_table,confident_interval),3),
      caption ="Estimated Coefficients of Imputations Using the Data in Wide Format")
### ------------------------------------------------------------
### QB1
library(wgeesel)
data(imps)

imps$Drug <- factor(imps$Drug)
levels(imps$Drug) <- c("Placebo", "Txt")
imps$Sex <- factor(imps$Sex)
levels(imps$Sex) <- c("Female", "Male")

imps_wide <- reshape(imps[,c(-2,-6,-7)], timevar = "Week",
                     idvar = "ID", v.names = "Y", direction = "wide")

aggr_plot <- aggr(imps_wide,
                  col=c('navyblue','red'),
                  numbers=TRUE,
                  sortVars=F,
                  labels=names(imps_wide),
                  cex.axis=.7,
                  gap=3,
                  ylab=c("Histogram of missing data","Pattern"))


par(mfrow=c(2,2))
spineMiss(imps_wide[, c("Sex", "Y.3")])
spineMiss(imps_wide[, c("Sex", "Y.6")])
spineMiss(imps_wide[, c("Drug", "Y.3")])
spineMiss(imps_wide[, c("Drug", "Y.6")])

ids.miss <- unique(imps$ID[is.na(imps$Y)])
imps$missing <- "Complete"
imps$missing[imps$ID %in% ids.miss] <- "Missing"

ggplot(imps, aes(x = IMPS79, fill = missing)) +
  geom_density(alpha =0.25) +
  facet_grid(Sex ~ Drug + Week) +
```

```
  theme(legend.position="bottom")

### ------------------------------------------------------------
### QB2
imps_complete <- imps[!is.na(imps$Y),]

fit.2 <- geeglm(Y ~ Week + Drug + Sex, id=ID, data = imps_complete,
                family = binomial(link = "logit"), std.err = "san.se")
# summary(fit.2)

library(doBy)
fit2_summary <- round(coef(summary(fit.2)),3)

knitr::kable(fit2_summary, caption = "Marginal Model For All Available Data")
### ------------------------------------------------------------
### QB3
imps$lag1y <- ylag(imps$ID,imps$Y,1)

fit.3a <- wgee(Y ~ Week + Drug + Sex, id=imps$ID, data = imps,
                family="binomial",
                corstr ="independence", scale = NULL,
                mismodel = R ~ Week + Drug)
# summary(fit.3a)

Coef.3a <- matrix(NA,nrow=length(fit.3a$beta),ncol=4)
Coef.3a[,1] <- c(fit.3a$beta)
Coef.3a[,2] <- sqrt(diag(fit.3a$var))
Coef.3a[,3] <- Coef.3a[,1]/Coef.3a[,2]
Coef.3a[,4] <- round(2*pnorm(abs(Coef.3a[,3]), lower.tail=F), digits=8)
Coef.3a <- round(Coef.3a,3)
colnames(Coef.3a) <- c("Estimates","Robust SE", "z value", "Pr(>|z|)")
rownames(Coef.3a) <-rownames(fit.3a$beta)
knitr::kable(Coef.3a, caption = "Marginal Model For Missing depending on Time and Treatment")


fit.3b <- wgee(Y ~ Week + Drug + Sex, id=imps$ID, data = imps,
                family="binomial",
                corstr ="independence", scale = NULL,
                mismodel = R ~ Week + Drug + Sex + lag1y)
# summary(fit.3b)

Coef.3b <- matrix(NA,nrow=length(fit.3b$beta),ncol=4)
Coef.3b[,1] <- c(fit.3b$beta)
Coef.3b[,2] <- sqrt(diag(fit.3b$var))
Coef.3b[,3] <- Coef.3b[,1]/Coef.3b[,2]
Coef.3b[,4] <- round(2*pnorm(abs(Coef.3b[,3]), lower.tail=F), digits=8)
Coef.3b <- round(Coef.3b,3)
colnames(Coef.3b) <- c("Estimates","Robust SE", "z value", "Pr(>|z|)")
rownames(Coef.3b) <-rownames(fit.3b$beta)
knitr::kable(Coef.3b, caption = "Marginal Model For Missing depending on Time, Treatment, Sex, and Most
### ------------------------------------------------------------
### QB4
```