Diezhang Wu and Linxi Han
BIST5092
2021 Summer
MS-PhD Project on Phase III Clinical trial

1. Introduction and Study Objectives

Psoriasis is an inflammatory skin disease that causes red, flaky, crusty patches of skin covered with silvery scales. It occurs in 2 to 3% of the worldwide population. Even though there exists several therapies, many patients remain untreated or do not have an adequate response. Early clinical trials suggested that drug X has efficacy in the treatment of this skin disease.

To confirm the findings of early clinical trials with drug X, a phase III trial was conducted. In this phase III trial, patients with moderate-to-severe plaque psoriasis were randomly selected to receive drug X (210mg and 140mg), ustekinumab or placebo. The main purposes of this phase III clinical trial are to assess the safety and efficacy of drug X at two different doses (210mg and 140mg) compared with placebo and ustekinumab.

A major problem in the analysis of clinical trials is missing data caused by patients dropping out of the study before completion. The reason for dropout may be study-related (eg, adverse event, death, unpleasant study procedures, lack of improvement) or study-unrelated (eg, moving away, unrelated disease). This problem is especially a concern for a slow-acting treatment or a highly toxic drug. Missing data problems in clinical trials may result in biased treatment comparisons and also impact the overall statistical power of the study. There are three commonly seen mechanisms for missing data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Missing completely at random (MCAR) is defined as the process in which the probability is independent of both observed and unobserved measurements. The second mechanism, which is more restrictive than MCAR, is missing at random (MAR). Under MAR, the probability of dropout depends on observed data, but not the unobserved data. When observations are neither MCAR nor MAR, they are classified as missing not at random (MNAR). It means that the probability of an observation being missing depends on unobserved measurements.

In the following, we focus on a sensitivity analysis in the context of missing data. We will simulate missing data according to these three missing mechanisms and analyze data using different approaches. Once we obtain all of the results, we will compare the results of three missing mechanisms with the results of the full dataset (primary

analysis). The goal of the sensitivity analysis is to evaluate the robustness of our results and check whether the missing values have significant negative influence on our results.

2. Dataset description

The simulated data contains complete data. It means that every patient has data collected for the endpoint at visit 6. There are three datasets we used in our analysis, including ADSL, ADPA and ADAE. The Subject-Level Analysis Dataset (ADSL) is used to provide the variables that describe the attributes of a subject (e.g., age, sex, weight and PASI score). It is one record per study identifier per subject identifier. The structure of Efficacy Analysis Dataset (ADPA) is one record per study identifier per subject identifier per parameter code per analysis visit. Key variables we used in this dataset contain AVISIT, PARAMCD, TRTPN and PCHGCA1N. The dataset ADAE is the Adverse Event Analysis Dataset that captures adverse event's name, start and end date, serious level and relation to the treatment, which are all important for the drug's safety concern.

3. Methods

For analyzing the treatment effect, we performed logistic regression on the 75% skin clearance (PCHGCA1N) with planned treatment (TRTPN) and sex from the full dataset. The planned treatment is used to comply with the treatment policy strategy.

To mimic the MCAR mechanism, we randomly selected 10%, 20% and 30% data to be missing respectively.

The assumption for MAR is that the missingness is related to certain aspects of the observed data. We assumed that if patients experienced a less than 10% improvement (PCHG < 10), their probability of missing the next visit is 30%, otherwise this probability is 5%. When the measurement of one visit is missing, all subsequent visits will also be missing. The missingness was imputed for visit 4, visit 5 and visit 6.

MNAR refers to the missingness caused by certain aspects that are not explained by the observed data. To generate the mechanism of missing, we created toxicity scores based on the ADAE dataset that contains detailed information about the adverse events that occurred during the trial. The basic idea is that patients with higher toxicity scores are more likely to drop out. This score considers two components of the adverse event, the duration of the event and its severity. The following equations show the step by step calculation of this score.

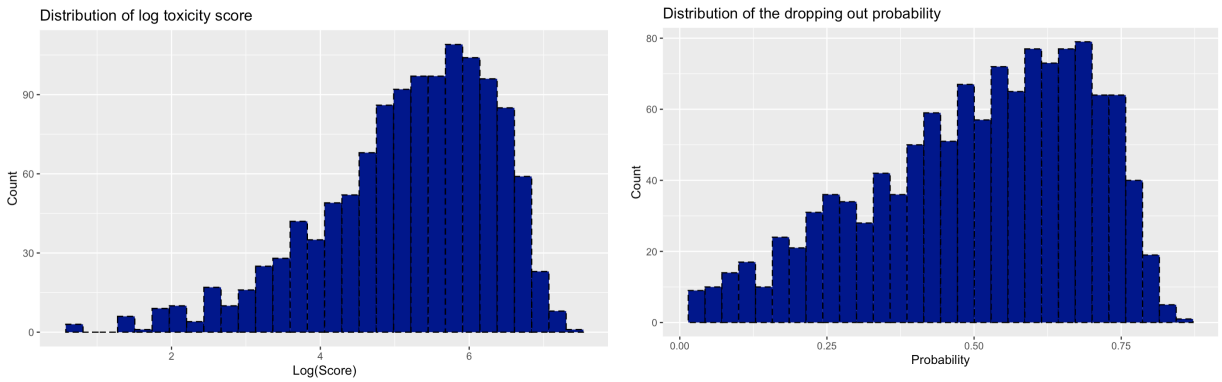
*Duration of the adverse event = event end date (*AEENDY*) - event start date (*AESTFY*)*

Severity score = 5 * AEREL + *3* * AESER + *2* * AESEV

*Toxicity score = Duration of the adverse event * Severity score*

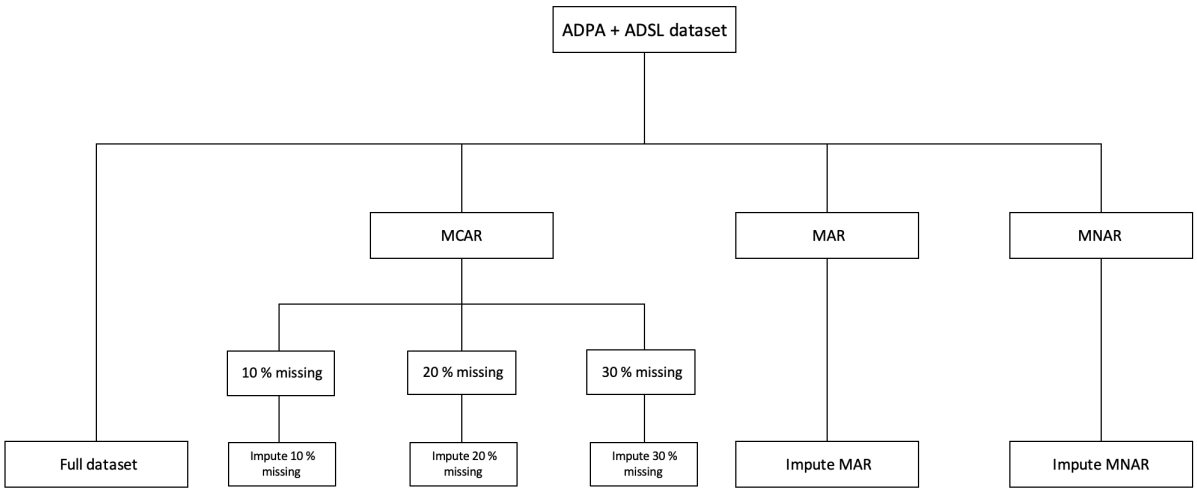
For AEREL and AESER, 1 is assigned to events recorded with “Y” and 0 otherwise. Variable AESEV is recoded as 1, 2, and 3 for “MILD”, “MODERATE” and “SEVERE”, respectively. The severity score is generated from a weighted combination of the three previously mentioned covariates. We consider the order of importance on dropout to be: if the adverse event is directly caused by the treatment > if the adverse event is serious > how severe this event is. Toxicity score for patients with multiple adverse events is simply the sum of individual events. We then take a log transformation of the score as it is extremely right skewed. The probability of dropout is generated from a normal distribution with the mean equal to the mean of the log toxicity score. The standard deviation is set to be larger than the standard deviation of the log toxicity score to avoid extremely high probability values. The simulated dropout probability has a range of 0.019 to 0.85. Figure 1 shows the distribution of the log toxicity score and simulated probability. Same as in MAR, the missingness is generated based on the simulated probability.

Figure 1. Distribution of log toxicity score and simulated dropout probability



Datasets containing missing values generated by all three mechanisms are imputed by recoding all missing values as treatment failures (PCHGCA1N = 0). The same logistic regression model is applied on the imputed datasets as well as the original full dataset (PCHGDA1N ~ TRTPN + SEX). Figure 2 illustrates the model generation process.

Figure 2. Diagram of the model generation process



The R markdown with detailed data imputation and analysis procedures is attached.

4. Results and discussion

The coefficient estimates from logistic regression are shown in Table 1. Results from the full dataset model show that all three treatments are effective at improving patients' skin conditions, and all effects are strongly statistically significant ($p < 2e\text{-}16$). Compared to patients taking placebo, the odds for patients taking active control (ustekinumab), 140mg, and 210mg of drug X to reach at least 75% skin clearance are $e^{3.2724} = 26.37$, $e^{3.1169} = 22.57$, $e^{4.2609} = 70.87$ times accordingly, after adjusted for a potential confounder sex.

The above result is further supported by the following sensitivity analysis in which we knocked out certain percentages of the data by all three proposed mechanisms of missing. We are still seeing significant treatment effects for models 2 through 6. Despite the magnitudes of the treatment effects being slightly reduced, the strength of the signals is still significant with extreme p-values (< 2e-16). Even when we assign the missing values with the worst outcome which is treatment failure, the treatment effects still stand out, as explained by models 7 to 11. The effect of sex remained insignificant for all models which is desirable. This sensitivity analysis reassures our primary findings of significant treatment effects.

To summarize, the current study proved the efficacy of drug X on clearing the skin conditions of psoriasis patients with proposed dosage against placebo. This conclusion is also endorsed by the sensitivity analysis.

Table 1. Coefficient estimates for all models

Model number	Model Description	Active control	140 mg	210 mg	Sex Male	
		log(OR)	log(OR)	log(OR)	log(OR)	p-value
1	Full dataset	3.2724	3.1169	4.2609	0.194	0.117
2	MCAR 10%	3.1982	3.062	4.1797	0.156	0.234
3	MCAR 20%	3.2763	3.154	4.2759	0.1718	0.221
4	MCAR 30%	3.1821	3.1922	4.325	0.2061	0.172
5	MAR	2.8004	2.6614	3.8378	0.235	0.0773
6	MNAR	3.1769	3.184	4.1499	0.222	0.138
7	MCAR 10% imputed	3.0016	2.864	3.7679	0.1499	0.195
8	MCAR 20% imputed	2.8794	2.8523	3.5386	0.1508	0.176
9	MCAR 30% imputed	2.6584	2.669	3.3617	0.1022	0.353
10	MAR imputed	3.0741	2.9271	3.9604	0.2029	0.0801
11	MNAR imputed	2.3862	2.5904	2.8242	0.0561	0.605

* p-value of all treatmentment effects from all models are significant $p < 2$ e-16

References

EMA (2010). Guideline on Missing Data in Confirmatory Clinical Trials. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-missing-data-confirmatory-clinical-trials_en.pdf

Myers, W. R. (2000). Handling missing data in clinical trials: an overview. *Drug information journal: DIJ/Drug Information Association*, *34*(2), 525-533

Lebwohl, M., Strober, B., Menter, A., Gordon, K., Weglowska, J., Puig, L., ... & Nirula, A. (2015). Phase 3 studies comparing brodalumab with ustekinumab in psoriasis. *New England Journal of Medicine*, *373*(14), 1318-1328.

BIST 5092 Phase III Project

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(stringr)
library(ggplot2)
```

import ADPA and ADSL datasets, select measurements at visit 6

```
adpa <- read.csv("../data_phase 3/ADPA.csv")
pasi <- adpa %>% filter(PARAMCD == "PASI")
pasi6 <- adpa %>% filter(PARAMCD == "PASI", AVISIT == "VISIT 6")
adsl <- read.csv("../data_phase 3/ADSL.csv")
adae <- read.csv("../data_phase 3/ADAE.csv")
```

perform logistic regression to assess the treatment effect on PASI75 adjusted by sex with the full dataset

```
dat_full <- merge(adsl, pasi6, by = "SUBJID")
summary(glm(PCHGCA1N ~ SEX + as.factor(TRTPN), data = dat_full, family = binomial))
```

```
##
## Call:
## glm(formula = PCHGCA1N ~ SEX + as.factor(TRTPN), family = binomial,
##     data = dat_full)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0185  -0.4241   0.5286   0.8783   2.2944
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -2.5574     0.2248 -11.377   <2e-16 ***
## SEXM                 0.1940     0.1239   1.566    0.117
## as.factor(TRTPN)2    3.2724     0.2438  13.421   <2e-16 ***
## as.factor(TRTPN)3    3.1169     0.2257  13.811   <2e-16 ***
## as.factor(TRTPN)4    4.2609     0.2395  17.791   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2396.1  on 1830  degrees of freedom
## Residual deviance: 1804.7  on 1826  degrees of freedom
## AIC: 1814.7
##
## Number of Fisher Scoring iterations: 5
```

MCAR

create functions to generate dataset with certain percent missing

```
generate_missing <- function(data, percent){
  id <- sample(1:nrow(data), round(nrow(data)*percent), replace = FALSE)
  data_par1 <- data[id,]
  data_par1$AVAL <- NA
  data_par2 <- data[-id,]
  data_final <- rbind(data_par1, data_par2)
  data_final$PCHGCA1N[which(is.na(data_final$AVAL))] <- NA

  return(data_final)
}
```

1. generate dataset with 10% missing and perform logistic regression

```
set.seed(1234)
pasi6_final1 <- generate_missing(pasi6, 0.1)
dat1 <- merge(adsl, pasi6_final1, by = "SUBJID")
summary(glm(PCHGCA1N ~ SEX + as.factor(TRTPN), data = dat1, family = binomial))
```

```
##
## Call:
## glm(formula = PCHGCA1N ~ SEX + as.factor(TRTPN), family = binomial,
##     data = dat1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0105  -0.4370   0.5332   0.8755   2.2534
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.4568     0.2306 -10.653   <2e-16 ***
## SEXM                0.1560     0.1309   1.191    0.234
## as.factor(TRTPN)2   3.1982     0.2515  12.715   <2e-16 ***
## as.factor(TRTPN)3   3.0620     0.2323  13.182   <2e-16 ***
## as.factor(TRTPN)4   4.1797     0.2464  16.965   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 2149.9  on 1647  degrees of freedom
## Residual deviance: 1628.0  on 1643  degrees of freedom
##   (183 observations deleted due to missingness)
## AIC: 1638
##
## Number of Fisher Scoring iterations: 4
```

2. Impute the missingness and reanalyze the data

```
dat1_im <- dat1
dat1_im$PCHGCA1N[which(is.na(dat1_im$AVAL))] <- 0
summary(glm(PCHGCA1N ~ SEX + as.factor(TRTPN), data = dat1_im, family = binomial))
```

```
##
## Call:
## glm(formula = PCHGCA1N ~ SEX + as.factor(TRTPN), family = binomial,
##     data = dat1_im)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7759  -1.3036   0.6803   0.9433   2.3003
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.5722     0.2261 -11.375   <2e-16 ***
## SEXM                0.1499     0.1157   1.295    0.195
## as.factor(TRTPN)2   3.0016     0.2439  12.304   <2e-16 ***
## as.factor(TRTPN)3   2.8640     0.2281  12.558   <2e-16 ***
## as.factor(TRTPN)4   3.7679     0.2344  16.078   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2493.8  on 1830  degrees of freedom
## Residual deviance: 2020.7  on 1826  degrees of freedom
## AIC: 2030.7
##
## Number of Fisher Scoring iterations: 5
```

3. repeat 1 & 2 with 20% missingness

```
## with 20% missingness
set.seed(1234)
pasi6_final2 <- generate_missing(pasi6, 0.2)
dat2 <- merge(adsl, pasi6_final2, by = "SUBJID")
summary(glm(PCHGCA1N ~ SEX + as.factor(TRTPN), data = dat2, family = binomial))
```

```
##
## Call:
## glm(formula = PCHGCA1N ~ SEX + as.factor(TRTPN), family = binomial,
##     data = dat2)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -2.0328  -0.4278   0.5205   0.8583   2.2780
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.5170     0.2510 -10.026   <2e-16 ***
## SEXM                0.1718     0.1402   1.225    0.221
## as.factor(TRTPN)2   3.2763     0.2740  11.956   <2e-16 ***
## as.factor(TRTPN)3   3.1540     0.2528  12.475   <2e-16 ***
## as.factor(TRTPN)4   4.2759     0.2684  15.932   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1898.6  on 1464  degrees of freedom
## Residual deviance: 1425.3  on 1460  degrees of freedom
##   (366 observations deleted due to missingness)
## AIC: 1435.3
##
## Number of Fisher Scoring iterations: 5
```

```
## with 20% missingness imputed
dat2_im <- dat2
dat2_im$PCHGCA1N[which(is.na(dat2_im$AVAL))] <- 0
summary(glm(PCHGCA1N ~ SEX + as.factor(TRTPN), data = dat2_im, family = binomial))
```

```
##
## Call:
## glm(formula = PCHGCA1N ~ SEX + as.factor(TRTPN), family = binomial,
##     data = dat2_im)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5847  -1.2246   0.8189   1.0687   2.3790
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.7691     0.2429 -11.399   <2e-16 ***
## SEXM                0.1508     0.1115   1.352    0.176
## as.factor(TRTPN)2   2.8794     0.2588  11.126   <2e-16 ***
## as.factor(TRTPN)3   2.8523     0.2451  11.635   <2e-16 ***
## as.factor(TRTPN)4   3.5386     0.2477  14.285   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2535.6  on 1830  degrees of freedom
## Residual deviance: 2140.8  on 1826  degrees of freedom
## AIC: 2150.8
##
## Number of Fisher Scoring iterations: 5
```

repeat 1 & 2 with 30% missingness

```
## with 30% missingness
set.seed(1234)
pasi6_final3 <- generate_missing(pasi6, 0.3)
dat3 <- merge(adsl, pasi6_final3, by = "SUBJID")
summary(glm(PCHGCA1N ~ SEX + as.factor(TRTPN), data = dat3, family = binomial))
```

```
##
## Call:
## glm(formula = PCHGCA1N ~ SEX + as.factor(TRTPN), family = binomial,
##     data = dat3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0680  -0.4346   0.5008   0.8330   2.2785
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.5182     0.2668  -9.438   <2e-16 ***
## SEXM                0.2061     0.1511   1.364    0.172
## as.factor(TRTPN)2   3.1821     0.2894  10.996   <2e-16 ***
## as.factor(TRTPN)3   3.1922     0.2684  11.893   <2e-16 ***
## as.factor(TRTPN)4   4.3250     0.2849  15.182   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1651.6  on 1281  degrees of freedom
## Residual deviance: 1225.1  on 1277  degrees of freedom
##   (549 observations deleted due to missingness)
## AIC: 1235.1
##
## Number of Fisher Scoring iterations: 5
```

```
## with 30% missingness imputed
dat3_im <- dat3
dat3_im$PCHGCA1N[which(is.na(dat3_im$AVAL))] <- 0
summary(glm(PCHGCA1N ~ SEX + as.factor(TRTPN), data = dat3_im, family = binomial))
```

```
##
## Call:
## glm(formula = PCHGCA1N ~ SEX + as.factor(TRTPN), family = binomial,
##     data = dat3_im)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4471  -1.1401  -0.3355   0.9689   2.4106
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.8492     0.2535 -11.241   <2e-16 ***
## SEXM                0.1022     0.1101   0.929    0.353
## as.factor(TRTPN)2   2.6584     0.2691   9.880   <2e-16 ***
```

```
## as.factor(TRTPN)3   2.6690     0.2561  10.422   <2e-16 ***
## as.factor(TRTPN)4   3.3617     0.2572  13.072   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2525.8  on 1830  degrees of freedom
## Residual deviance: 2193.2  on 1826  degrees of freedom
## AIC: 2203.2
##
## Number of Fisher Scoring iterations: 5
```

MAR

```
# select subject id, visit number, percent change from baseline at each visit
# reshape the dataset into wide format
pasi_inter <- pasi %>%
  select(SUBJID, AVISIT, PCHG) %>%
  spread(AVISIT, PCHG)

# generate missing data conditioned on previous visits
# in miss4, miss5 and miss6, 0 is missing, 1 is not missing
set.seed(1234)
for (i in 1:nrow(pasi_inter)){
  if (pasi_inter$`VISIT 3`[i] < 10){
    pasi_inter$miss4[i] <- rbinom(1, 1, 0.7)
  } else{ pasi_inter$miss4[i] <- rbinom(1, 1, 0.95)}
}

for (i in 1:nrow(pasi_inter)){
  if (pasi_inter$miss4[i] == 0){
    pasi_inter$miss5[i] = 0
  } else if (pasi_inter$`VISIT 4`[i] < 10){
    pasi_inter$miss5[i] <- rbinom(1, 1, 0.7)
  } else{ pasi_inter$miss5[i] <- rbinom(1, 1, 0.95)}
}

for (i in 1:nrow(pasi_inter)){
  if (pasi_inter$miss4[i] == 0){
    pasi_inter$miss6[i] = 0
  } else if (pasi_inter$`VISIT 5`[i] < 10){
    pasi_inter$miss6[i] <- rbinom(1, 1, 0.7)
  } else{ pasi_inter$miss6[i] <- rbinom(1, 1, 0.95)}
}

# check percent of missing data at visit 6
1 - sum(pasi_inter$miss6)/nrow(pasi_inter) # about 17% missingness
```

```
## [1] 0.1693064
```

```
# generate missingness in the pasi6 data according to the missigness pattern obtained in the previous s
pasi6_mar <- pasi6
pasi6_mar$PCHGCA1N[which(pasi_inter$miss6 == 0)] <- NA # PCHGCA1N contains visit 6 results with missing
```

```r
pasi6_mar$im6 <- pasi6_mar$PCHGCA1N
pasi6_mar$im6[which(is.na(pasi6_mar$PCHGCA1N))] <- 0 # impute all missing data as non-responders

dat_mar <- merge(adsl, pasi6_mar, by = "SUBJID")
```

```r
# logistic regression on PASI75 with missing values excluded
summary(glm(PCHGCA1N ~ SEX + as.factor(TRTPN), data = dat_mar, family = binomial))
```

```
##
## Call:
## glm(formula = PCHGCA1N ~ SEX + as.factor(TRTPN), family = binomial,
##     data = dat_mar)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0431  -0.5322   0.5146   0.8693   2.1126
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.1180     0.2494  -8.491   <2e-16 ***
## SEXM                0.2350     0.1330   1.767   0.0773 .
## as.factor(TRTPN)2   2.8004     0.2693  10.400   <2e-16 ***
## as.factor(TRTPN)3   2.6614     0.2509  10.608   <2e-16 ***
## as.factor(TRTPN)4   3.8378     0.2645  14.511   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1893.6  on 1520  degrees of freedom
## Residual deviance: 1553.0  on 1516  degrees of freedom
##   (310 observations deleted due to missingness)
## AIC: 1563
##
## Number of Fisher Scoring iterations: 4
```

```r
# logistic regression on PASI75 with missing data imputed as 0
summary(glm(im6 ~ SEX + as.factor(TRTPN), data = dat_mar, family = binomial))
```

```
##
## Call:
## glm(formula = im6 ~ SEX + as.factor(TRTPN), family = binomial,
##     data = dat_mar)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8053  -1.2527   0.6605   0.9640   2.3722
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.7518     0.2393  -11.50   <2e-16 ***
## SEXM                0.2029     0.1159    1.75   0.0801 .
## as.factor(TRTPN)2   3.0741     0.2554   12.04   <2e-16 ***
## as.factor(TRTPN)3   2.9271     0.2405   12.17   <2e-16 ***
## as.factor(TRTPN)4   3.9604     0.2472   16.02   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2502.7  on 1830  degrees of freedom
## Residual deviance: 2004.2  on 1826  degrees of freedom
## AIC: 2014.2
##
## Number of Fisher Scoring iterations: 5
```

MNAR

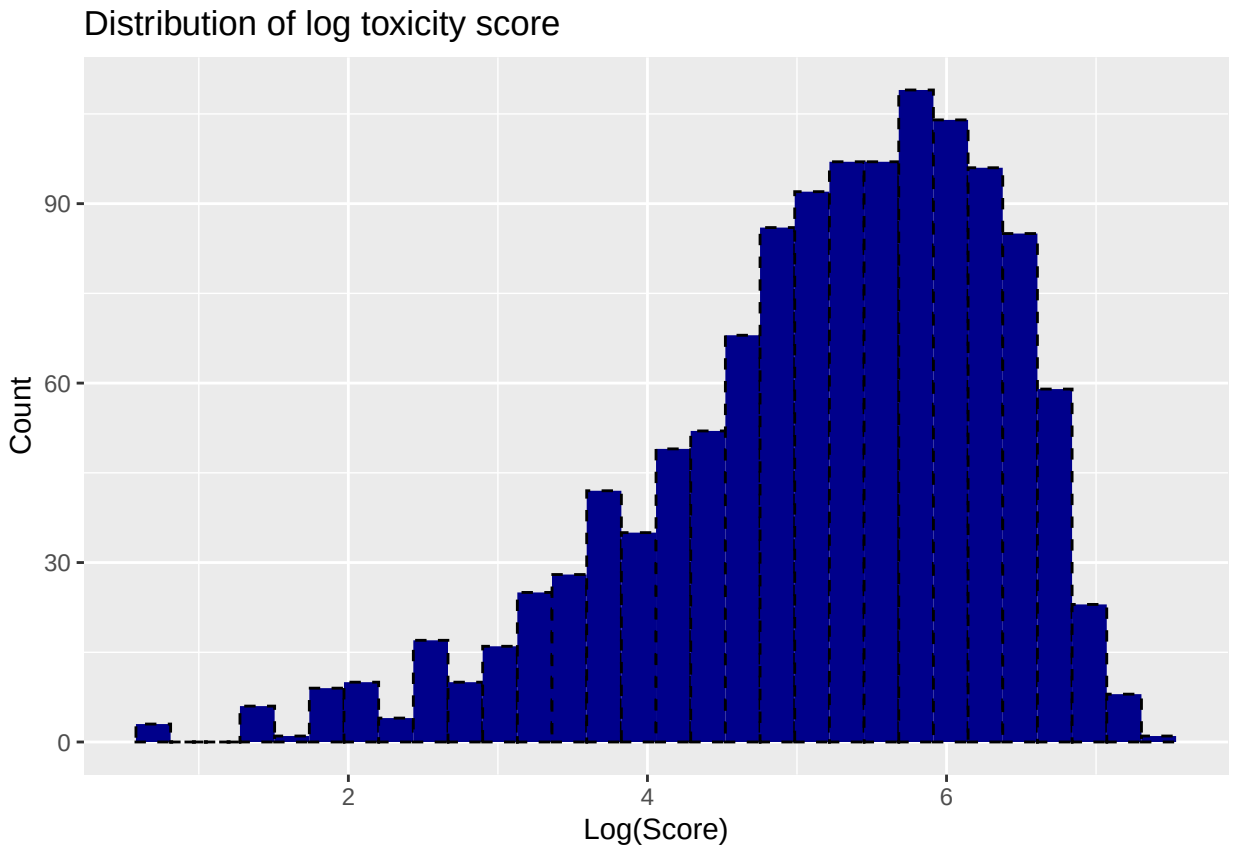
```
# sum the duration of each adverse event to generate total days of adverse event the patient had during
adae$duration <- adae$AEENDY - adae$AESTDY
#adduration <- adae %>% group_by(USUBJID) %>% summarise(duration = sum(duration))

# create severity score for each patient with AE using AESER, AESEV, and AEREL
# AESER_rate = 1 for Y, AESEV_rate : 1 for mild, 2 for moderate, 3 for severe, AEREL_rate = 1 for Y
adae$AESER_rate <- ifelse(adae$AESER == "Y", 1, 0)
adae$AESEV_rate <- ifelse(adae$AESEV == "MILD", 1, ifelse(adae$AESEV == "MODERATE", 2, 3))
adae$AEREL_rate <- ifelse(adae$AEREL == "Y", 1, 0)

# severity score score per event = AEREL_rate * 5 + AESER_rate * 3 + AESEV_rate * 2
adae$serscore <- adae$AEREL_rate * 5 + adae$AESER_rate * 3 + adae$AESEV_rate * 2

# toxic score = severity score * duration
adae$toxscore <- adae$serscore * adae$duration

# compute toxicity scale for each subject by sum the toxic score from each adverse event
toxscale <- adae %>% group_by(USUBJID) %>% summarise(toxscore = sum(toxscore))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
# check the distribution of the toxic scale
# hist(toxscale$toxscore)
toxscale$logscore <- log(toxscale$toxscore)
toxscale %>% ggplot(aes(x = logscore)) +
  geom_histogram(color="black", fill="darkblue", linetype = "dashed") +
  labs(title = "Distribution of log toxicity score", x = "Log(Score)", y = "Count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```

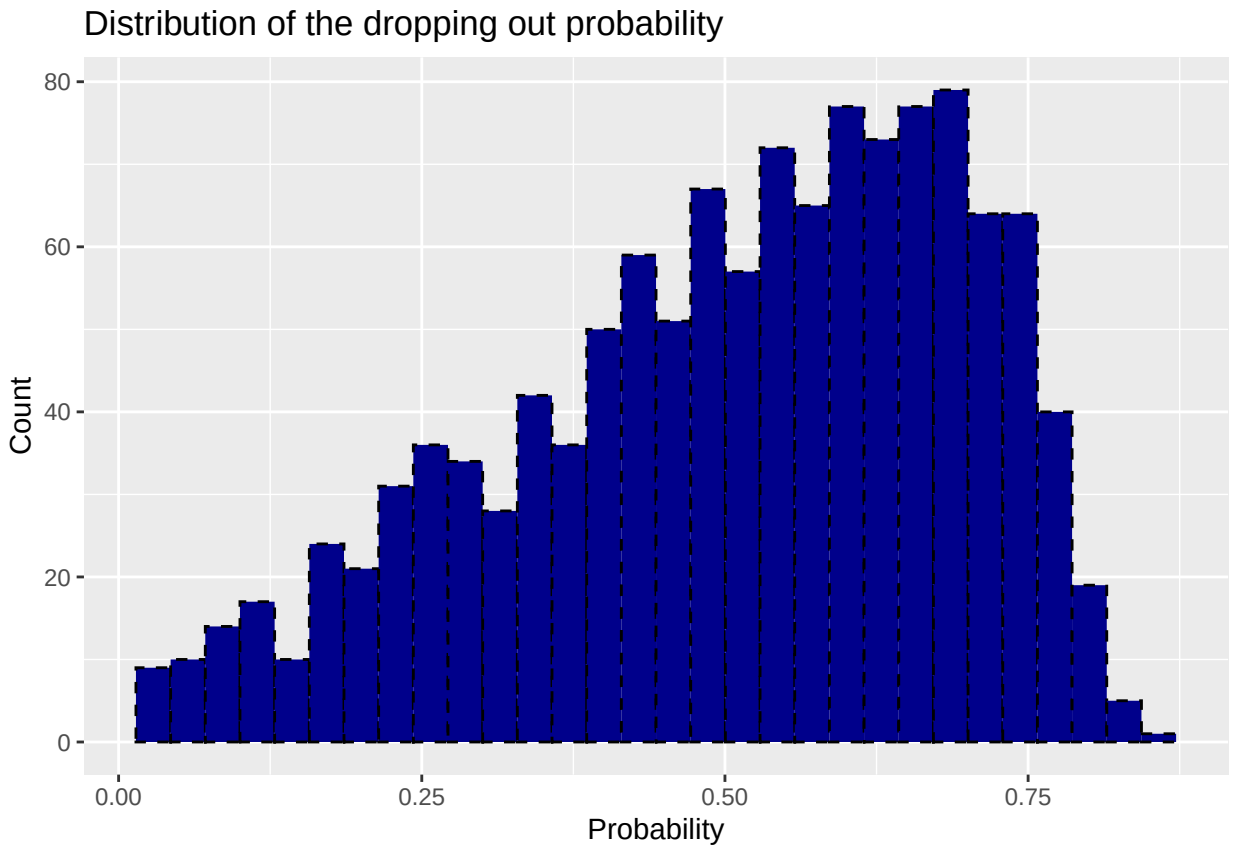



```
# assume the probability of dropping out is positively related to the toxicity scale and has an uniform
# the probability then can be simply determined by dividing the toxicity score with a common denominato
toxscale <- toxscale[which(toxscale$logscore != -Inf),]
mu <- mean(toxscale$logscore)
sigma <- sd(toxscale$logscore) + 1
toxscale$problog2 <- pnorm(toxscale$logscore, mu, sigma)

# impute the missingness of the subjects with adverse events with their probability of dropping out
aemissing <- toxscale
set.seed(1234)
for (i in 1:nrow(aemissing)){aemissing$misslog2[i] <- rbinom(1, 1, 1 - aemissing$problog2[i])}
```

```
## Warning: Unknown or uninitialised column: `misslog2`.
```

```
# check the probability distribution
toxscale %>% ggplot(aes(x = problog2)) +
  geom_histogram(color="black", fill="darkblue", linetype = "dashed") +
  labs(title = "Distribution of the dropping out probability", x = "Probability", y = "Count")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
aemissing$SUBJID <- as.numeric(substr(aemissing$USUBJID, nchar(aemissing$USUBJID) - 4 + 1, nchar(aemiss

# check the percent of missingness
# table(aemissing$misslog2) # 33% missing values
# 605 / nrow(pasi6)

# join the aemissing table to pasi6 and set visit6 results to be missing based on results from the prev
pasi6_mnar <- left_join(pasi6, aemissing, by = "SUBJID")
pasi6_mnar$PCHGCA1N[which(pasi6_mnar$misslog2 == 0)] <- NA
pasi6_mnar$im6 <- pasi6_mnar$PCHGCA1N
pasi6_mnar$im6[which(is.na(pasi6_mnar$PCHGCA1N))] <- 0
dat_mnar <- merge(pasi6_mnar, adsl, by = "SUBJID")

# perform logistic regression on pasi75 with incomplete dataset and imputed dataset
summary(glm(PCHGCA1N ~ SEX + as.factor(TRTPN), data = dat_mnar, family = binomial))
```

```
##
## Call:
## glm(formula = PCHGCA1N ~ SEX + as.factor(TRTPN), family = binomial,
##     data = dat_mnar)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9875  -0.4321   0.5466   0.8402   2.2899
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)        -2.5463     0.2541 -10.020   <2e-16 ***
## SEXM                0.2220     0.1495   1.485    0.138
## as.factor(TRTPN)2   3.1769     0.2817  11.280   <2e-16 ***
## as.factor(TRTPN)3   3.1840     0.2559  12.442   <2e-16 ***
## as.factor(TRTPN)4   4.1499     0.2742  15.136   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1649.5  on 1242  degrees of freedom
## Residual deviance: 1224.4  on 1238  degrees of freedom
##   (588 observations deleted due to missingness)
## AIC: 1234.4
##
## Number of Fisher Scoring iterations: 5
```

```
summary(glm(im6 ~ SEX + as.factor(TRTPN), data = dat_mnar, family = binomial))
```

```
##
## Call:
## glm(formula = im6 ~ SEX + as.factor(TRTPN), family = binomial,
##     data = dat_mnar)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2519  -1.1519  -0.3694   1.1280   2.3543
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -2.70675    0.24169 -11.199   <2e-16 ***
## SEXM               0.05606    0.10849   0.517    0.605
## as.factor(TRTPN)2  2.38617    0.25901   9.213   <2e-16 ***
## as.factor(TRTPN)3  2.59042    0.24502  10.572   <2e-16 ***
## as.factor(TRTPN)4  2.82420    0.24511  11.522   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2493.1  on 1830  degrees of freedom
## Residual deviance: 2247.2  on 1826  degrees of freedom
## AIC: 2257.2
##
## Number of Fisher Scoring iterations: 5
```