Missing Data Handling for Phase III Trial Data

Ruishan Lin Zelin Wang Xiangyi Xi

08/11/2023

Disclaimer

- **Disclaimer of data**: The data is simulated and is not real data from clinical study. They are for educational and exercise purpose only.
- **Disclaimer**: The design of the trial is created based on publicly available open-source information in immunology therapeutic area. It is for educational purpose only. The presentations reflect the views of the speakers based on their understanding of the open source information and simulated data, and have not been evaluated by University of connecticut. The materials cannot be used for promotional activities. They are not intended to diagnose, treat, cure or prevent any disease.

Overview

1. Introduction
    1. Background
    2. Motivation
    3. Goal
2. Data Analysis
3. Simulation Studies
    1. MCAR
    2. MAR
    3. MNAR
4. Conclusions and Discussions

Background

Phase III study

- Tests on new and wider demographic
- Tests for long term effectiveness and comparisons with other medications

Figure: Psoriasis

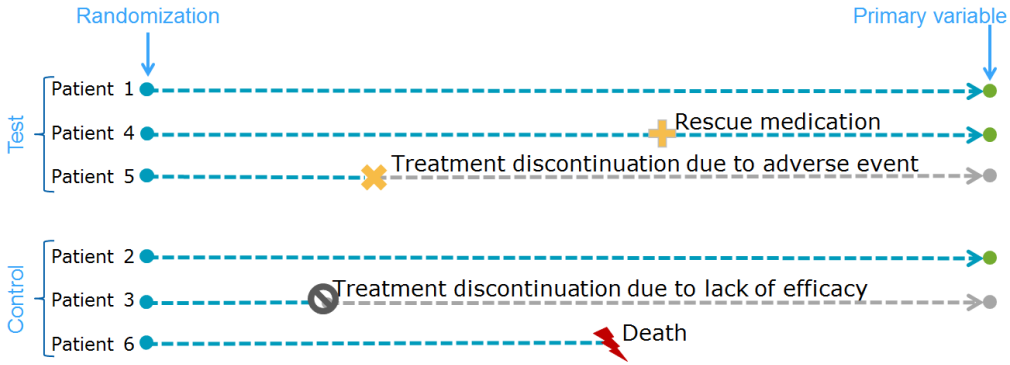


Figure: ICEs

Background

Missing Data Mechanisms

Let $Y = (y_{ij})$ denote an $(n \times K)$ dataset without missing values. Define the *missing-data indicator matrix* $M = (m_{ij})$ such that

$$\begin{cases} y_{ij} \text{ is missing if } m_{ij} = 0 \\ y_{ij} \text{ is present if } m_{ij} = 1 \end{cases}$$

The missing-data mechanism is characterized by the conditional distribution of M given Y, say $f(M|Y, \phi)$, where ϕ denotes unknown parameters. Based on $f(M|Y, \phi)$, there are three missing meachanism:

- Missing completely at random (MCAR): $f(M|Y, \phi) = f(M|\phi)$ for all Y, ϕ
- Missing at random (MAR): $f(M|Y, \phi) = f(M|Y_{obs}, \phi)$ for all Y_{mis}, ϕ
- Missing not at random (MNAR): $f(M|Y, \phi) = f(M|Y_{mis}, \phi)$ for all Y_{mis}, ϕ

Background

Missing completely at random (MCAR)

- Probability of missing unrelated to any variable under study

  Subset with missing data = simple random sample
- Assumption often made, but not always reasonable
- Can partially test for MCAR by checking associating between response and variables of interest
- Example: Random failure of the experimental instrument (e.g. test tube break, equipment failure)

Background

Missing at random (MAR)

- Probability of missing depends on some observed variable(s) but not the missing one(s)

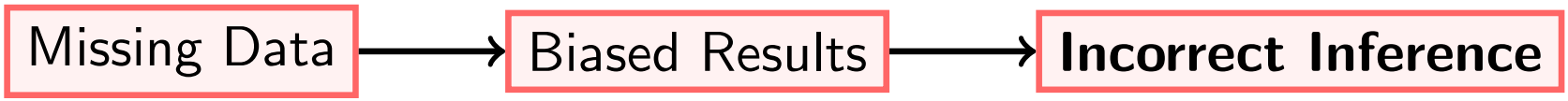
  Within the observed variable, subset = random sample from that group
- More reasonable assumption - most methods assume this mechanism
- Cannot test assumption
- Example: Dropout based on known baseline characteristics

Background

Missing not at random (MNAR)

- Probability of missing depends on unobserved predictors
- Impossible to test
- Example: Dropout based on the treatment outcome

Motivation & Goal



Missing Data Approaches

- Complete case analysis
- Single imputation methods e.g. Last Observation Carried Forward (LOCF)
- Multiple imputation
- Model based approaches

Compare multiple approaches for types of missing data and the complete data

An Overview of the Phase III Data

		PASI	
Trtp	Sex	0	1
1	F	102	10
	M	182	15
2	F	29	64
	M	61	146
3	F	78	123
	M	126	283
4	F	26	152
	M	58	375

Frequency Table: Complete Data at Visit 6

Total Number: 1831

Trtp meaning:

1. Placebo
2. Active control
3. Test drug 140mg
4. Test drug 210mg

Frequency Tables - MCAR 10% Missing

		PASI	
Trtp	Sex	0	1
1	F	93	9
	M	171	12
2	F	26	60
	M	53	122
3	F	71	109
	M	114	250
4	F	25	138
	M	53	342

Completers Data
Total Number: 1648

		PASI	
Trtp	Sex	0	1
1	F	103	9
	M	185	12
2	F	33	60
	M	85	122
3	F	92	109
	M	159	250
4	F	40	138
	M	92	342

Imputed Data
(Composite)

		PASI	
Trtp	Sex	0	1
1	F	103	9
	M	184	13
2	F	33	60
	M	82	125
3	F	90	111
	M	147	262
4	F	36	142
	M	82	352

Imputed Data
(LOCF)

Approach Matters!!!

Frequency - Tables MCAR 20% Missing

		PASI	
Trtp	Sex	0	1
1	F	74	9
	M	147	12
2	F	24	53
	M	50	122
3	F	58	103
	M	103	216
4	F	23	118
	M	49	304

Completers Data
Total Number: 1465

		PASI	
Trtp	Sex	0	1
1	F	103	9
	M	185	12
2	F	40	53
	M	85	122
3	F	98	103
	M	193	216
4	F	60	118
	M	130	304

Imputed Data
(Composite)

		PASI	
Trtp	Sex	0	1
1	F	103	9
	M	184	13
2	F	39	54
	M	79	128
3	F	92	109
	M	177	232
4	F	54	124
	M	106	328

Imputed Data
(LOCF)

Frequency Tables - MCAR 30% Missing

		PASI	
Trtp	Sex	0	1
1	F	66	9
	M	124	11
2	F	21	41
	M	45	99
3	F	56	89
	M	94	201
4	F	20	113
	M	36	257

Completers Data
Total Number: 1282
Decreased
Sample Size

		PASI	
Trtp	Sex	0	1
1	F	103	9
	M	186	11
2	F	52	41
	M	108	99
3	F	112	89
	M	208	201
4	F	65	113
	M	177	257

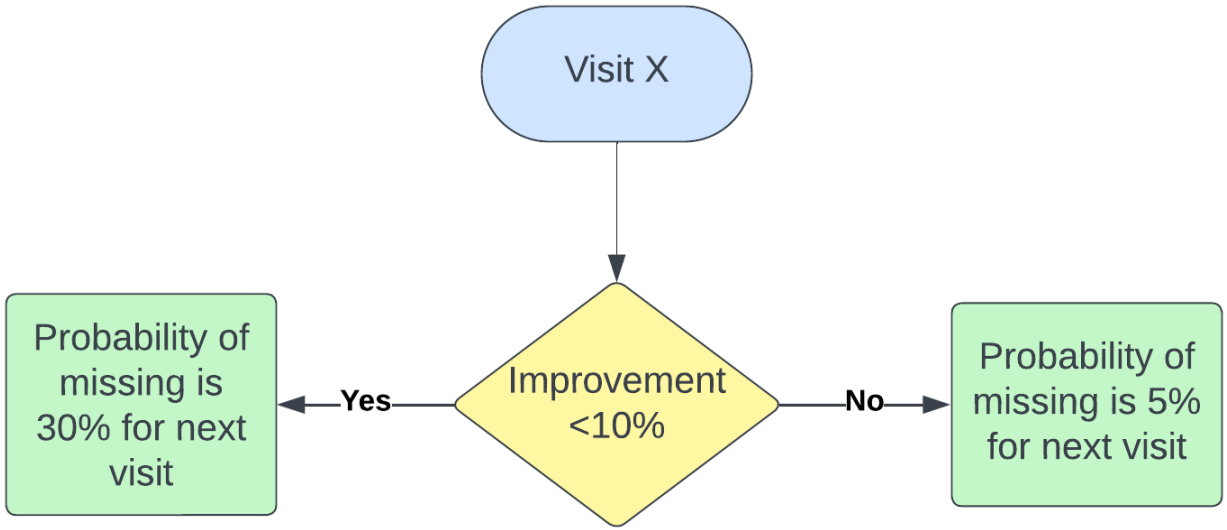
Imputed Data
(Composite)
Underestimated
Treatment Effects

		PASI	
Trtp	Sex	0	1
1	F	103	9
	M	184	13
2	F	47	46
	M	98	109
3	F	102	99
	M	187	222
4	F	56	122
	M	129	305

Imputed Data
(LOCF)
Constant Response
Assumption

MAR

- Probability of missing depends on some observed variable(s) but not the missing one(s)
- Within the observed variable, subset = random sample from that group



Frequency Tables - MAR

		PASI	
Trtp	Sex	0	1
1	F	36	9
	M	76	15
2	F	27	49
	M	52	127
3	F	65	106
	M	93	236
4	F	21	126
	M	47	326

Completers Data
Total Number: 1411

		PASI	
Trtp	Sex	0	1
1	F	103	9
	M	182	15
2	F	44	49
	M	80	127
3	F	95	106
	M	173	236
4	F	52	126
	M	108	326

Imputed Data
(Composite)

		PASI	
Trtp	Sex	0	1
1	F	103	9
	M	182	15
2	F	43	50
	M	76	131
3	F	92	109
	M	160	249
4	F	44	134
	M	90	344

Imputed Data
(LOCF)

Frequency Tables - MAR

		PASI	
Trtp	Sex	0	1
1	F	36	9
	M	76	15
2	F	27	49
	M	52	127
3	F	65	106
	M	93	236
4	F	21	126
	M	47	326

Completers Data
Total Number: 1411

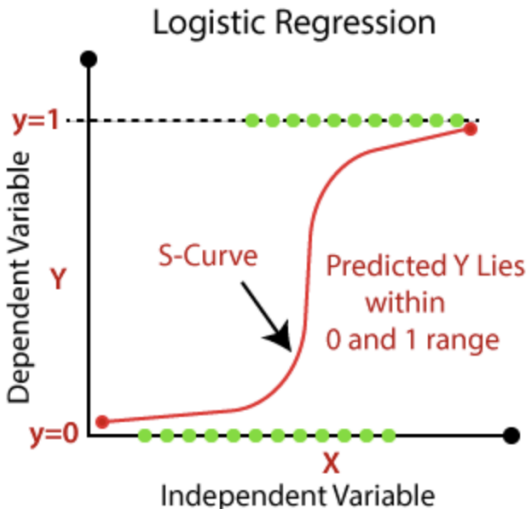
		PASI	
Trtp	Sex	0	1
1	F	103	9
	M	182	15
2	F	44	49
	M	80	127
3	F	95	106
	M	173	236
4	F	52	126
	M	108	326

Imputed Data
(Composite)

		PASI	
Trtp	Sex	0	1
1	F	103	9
	M	182	15
2	F	43	50
	M	76	131
3	F	92	109
	M	160	249
4	F	44	134
	M	90	344

Imputed Data
(LOCF)

Logistic Regression



$$\text{Odds} = \frac{\text{Probability Event Occurs(p)}}{\text{Probability Event Does Not Occur(1-p)}}$$

$$\text{Odds Ratio} = \frac{\text{Odds of an Event (Condition A)}}{\text{Odds of an Event(Condition B)}}$$

Logistic Regression Equation in R

glm(PASI75 Response $\sim$ *Treatment* + *Sex*, *family* = *binomial*())

Logistic Regression and Odds Ratio

Logistic Regression Let p denote the probability that the binary dependent variable Y (PASI75 Response) equals 1. There are two independent variables, namely X_1 (SEX) and X_2 (TRTP). The multiple logistic regression model can be written as follows:

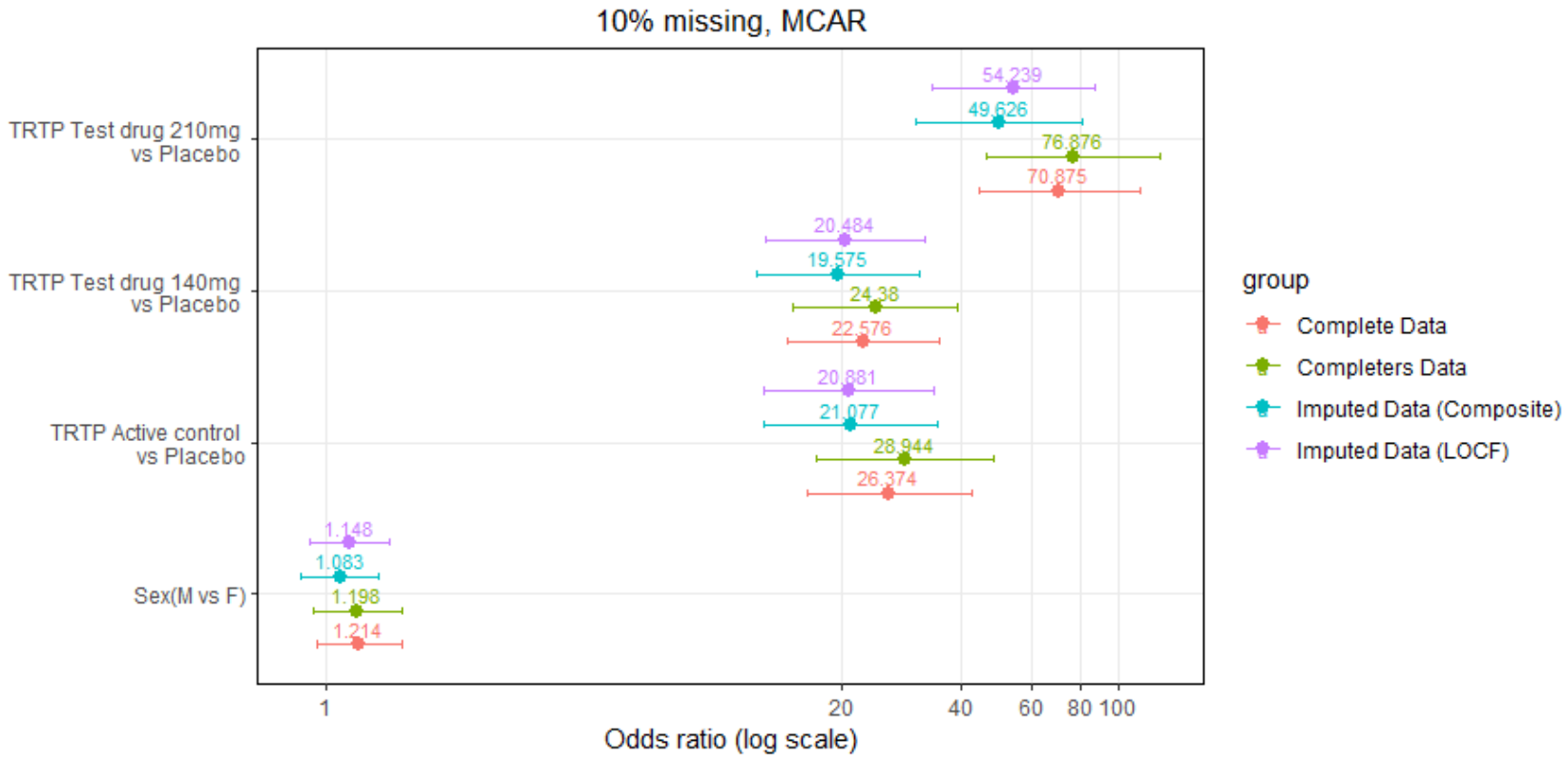
$$logit(p) = ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Odds Ratio

$$exp(\beta_1) = \frac{P(Y=1|X_1=1,X_2)/P(Y=0|X_1=1,X_2)}{P(Y=1|X_1=0,X_2)/P(Y=0|X_1=0,X_2)}$$

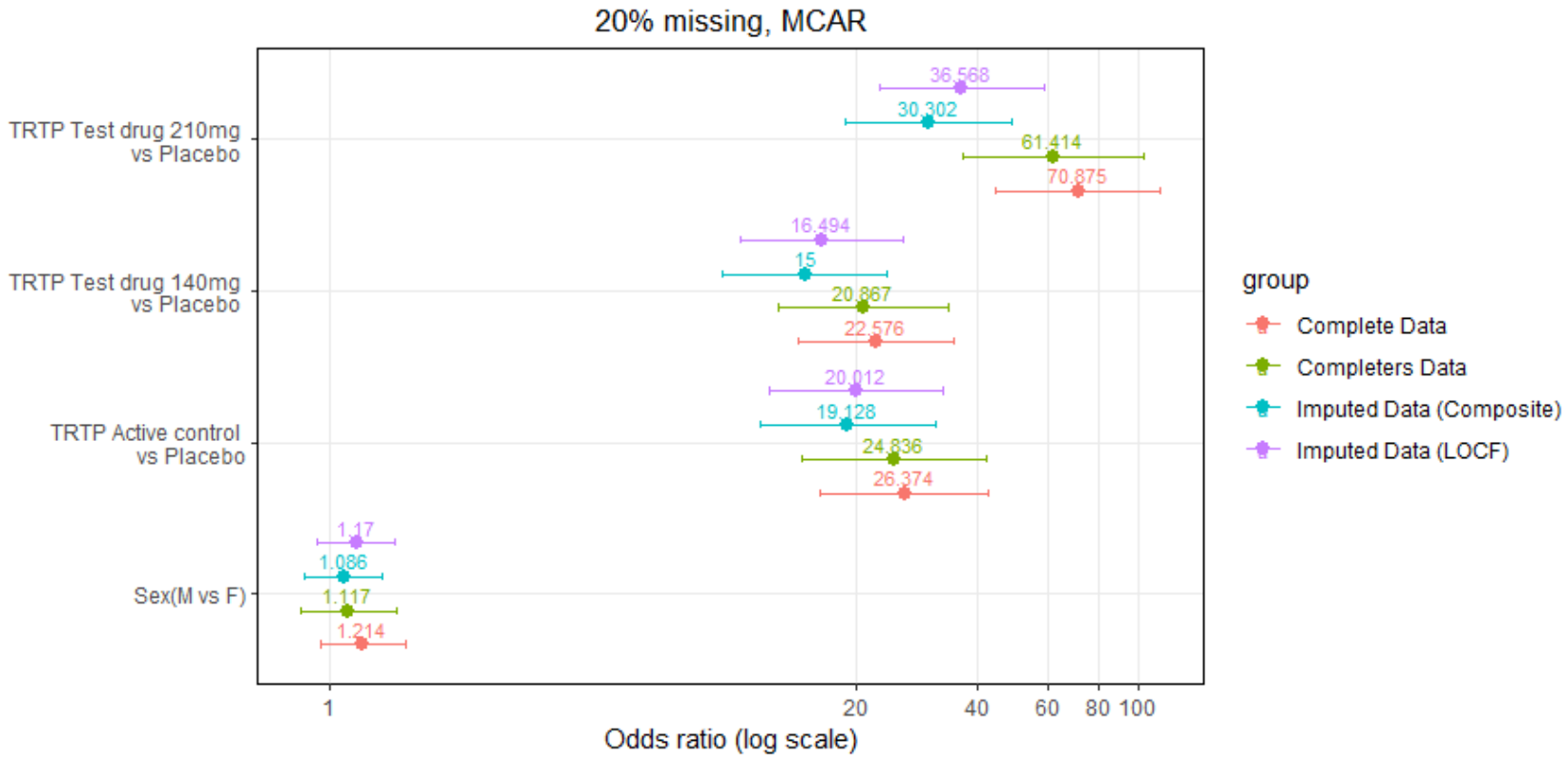
Odds Ratio

MCAR 10% Missing



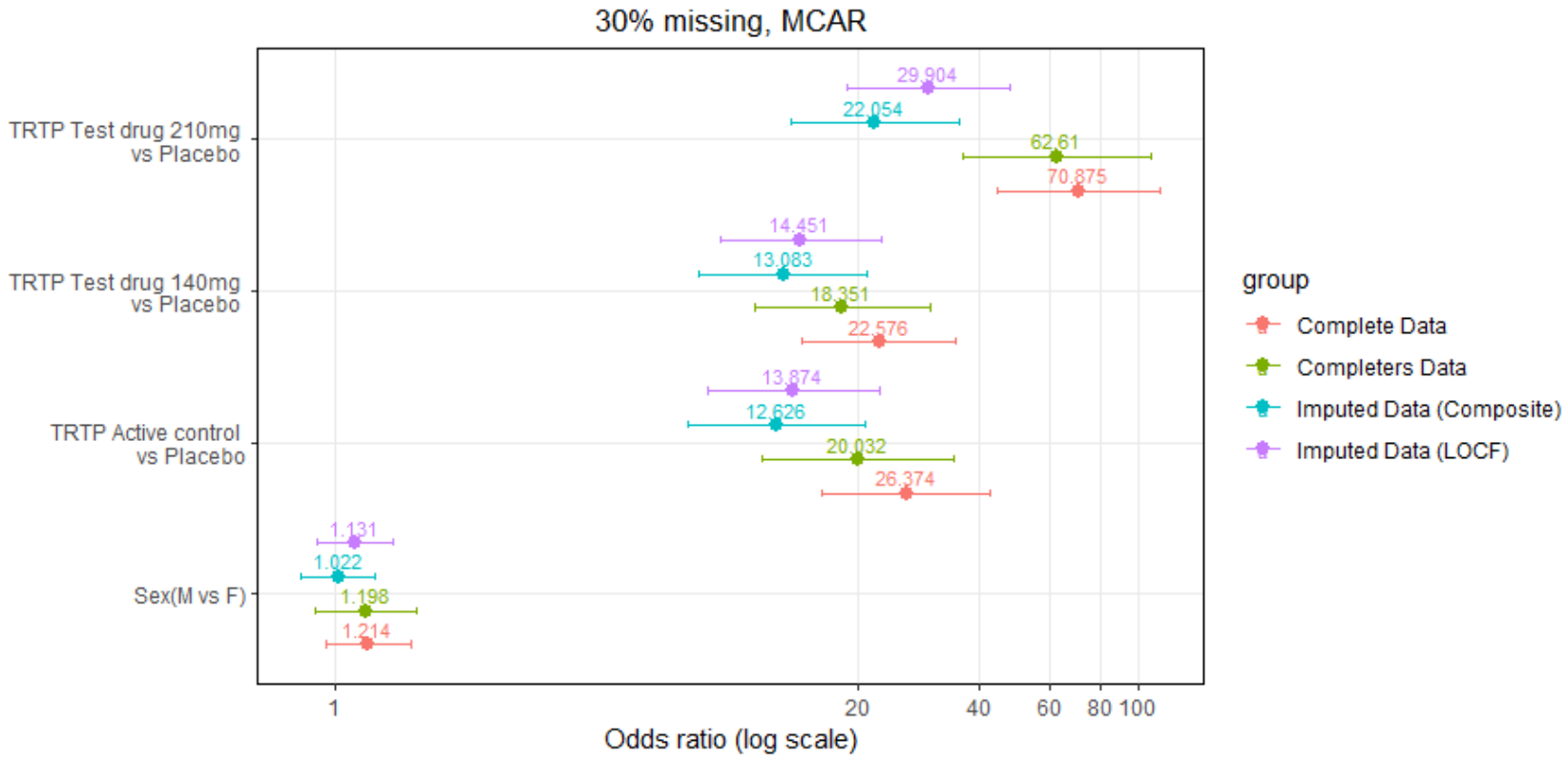
Odds Ratio

MCAR 20% Missing



Odds Ratio

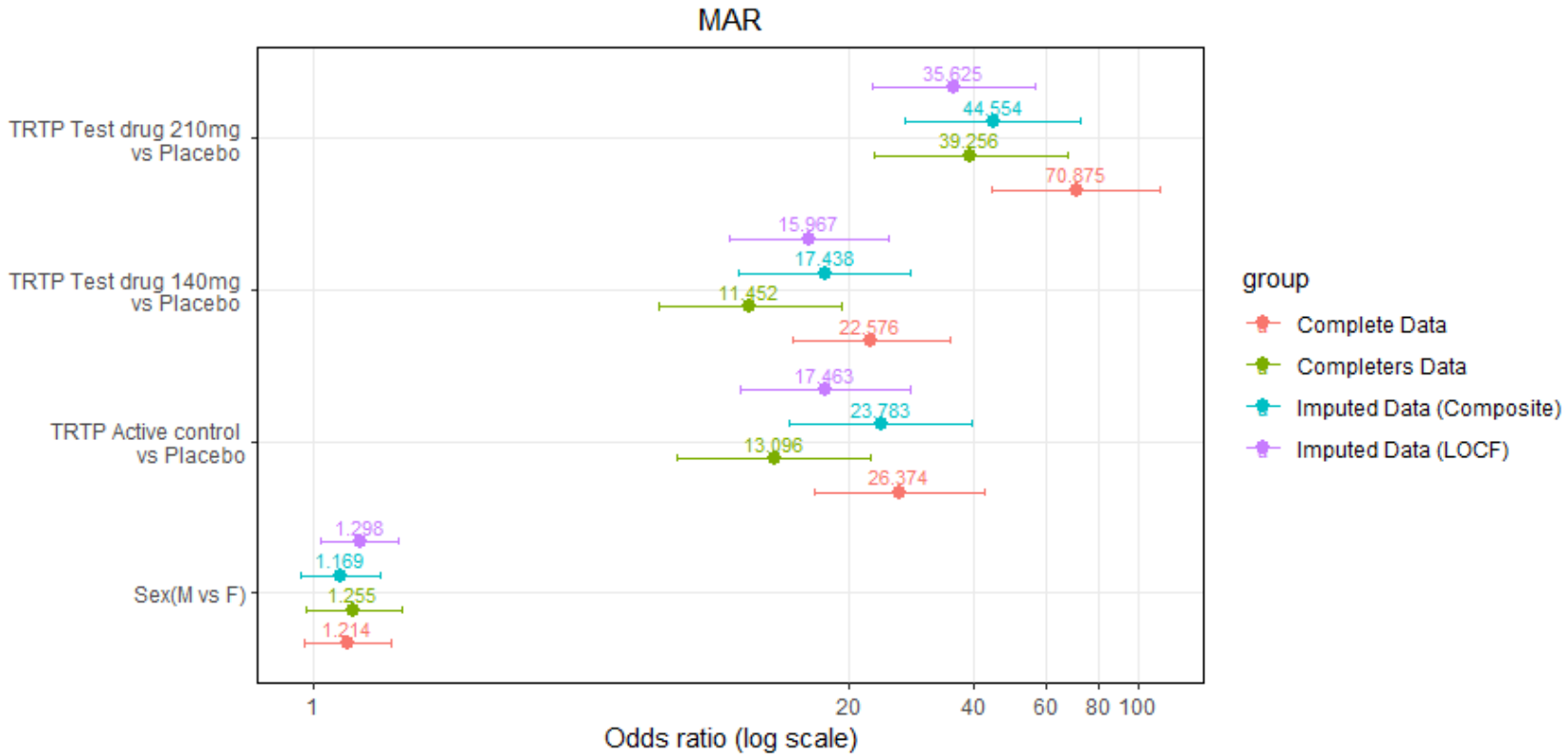
MCAR 30% Missing



Odds Ratio

MAR

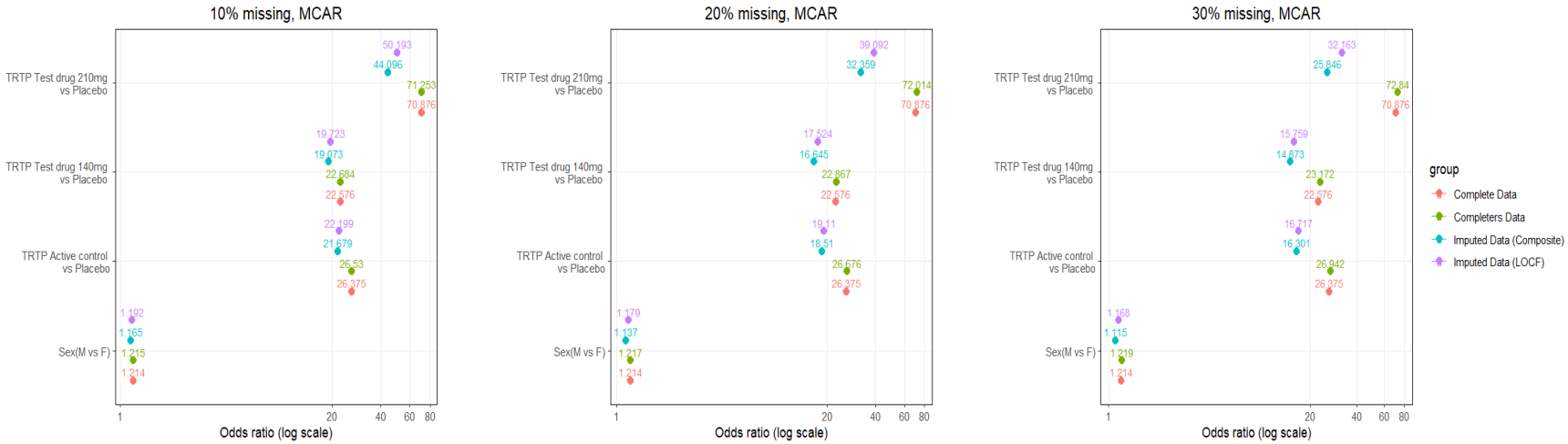
Underestimated



R Shiny Demo

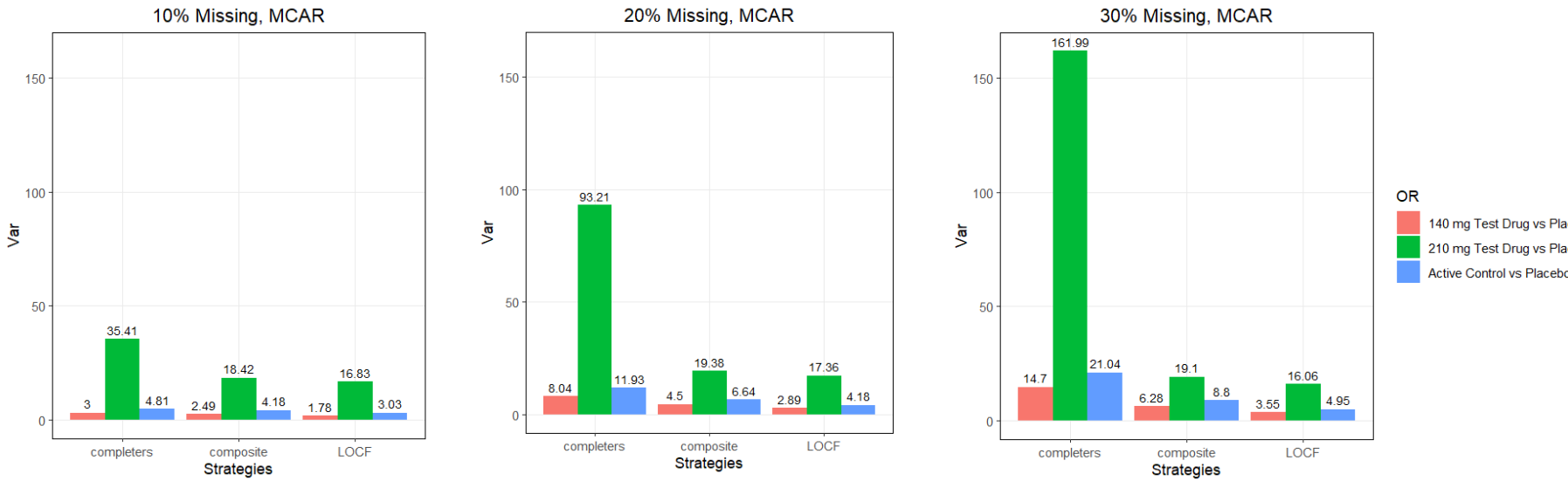
MCAR - Large Sample Results

Mean of Odds Ratios

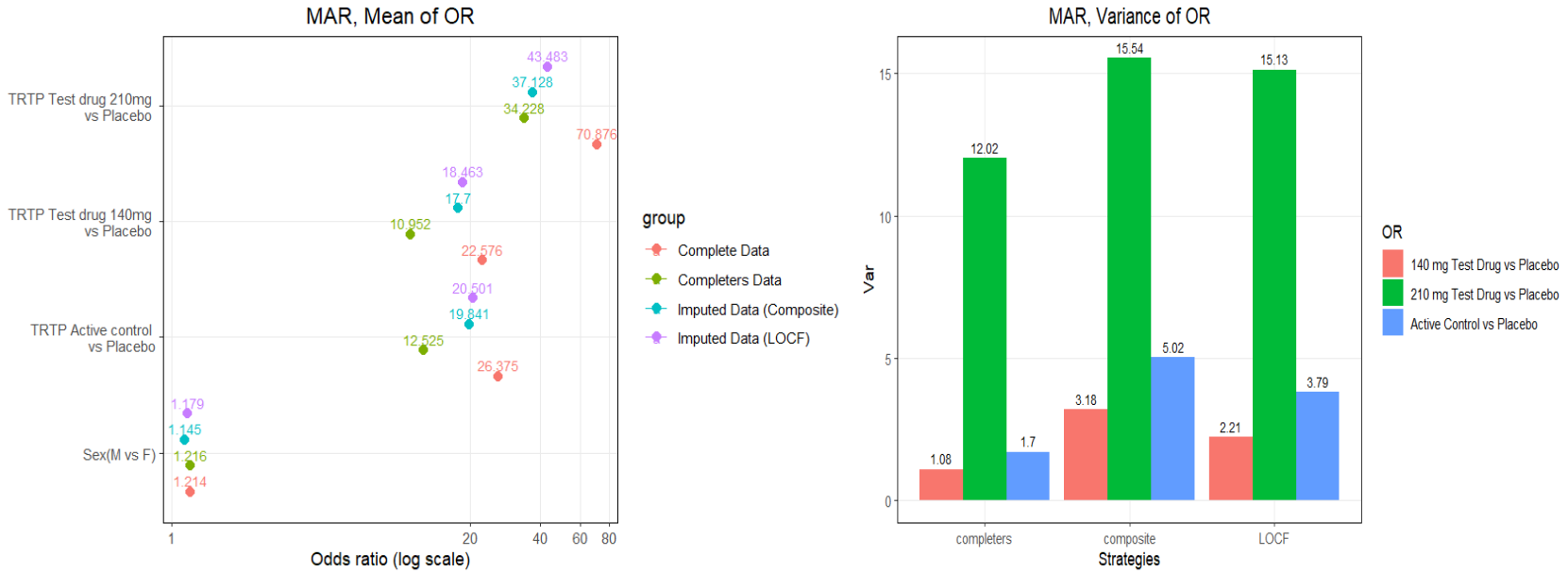


MCAR - Large Sample Results

Variance of Odds Ratios



MAR - Large Sample Results



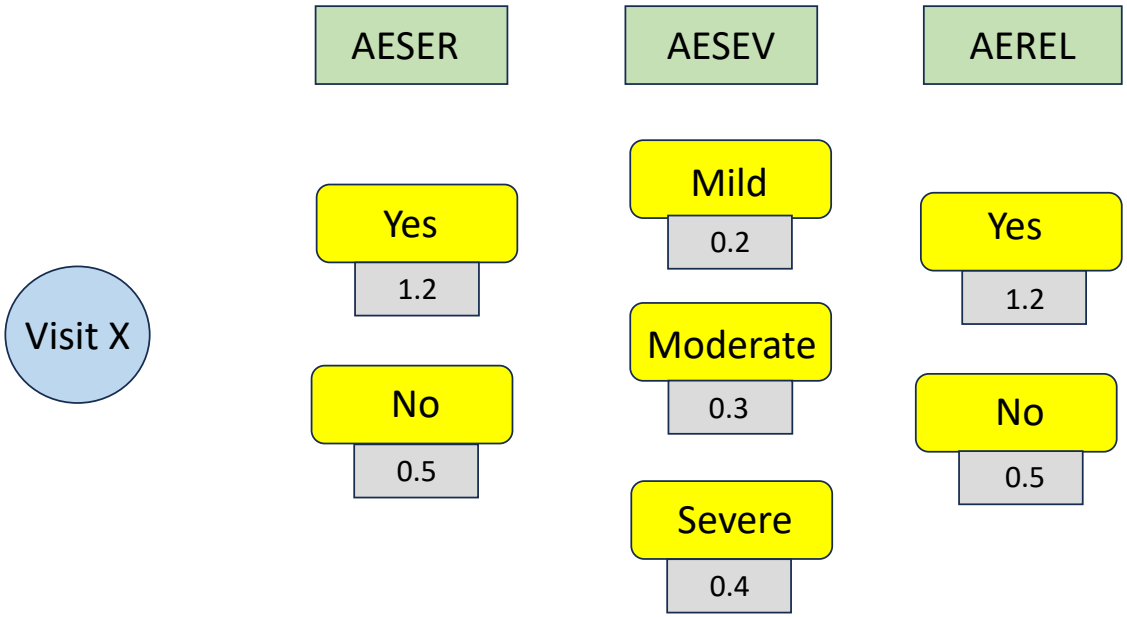
MNAR

- **MNAR(Missing not at random)**: Missingness depends on unobserved predictors.
  1. Toxicity Index based on AE
  2. Improvement in PASI based on visit 6
  3. Cumulative Improvement in PASI
  4. Weight, Age and Baseline

MNAR Scheme 1-Toxicity Index Based on AE

- **Toxicity index**: The initial predictor we consider is potential correlation between "Toxicity Index" and missing data. We hypothesize that the probability of missing data is positively correlated with the underlying "toxicity index". Subjects with higher toxicity may be more likely to experience adverse effects leading to data dropout.
- **AE Factors**: Based on AE data, when patients occur adverse event at their visit day, they have different possibilility to drop out, which depends on AESER (AE serious or not), AESEV (AE severity) and AEREL (Investigator's assessment of whether AE is related to the treatment or not).

Related Factors in AE



Possibility = AESER x AESEV x AEREL

Frequency Tables-MNAR-Toxicity Index

		PASI	
Trtp	Sex	0	1
1	F	89	8
	M	149	13
2	F	24	52
	M	45	113
3	F	64	105
	M	109	227
4	F	20	114
	M	43	287

Completers Data
Total Number: 1462

		PASI	
Trtp	Sex	0	1
1	F	104	8
	M	184	13
2	F	41	52
	M	94	113
3	F	96	105
	M	182	227
4	F	64	114
	M	147	287

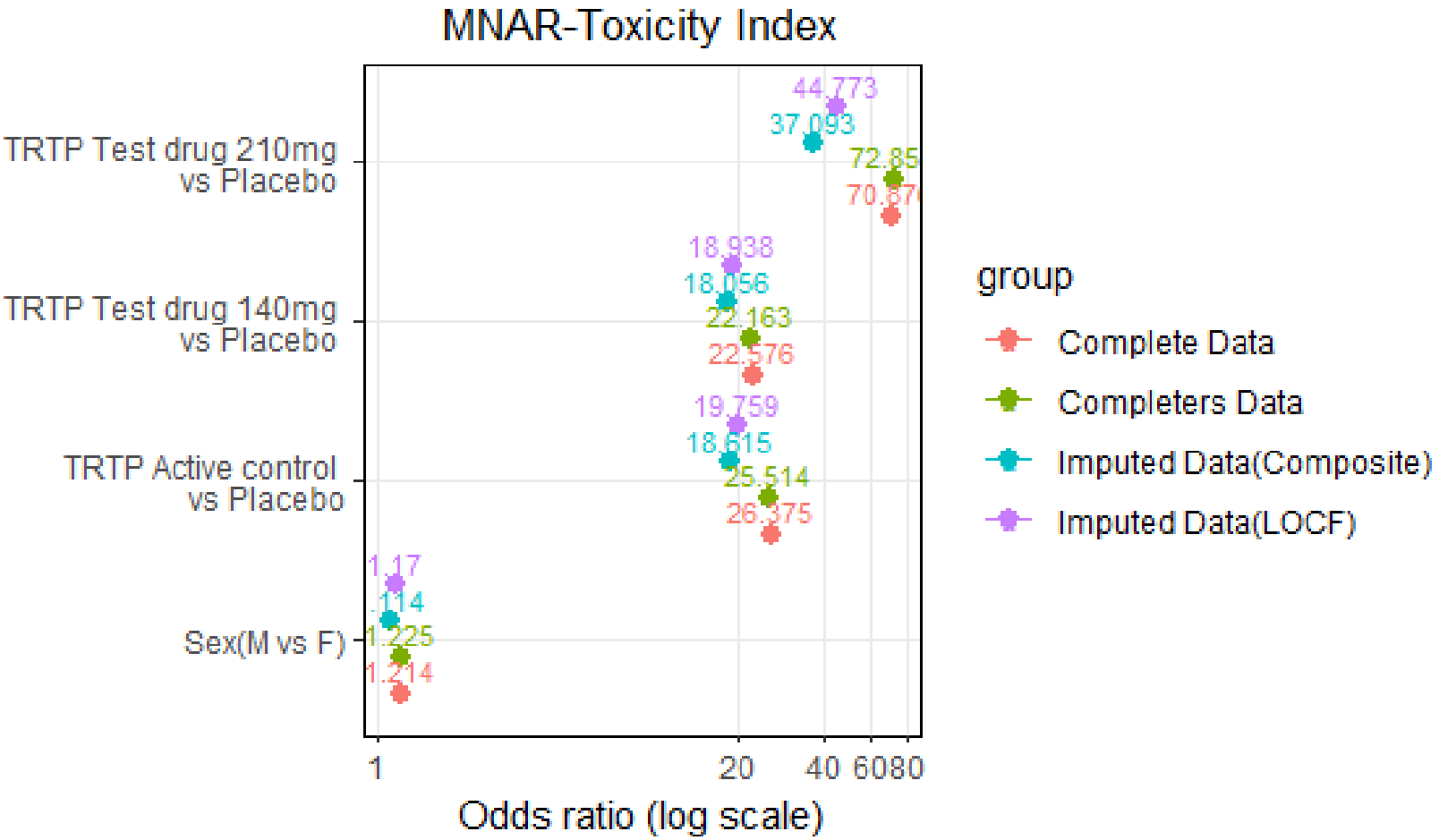
Imputed Data
(Composite)

		PASI	
Trtp	Sex	0	1
1	F	104	8
	M	182	15
2	F	41	52
	M	81	125
3	F	92	109
	M	156	253
4	F	38	140
	M	85	349

Imputed Data
(LOCF)

Odds Ratio

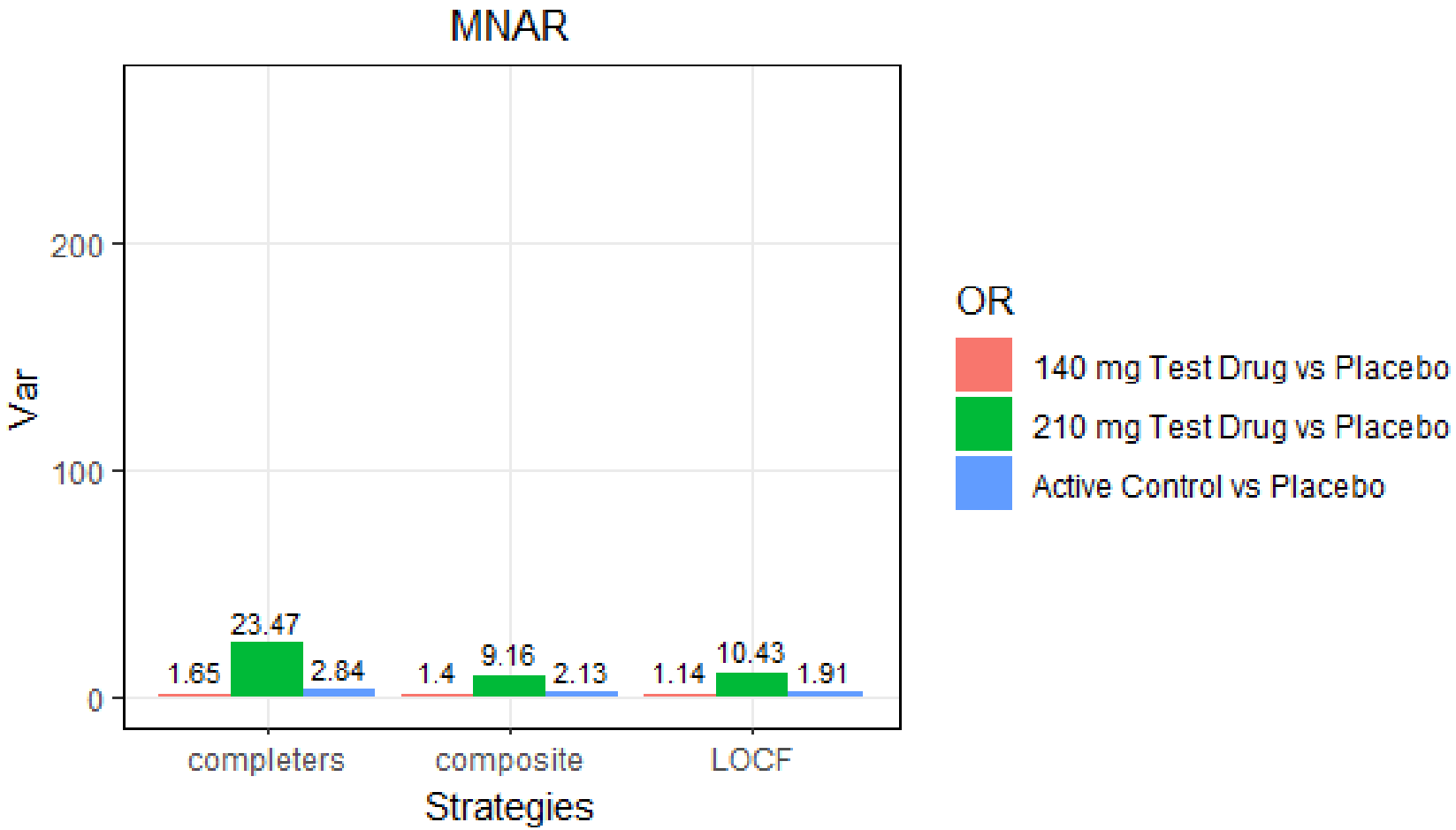
MNAR-Scheme 1 Toxicity Index



Odds Ratio

MNAR-Scheme 1 Toxicity Index

Variance of Odds Ratios



MNAR Scheme 2 - Improvement in PASI

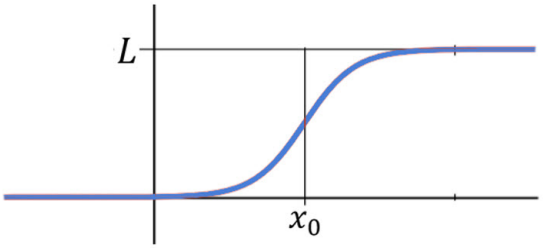
- **Improvement in PASI**: The probability of missing depends on the value of response at visit 6. We hypothesize that as the improvement in PASI increases compared to the baseline, the likelihood of drop out also increases. To model this relationship accurately, we employ a corrected logistic function to simulate the probability of improvement and its effect on drop out.

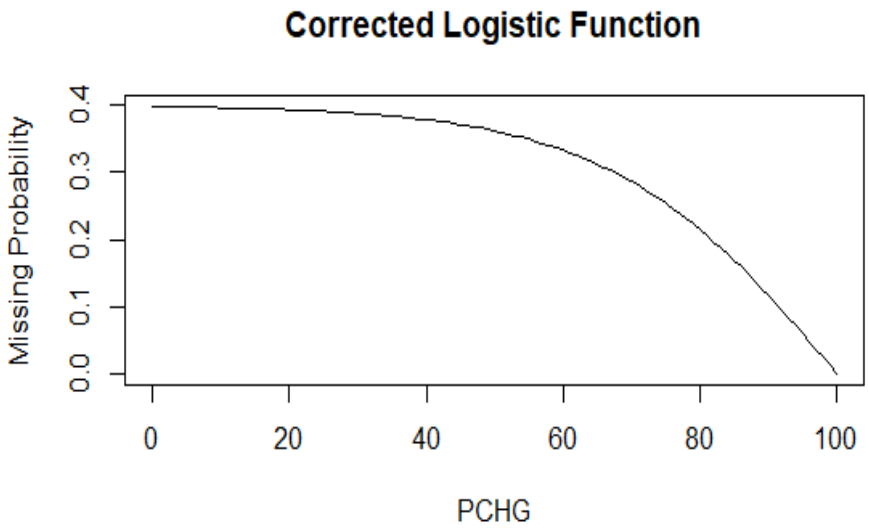
MNAR-Corrected Logistic Function

Logistic Function

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

x_0 = x value of midpoint
L = maximum value
k = growth rate





Formula: $f(x) = 0.8\left(\frac{1}{1+exp(-0.006(100-x))} - 0.5\right)$

Frequency Tables-MNAR-Improvement in PASI

		PASI	
Trtp	Sex	0	1
1	F	67	8
	M	113	15
2	F	22	60
	M	33	118
3	F	55	115
	M	94	257
4	F	20	142
	M	42	327

Completers Data
Total Number: 1488

		PASI	
Trtp	Sex	0	1
1	F	104	8
	M	182	15
2	F	33	60
	M	89	118
3	F	86	115
	M	152	257
4	F	36	142
	M	107	327

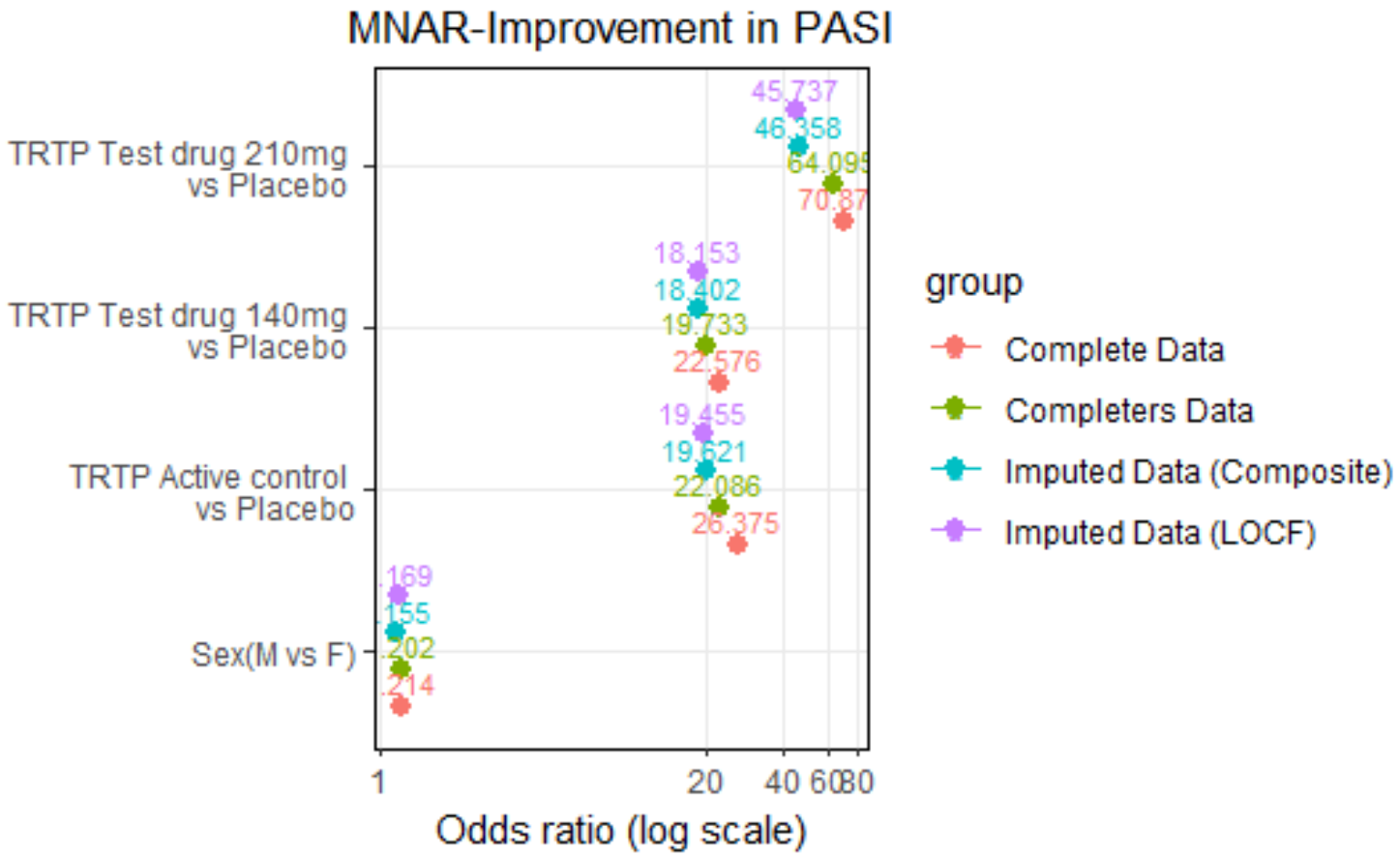
Imputed Data
(Composite)

		PASI	
Trtp	Sex	0	1
1	F	104	8
	M	182	15
2	F	33	60
	M	88	119
3	F	86	115
	M	149	260
4	F	36	142
	M	106	328

Imputed Data
(LOCF)

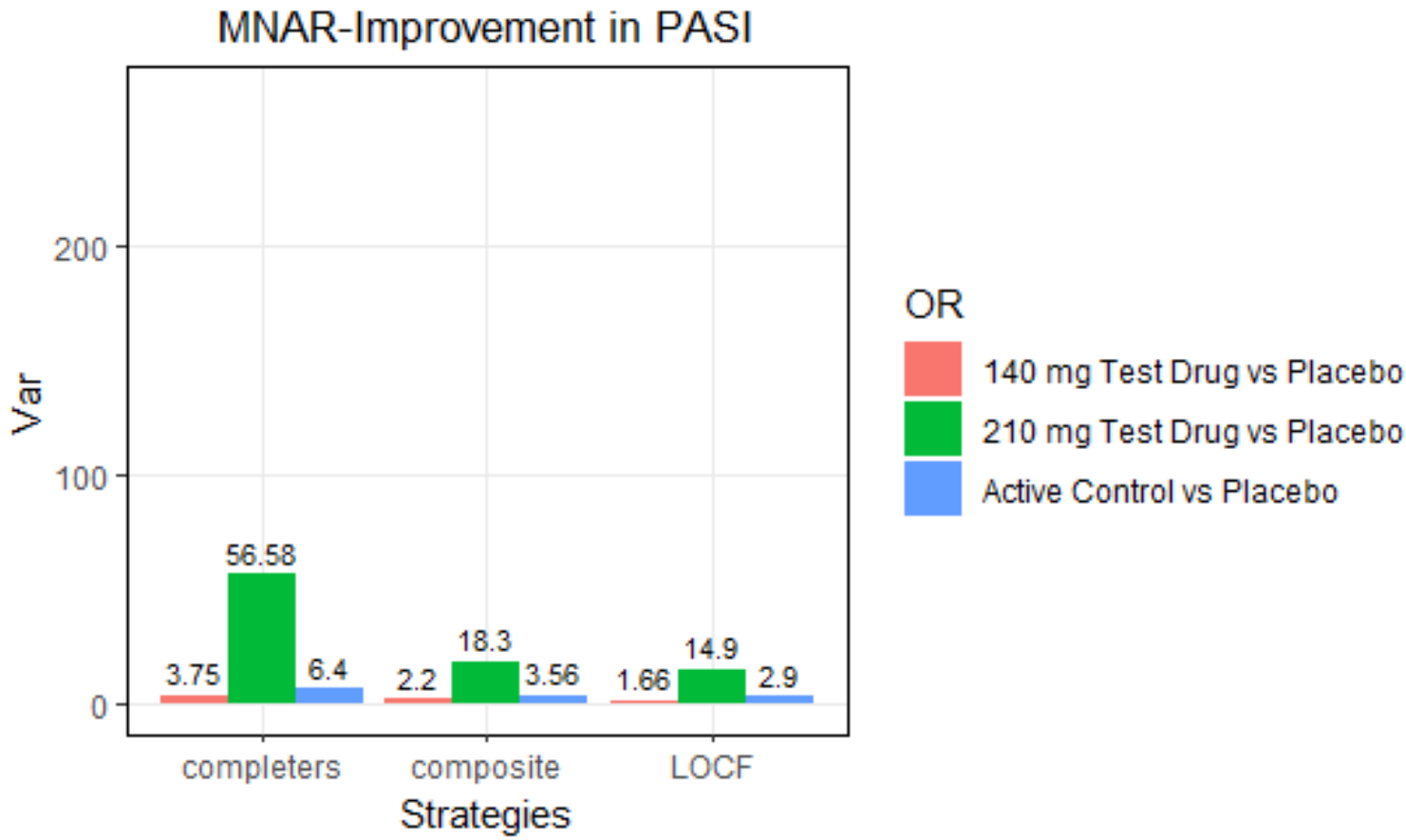
Odds Ratio

MNAR Scheme 2 - Improvement in PASI



Odds Ratio

MNAR Scehem 2 - Improvement in PASI



MNAR Scheme 3-Improvement in PASI(Cumulative)

- **Improvement in PASI(Cumulative)**: Just as in the Improvement in PASI, when we analyze the improvements observed during visits 3 to 6, we notice a trend that a more significant improvement tends to be associated with a higher likelihood of dropout. It's important to note that once a dropout occurs, any subsequent occurrences are also considered as dropouts.

Frequency Tables-MNAR-Improvement in PASI(Cumulative)

		PASI	
Trtp	Sex	0	1
1	F	83	10
	M	131	12
2	F	20	53
	M	44	116
3	F	66	103
	M	98	229
4	F	18	125
	M	40	301

Completers Data
Total Number: 1449

		PASI	
Trtp	Sex	0	1
1	F	102	10
	M	185	12
2	F	40	53
	M	91	116
3	F	98	103
	M	180	229
4	F	53	125
	M	133	301

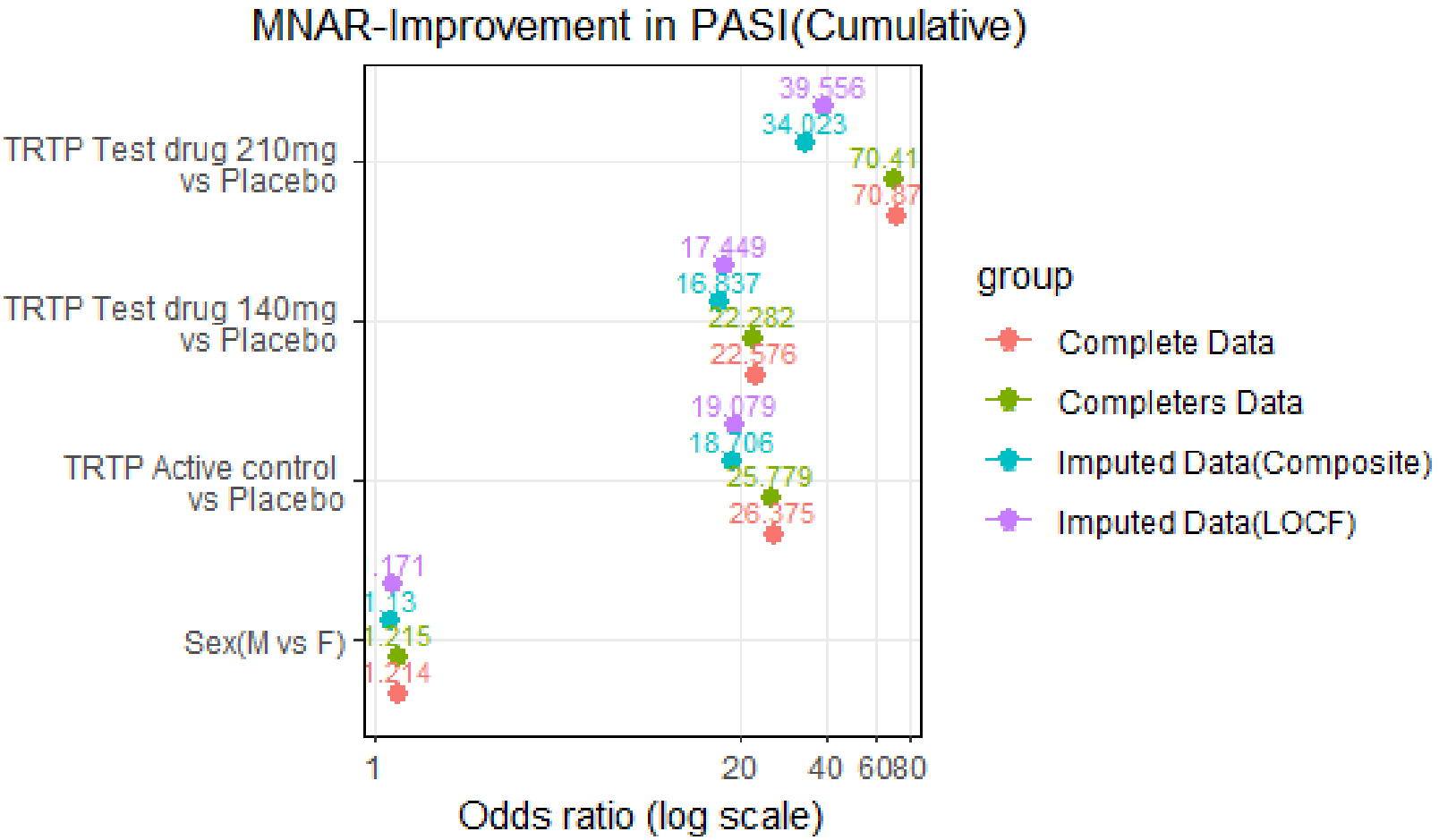
Imputed Data
(Composite)

		PASI	
Trtp	Sex	0	1
1	F	102	10
	M	185	12
2	F	40	53
	M	85	122
3	F	95	106
	M	168	241
4	F	49	129
	M	110	324

Imputed Data
(LOCF)

Odds Ratio

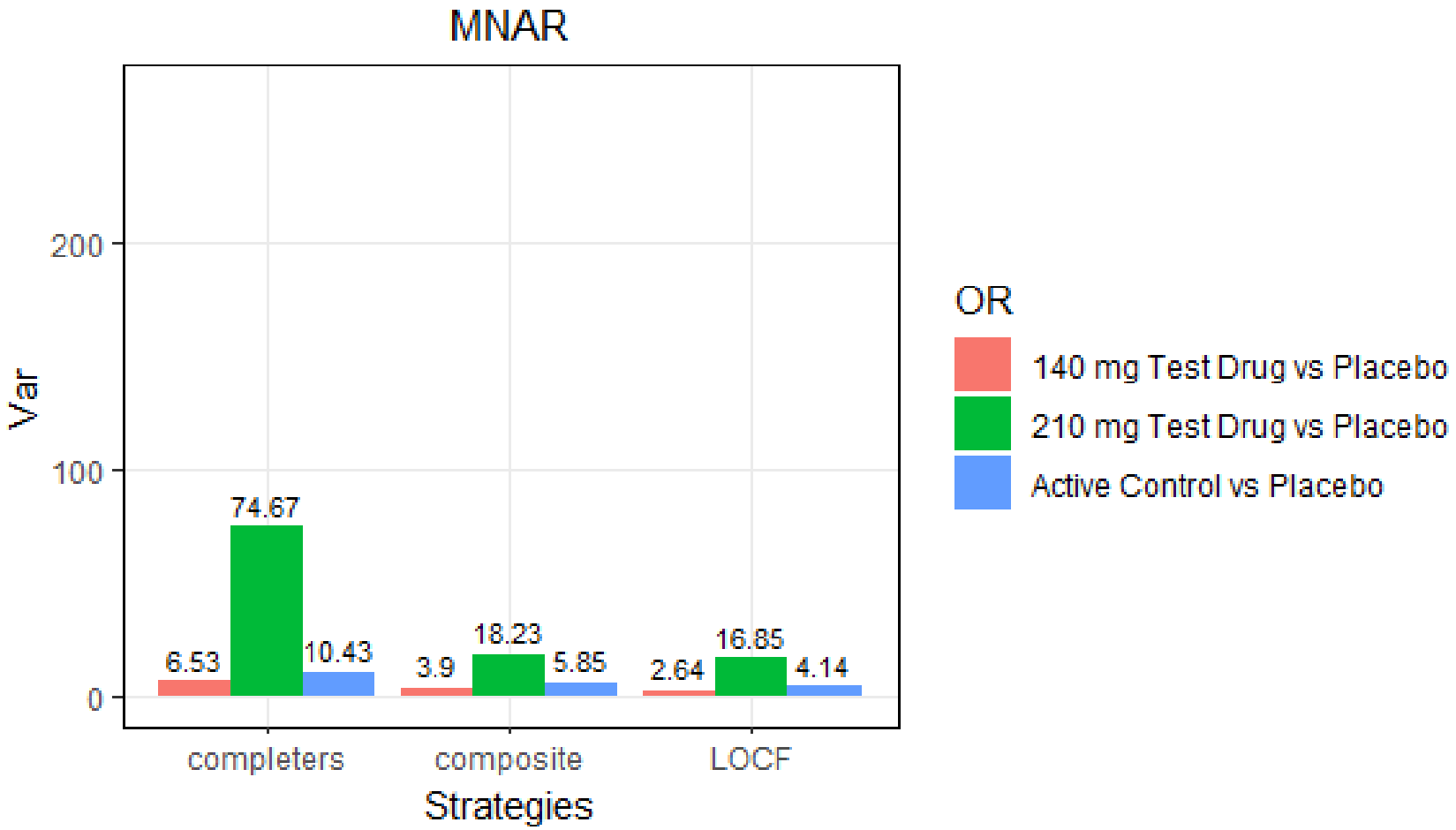
MNAR-Scheme 3 Improvement in PASI(Cumulative)



Odds Ratio

MNAR-Scheme 3 Improvement in PASI(Cumulative)

Variance of Odds Ratios



MNAR Scheme 4 - Weight, Age and Baseline

We implemented a weighted index for this MNAR scheme: Weight (0.3), Baseline Pasi Score (0.3), Age (0.4).

- **Weight:** individuals with a weight greater 100 kg demonstrate a higher dropout ratio compared to those below 100 kg. We set the probability of dropout for individuals with a weight greater 100 kg as 50% and 10% for those under 100 kg.
- **Age:** In terms of age, it has been observed that younger patients are the most prone to discontinuation, followed by elder patients. We set the probability of dropout for patients under 20 years old to be 50%, for patients between 60 and 80 years old to be 40%, for patients between 40 and 60 years old to be 10%, and for patients between 20 and 40 years old to be 5%.
- **Baseline:** Based on the initial PASI score, a lower score indicates milder symptoms, which in turn increases the likelihood of discontinuation. We set the probability of dropout to be $\frac{2}{\text{Baseline Pasi score}}$

Frequency Tables-MNAR - Weight Age and Baseline

		PASI	
Trtp	Sex	0	1
1	F	49	6
	M	95	12
2	F	13	77
	M	28	116
3	F	46	62
	M	70	161
4	F	11	83
	M	30	189

Completers Data
Total Number: 958

		PASI	
Trtp	Sex	0	1
1	F	107	5
	M	191	6
2	F	60	33
	M	130	77
3	F	139	62
	M	248	161
4	F	95	83
	M	245	189

Imputed Data
(Composite)

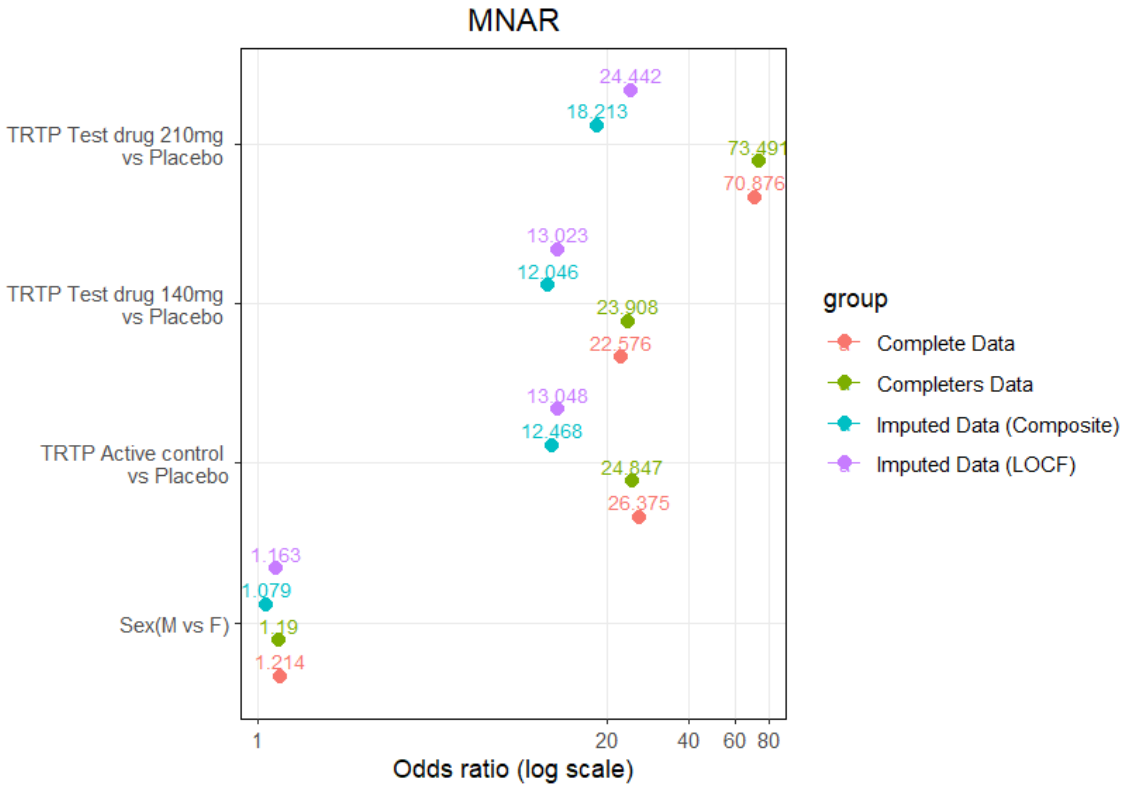
		PASI	
Trtp	Sex	0	1
1	F	107	5
	M	187	10
2	F	55	38
	M	115	92
3	F	126	75
	M	220	189
4	F	78	100
	M	178	256

Imputed Data
(LOCF)

Odds Ratio

MNAR Scheme 4-Weight, age and baseline

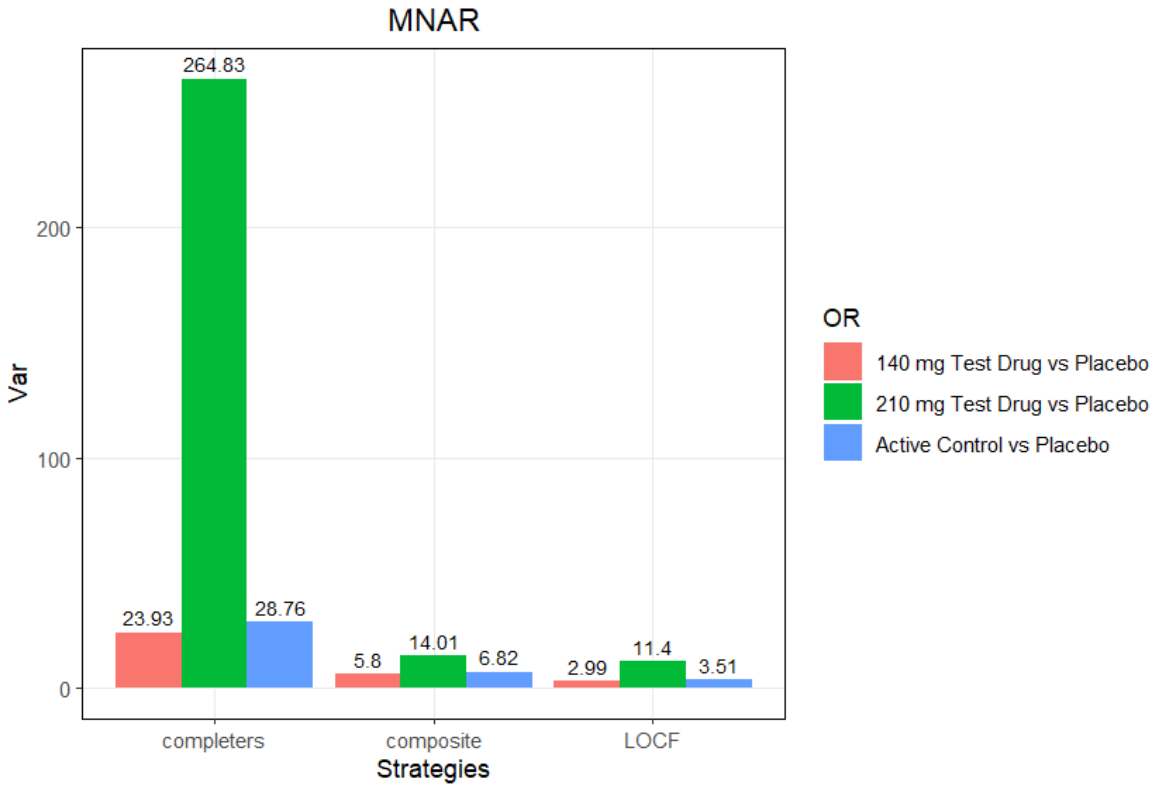
Mean of Odds Ratios



Odds Ratio

MNAR Scheme 4-Weight, age and baseline

Variance of Odds Ratios



Conclusions

MCAR

1. Including only completers yields the most accurate result.
2. As the level of missingness increases, the variance of OR increases.

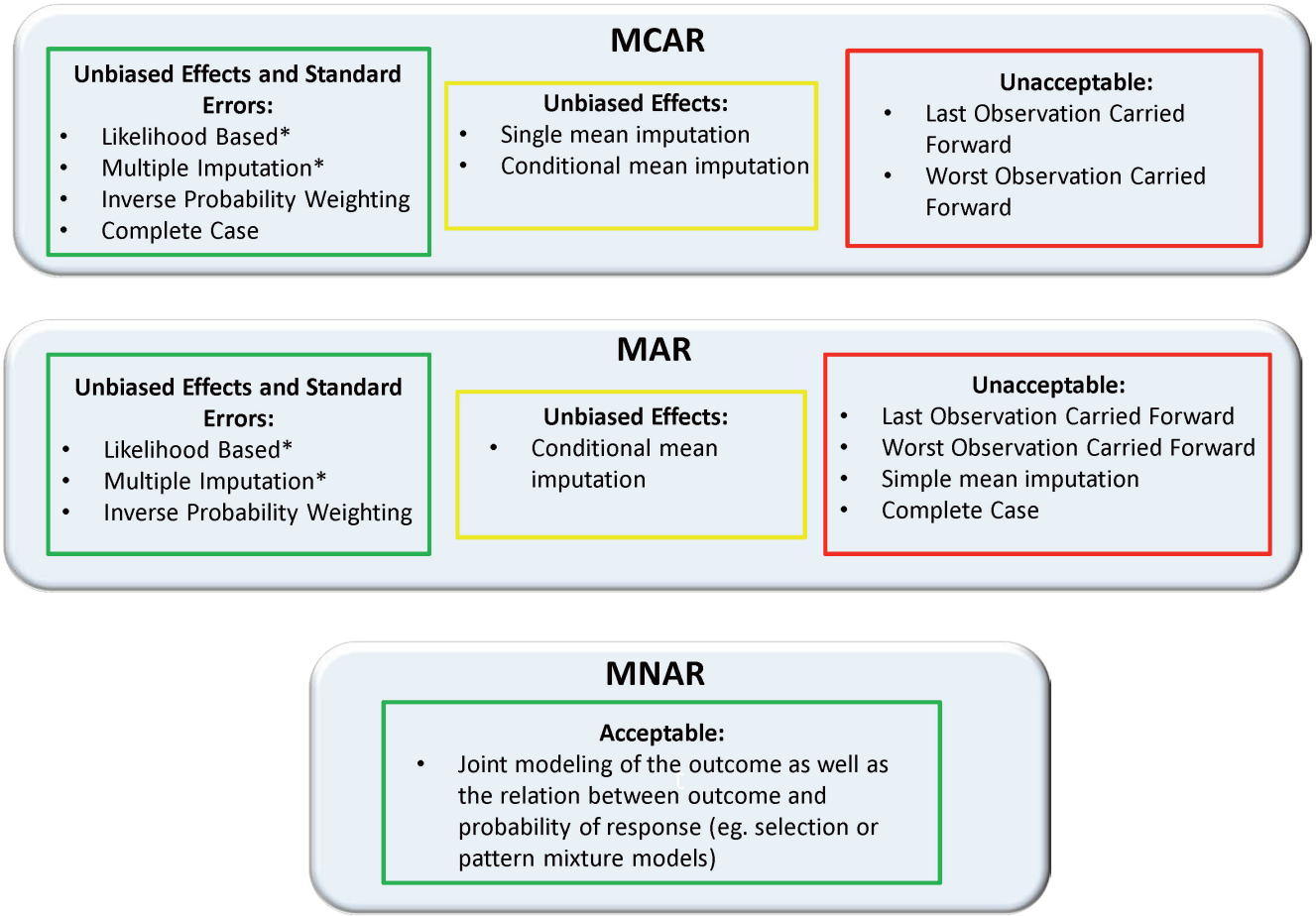
MAR

1. Imputation with LOCF yields the most accurate result, followed by imputation with composite strategy. The variance of the result is the highest
2. Including only completers underestimates the treatment effect size, but it yields the smallest variance of OR.

MNAR

1. Including only completers yields the most accurate result for all four MNAR schemes.
2. However, the variance of OR is the highest when we include only completers.

Conclusions Cont.



References

1. Dziura et al. *Strategies for dealing with Missing data in clinical trials: From design to Analysis*, Yale Journal of Biology and Medicine 86 (2013), pp.343-358.
2. Little et al. *The Prevention and Treatment of Missing Data in Clinical Trials*. The New England Journal of Medicine (2012). Special Report.
3. Little and Rubin. *Statistical Analysis with Missing Data, Second Edition*. Wiley (2002).
4. Mallinckrodt et al.(2020) *Estimands, Estimators and Sensitivity Analysis in Clinical Trials*. Routledge.
5. Schmidli. *Causal Reasoning and Strategies for Defining Estimands*. Novartis.