Handling Missing Data in Health Science Research

Day 1 - Part I

2022-06-21

Contents

About this workshop

Instructor

Aya Mitani, PhD, MPH Assistant Professor Division of Biostatistics Dalla Lana School of Public Health Website

Student Assistant

Mei Dong, MSc PhD student in Biostatistics Dalla Lana School of Public Health Website

Overview

Day I (2022 June 21 1-3p)

- Introduction
  - Missing data in health research
  - Missing data type
  - Missing data mechanism
- Understanding missingness in data
  - Visual inspection
  - Can it be ignored?

- Missing data in cross-sectional data
  - Single-level multiple imputation (MI)
  - MI of interaction variables

Day II (2022 June 23 1-3p)

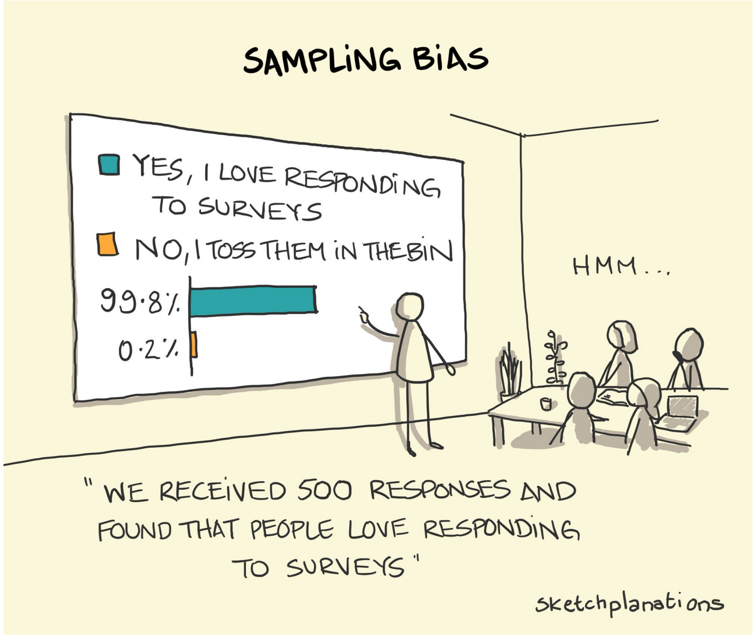
- Missing data in longitudinal data
  - IPW for drop-out in longitudinal data
  - Multiple imputation for multilevel data
- Methods for nonignorable missing data
- Consideration for fairness in missing data imputation

Housekeeping

- The workshop will **not** be recorded
- We will have a **10-minute break** in the middle of each session
- Ask questions!
  - Feel free to interrupt me by **raising your hand**
  - Put questions in the **chat** (Mei will field most of the questions)
  - Wait until the **end of each session** (I will stick around for a few minutes after 3p)

Missing data in health research

- It is **very** common for sets of quantitative data to be incomplete
  - i.e. Not all **planned** observations are actually made
- This is especially true when studies are conducted on **human** subjects, e.g.,
  - Epidemiologic studies
  - Clinical trials
  - Sample surveys
- Missing data can lead to **bias** and **lack of precision** in the analyses
- The default method is complete-case analysis or listwise deletion
  - Observations with at least one variable missing is deleted from the analysis


SAMPLING BIAS
🟦 YES, I LOVE RESPONDING TO SURVEYS
🟧 NO, I TOSS THEM IN THE BIN
99.8%
0.2%
HMM...
"WE RECEIVED 500 RESPONSES AND FOUND THAT PEOPLE LOVE RESPONDING TO SURVEYS"
sketchplanations

Reasons for missing data

- Item nonresponse
  - Respondent skips a question because they don't want to provide an answer
- Unit nonresponse
  - Refusal to complete the survey (see above cartoon)
- Data entry or coding error
  - Frequently occurs in administrative data
- Loss to follow-up
  - Informative or non-informative censoring
- By design
  - Bad survey design (unintentional)
  - Sampling (intentional)
- Many more

> Obviously the best way to treat missing data is not to have them. [Orchard and Woodbury (1972, p.697)]

Be explicit about missing data

- Researchers tend to downplay the issue of missing data
- Best practices when dealing with missing data
  - The presence of missing data should be explicitly stated in the text
  - If using default methods (listwise deletion or complete-case analysis), then mention it
  - Make sure all your tables have the same sample sizes
  - Use model-based missing data methods
- Not being explicit about missing data negatively affects the **reproducibility**, **replicability**, and **reliability** of your research

Example data: The Skin Cancer Prevention Study

The data are from a randomized, double-blind, placebo-controlled clinical trial of beta-carotene to prevent non-melanoma skin cancer in high risk subjects. Subjects were randomized to either receive placebo or 50mg of beta-carotene per day for 5 years. Subjects were examined once a year and biopsied if a cancer was suspected to determine the number of new skin cancers occurring since the last exam. The outcome variable is a count of the number of new skin cancers per year. The dataset contains 1683 observations. (https://pubmed.ncbi.nlm.nih.gov/2666024/)

Variable	Category	Explanation
`ID`	categorical	ID of participant
`center`	categorical	study center (1-4)
`age`	numeric	age (in years) at baseline
`skin`	binary	skin type, 1=burns, 0=otherwise
`gender`	categorical	1=male, 0=female
`exposure`	count	number of previous skin cancers
`treatment`	categorical	1=beta-carotene, 0=placebo
`year`	numeric	year of follow-up
`Y`	count	number of new skin cancers per year

```r
skin <- read.table("skin_data.txt", header = TRUE)
head(skin, 10)
```

ID	center	age	skin	gender	exposure	Y	treatment	year
100034	1	51	1	1	4	0	0	1
100034	1	51	1	1	4	1	0	2
100034	1	51	1	1	4	1	0	3
100034	1	51	1	1	4	1	0	4
100034	1	51	1	1	4	0	0	5
100045	1	68	1	0	2	0	0	1
100045	1	68	1	0	2	0	0	2
100045	1	68	1	0	2	0	0	3
100045	1	68	1	0	2	0	0	4
100045	1	68	1	0	2	0	0	5

```
tail(skin, 10)
```

ID	center	age	skin	gender	exposure	Y	treatment	year
420645	4	67	1	0	1	0	1	3
420656	4	41	1	1	1	0	1	1
420656	4	41	1	1	1	0	1	2
420656	4	41	1	1	1	0	1	3
420678	4	66	0	1	1	0	1	1
420678	4	66	0	1	1	0	1	2
420678	4	66	0	1	1	1	1	3
420689	4	62	0	0	1	0	1	1
420689	4	62	0	0	1	0	1	2
420689	4	62	0	0	1	0	1	3

Missing data in cross-sectional data

Notation

- $\boldsymbol{Y}$ is the $n \times p$ matrix containing the data values on p variables for n people
- $\boldsymbol{R}$ is the $n \times p$ matrix containing the response indicator (0 or 1)
  - If Y_{ij} is observed then $R_{ij} = 1$
  - If Y_{ij} is not observed then $R_{ij} = 0$

For example, if we have a data set containing 5 people and 4 variables that look like

id	v1	v2	v3	v4
1	50	1	5	1
2	65	1		0
3	48	1		
4	45	1	9	1
5	51	1	2	

Then,

$$\boldsymbol{Y} = \begin{pmatrix} 50 & 0 & 5 & 1 \\ 65 & 1 & . & 0 \\ 48 & 0 & . & . \\ 45 & 1 & 9 & 1 \\ 51 & 1 & 2 & . \end{pmatrix} \qquad \boldsymbol{R} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

- Further $\boldsymbol{Y} = (\boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}})$ contain the hypothetically complete data, where
  - $\boldsymbol{Y}_{\text{obs}}$ is the set of observed data
  - $\boldsymbol{Y}_{\text{mis}}$ has real values, but the values themselves are masked from us
  - Missingness indicators hid the true values that are **meaningful for analysis**

Missing data mechanism

- A hierarchy of **three** different types of missing data mechanisms can be distinguished by considering how $\boldsymbol{R}$ is related to $\boldsymbol{Y} = (\boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{mis}})$

Missing completely at random (MCAR)

- Probability that responses are missing does not depend on $\boldsymbol{Y}_{\text{obs}}$ or $\boldsymbol{Y}_{\text{mis}}$
- $\Pr(R = 0|Y_{\text{obs}}, Y_{\text{mis}}) = \Pr(R = 0)$
- The observed values can be thought of as a **random sample** of the full data

Missing at random (MAR)

- Probability that responses are missing may depend on observed information, $\boldsymbol{Y}_{\text{obs}}$
- $\Pr(R = 0|Y_{\text{obs}}, Y_{\text{mis}}) = \Pr(R = 0|Y_{\text{obs}})$

Missing not at random (MNAR)

- Probability that responses are missing depends on the set of observed information, $\boldsymbol{Y}_{\text{obs}}$, and the values that should have been obtained, $\boldsymbol{Y}_{\text{mis}}$
- $\Pr(R = 0|Y_{\text{obs}}, Y_{\text{mis}})$ does not simplify

- MCAR and MAR are often referred to as **ignorable** mechanisms
- MNAR is referred to as **nonignorable** mechanism
- Several tests exist to distinguish between MCAR and MAR, but they are not widely used and their practical value is unclear
  - Enders (2010, p. 17-21) evaluates two procedures
  - It is not possible to test MAR versus MNAR because the information that is needed for such a test is missing!

Year 1 data of skin cancer study

```r
skinbl <- read.table("skin_FULL.txt", header = TRUE)
head(skinbl, 10)
```

ID	center	age	skin	gender	exposure	Y	treatment	year
100034	1	51	1	1	4	0	0	1
100045	1	68	1	0	2	0	0	1
100056	1	58	1	0	7	1	0	1
100067	1	53	1	1	3	0	0	1
100102	1	55	0	0	2	0	0	1
100113	1	59	1	1	10	0	0	1
100124	1	56	0	1	5	0	0	1
100146	1	59	0	0	2	1	0	1
100214	1	64	1	1	2	0	0	1
100236	1	69	0	1	1	1	0	1

```
tail(skinbl, 10)
```

ID	center	age	skin	gender	exposure	Y	treatment	year
420410	4	63	1	1	11	0	1	1
420476	4	63	1	0	1	0	1	1
420487	4	38	0	1	1	0	1	1
420511	4	40	1	1	1	0	1	1
420555	4	47	0	1	2	0	1	1
420634	4	64	0	0	3	1	1	1
420645	4	67	1	0	1	0	1	1
420656	4	41	1	1	1	0	1	1
420678	4	66	0	1	1	0	1	1
420689	4	62	0	0	1	0	1	1

- For this workshop, we generated **three** different cross-sectional data sets
- The variables `age` and `skin` were simulated to be MCAR, MAR, and MNAR for each data set

Missing data mechanism	Description
Full	No missing data, all data are observed
MCAR	`age` or `skin` is missing completely at random for some individuals
MAR	Missingness of `age` and `skin` are related to all other covariates, but not with the outcome Y
MNAR	`age` is missing for all individuals with true age > 55 years `skin` is missing in 20% of individuals with true skin = 0

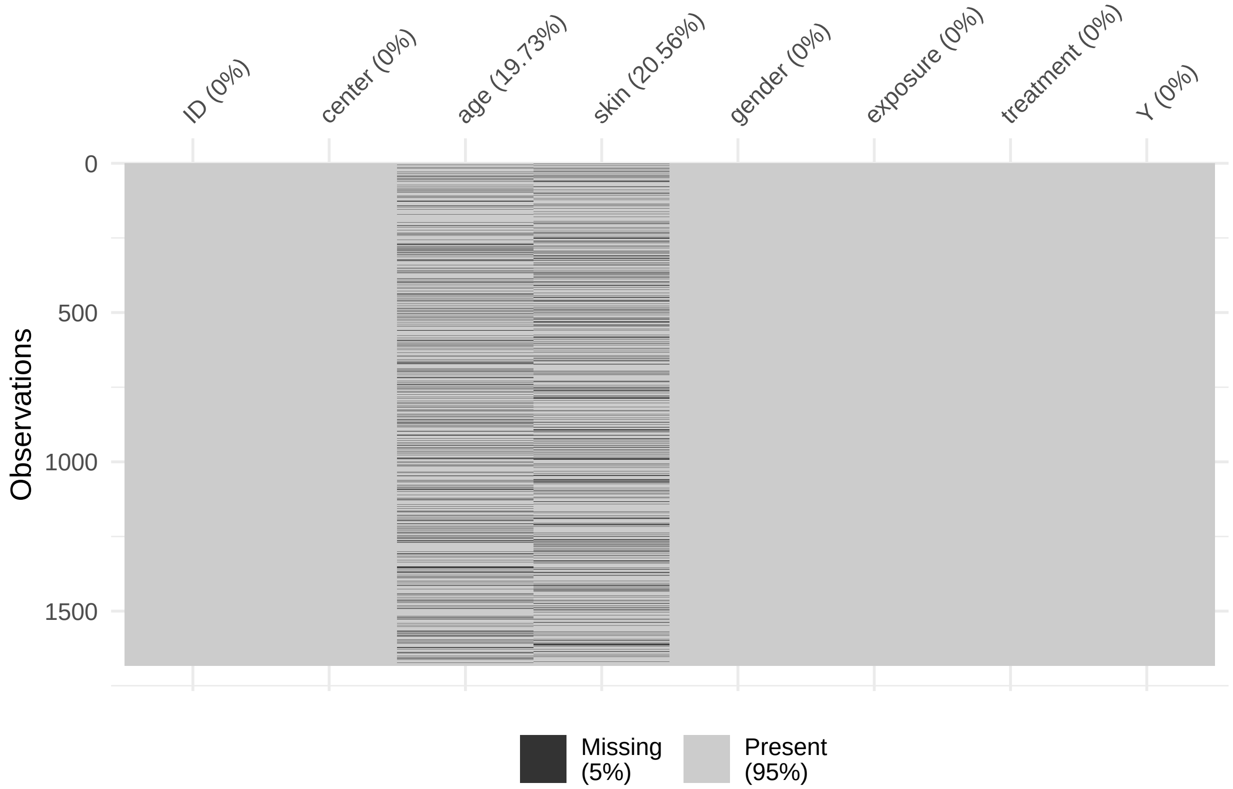
Missing data visualization

- Many packages in R are available to describe the prevalence and pattern of missing data
- We will go over three packages

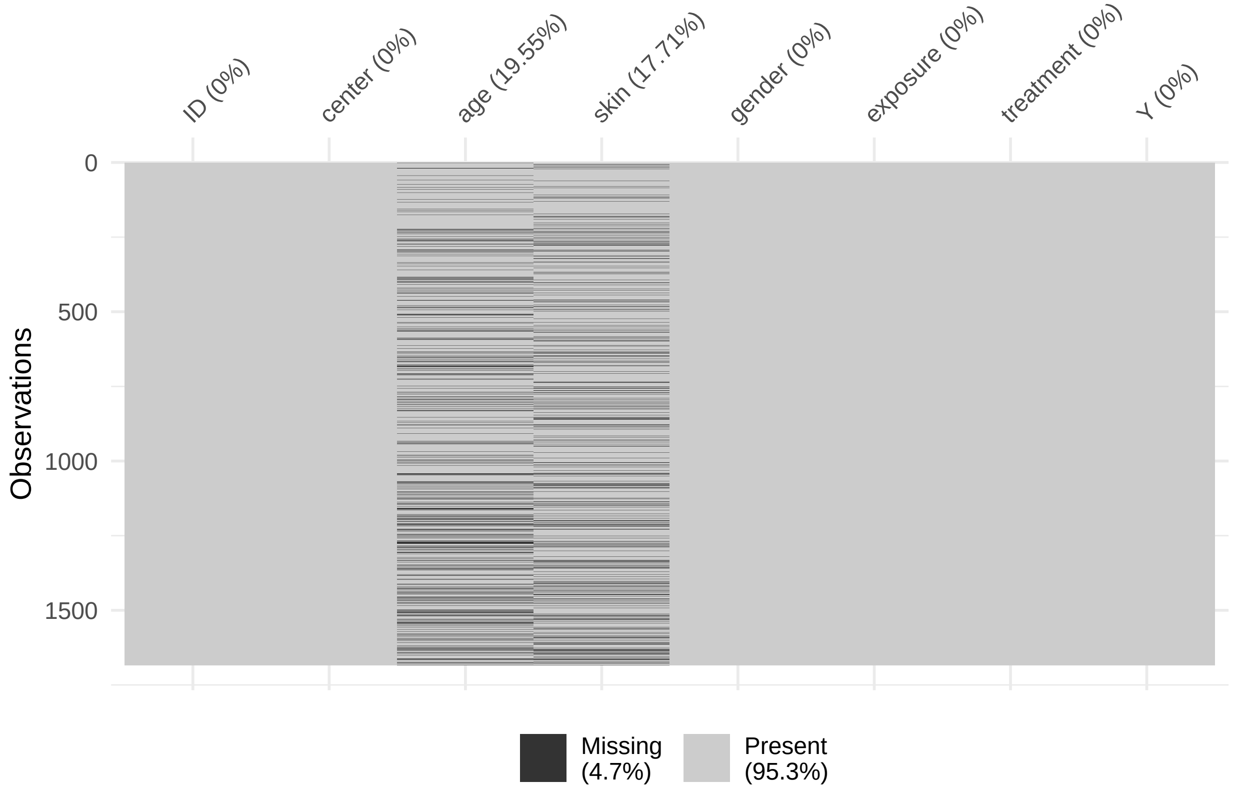
```
library(naniar)
library(VIM)
library(finalfit)
```

Heatplot of missingness across the entire data frame

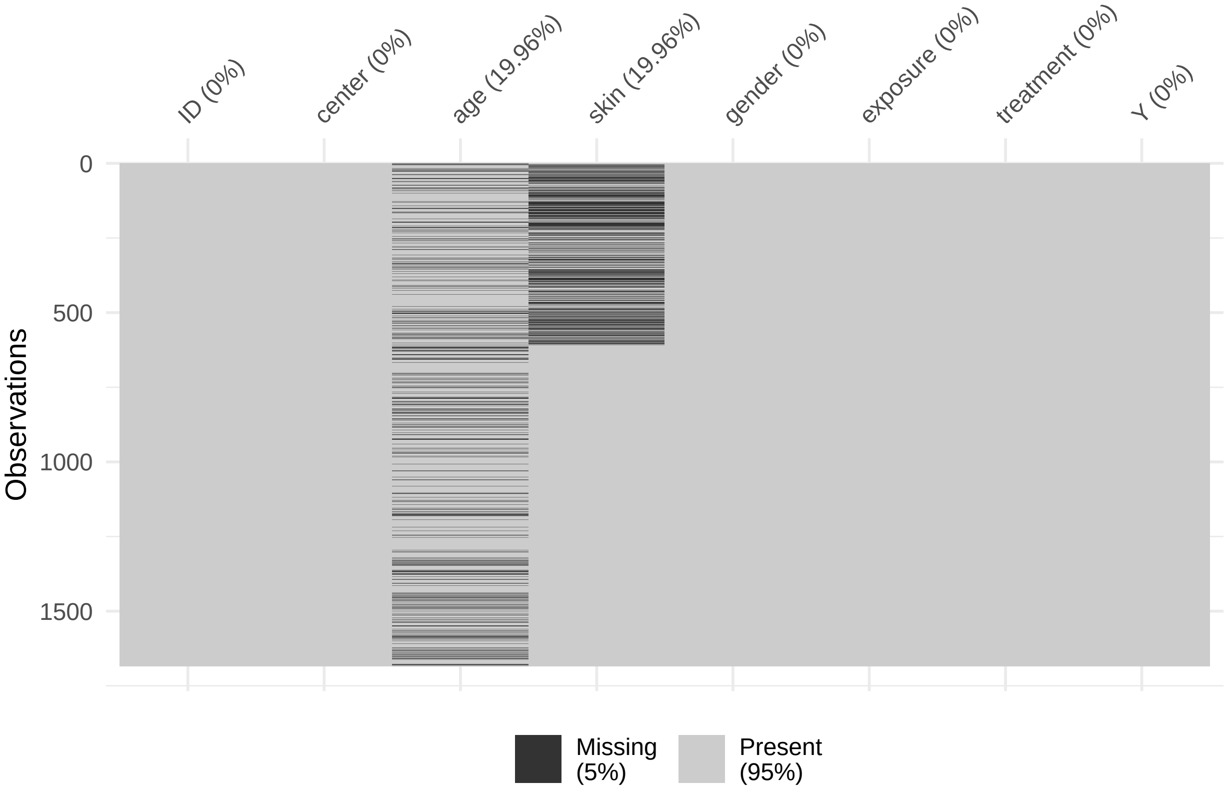
```
naniar::vis_miss(skin_mcar)
```



```
naniar::vis_miss(skin_mar)
```

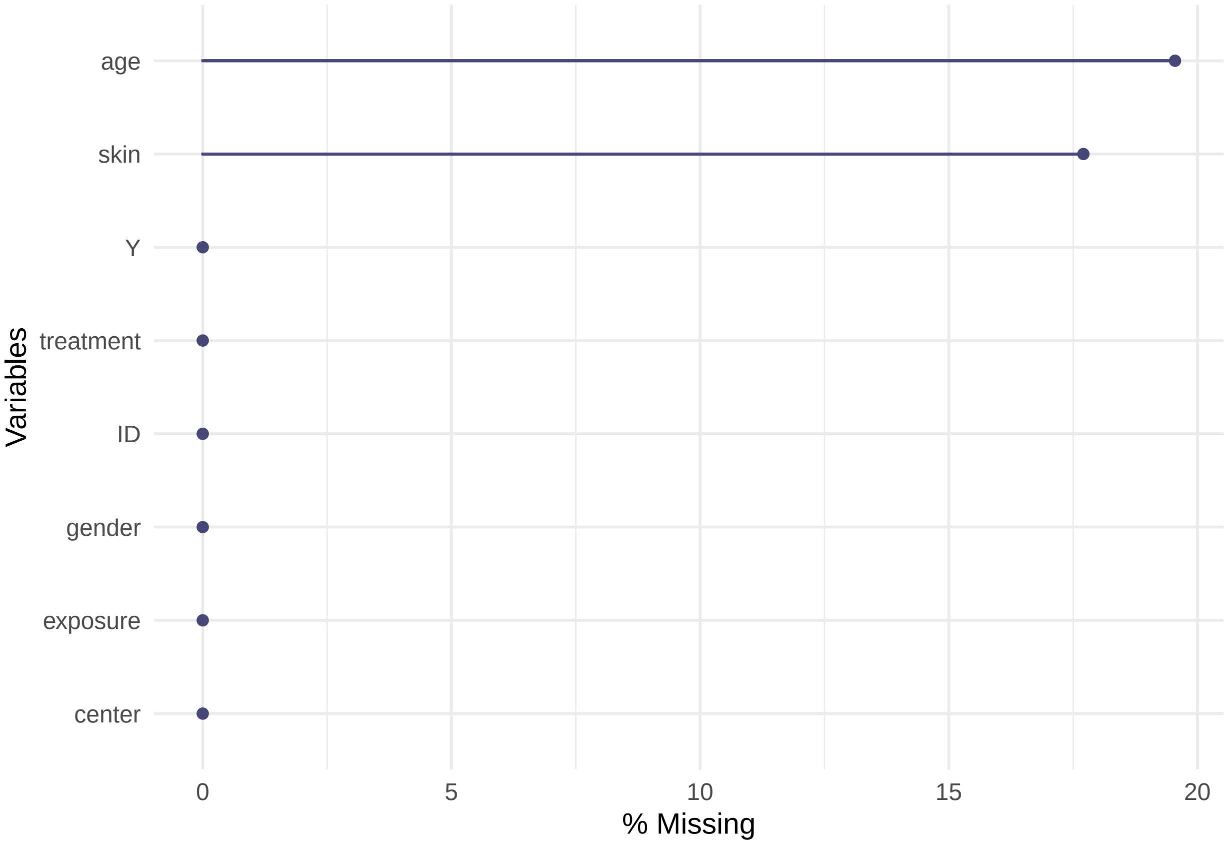


```
naniar::vis_miss(skin_mnar)
```

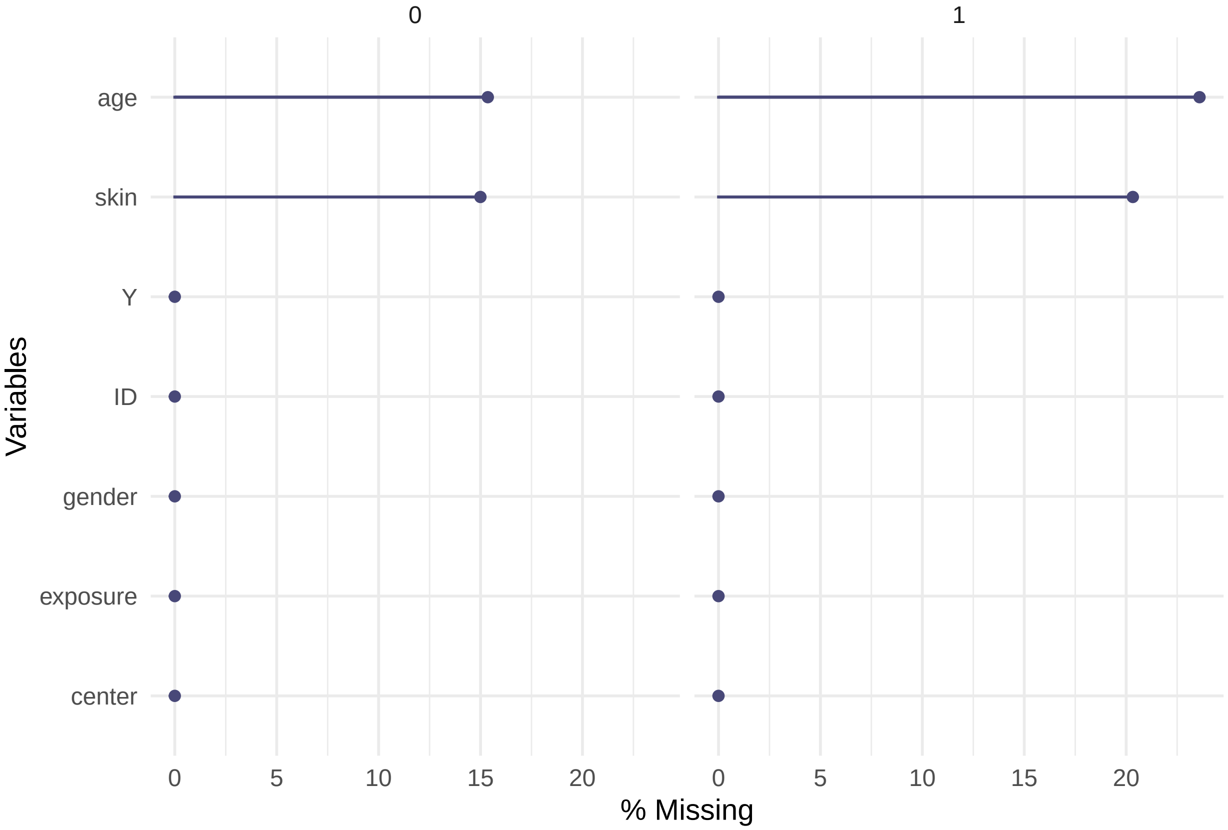


Lollipop plot of percentage of missingness for each variable

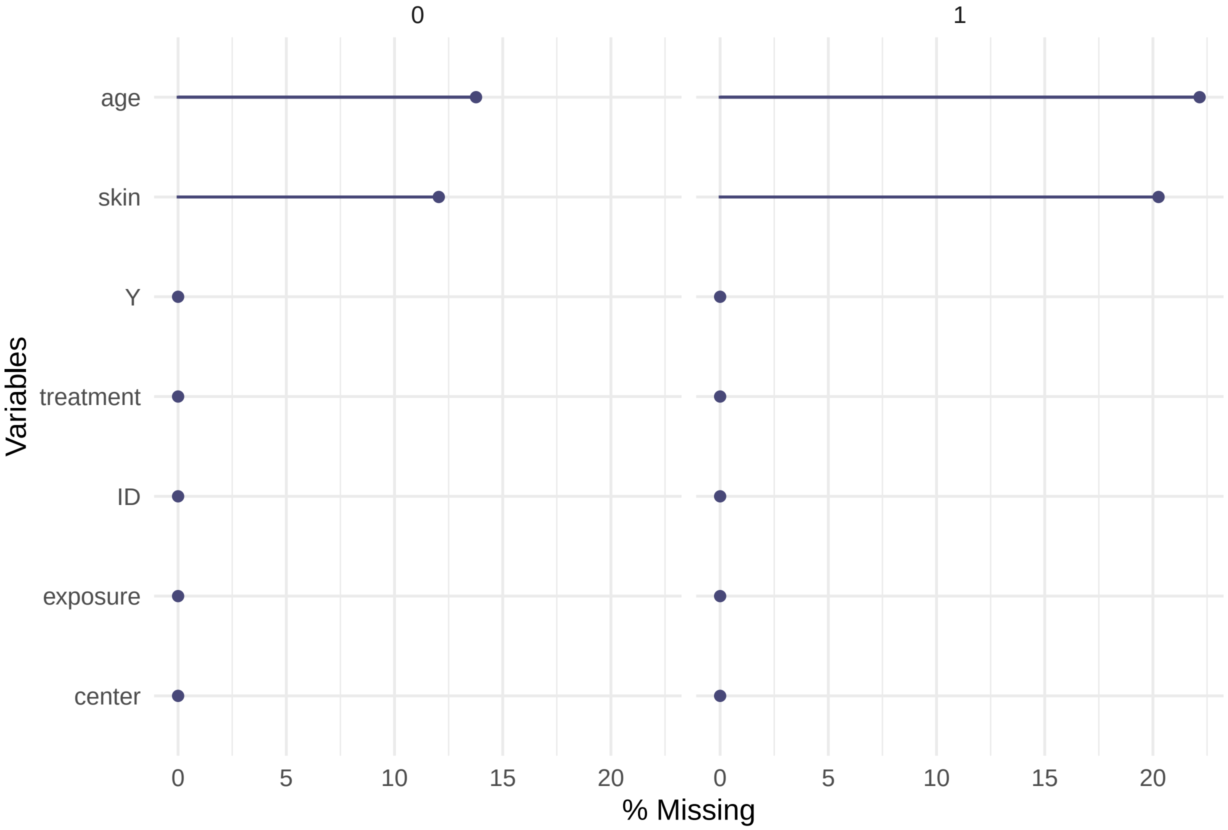
```
naniar::gg_miss_var(skin_mar, show_pct = TRUE)
```



```
# Stratified by treatment
naniar::gg_miss_var(skin_mar, show_pct = TRUE, facet = treatment)
```



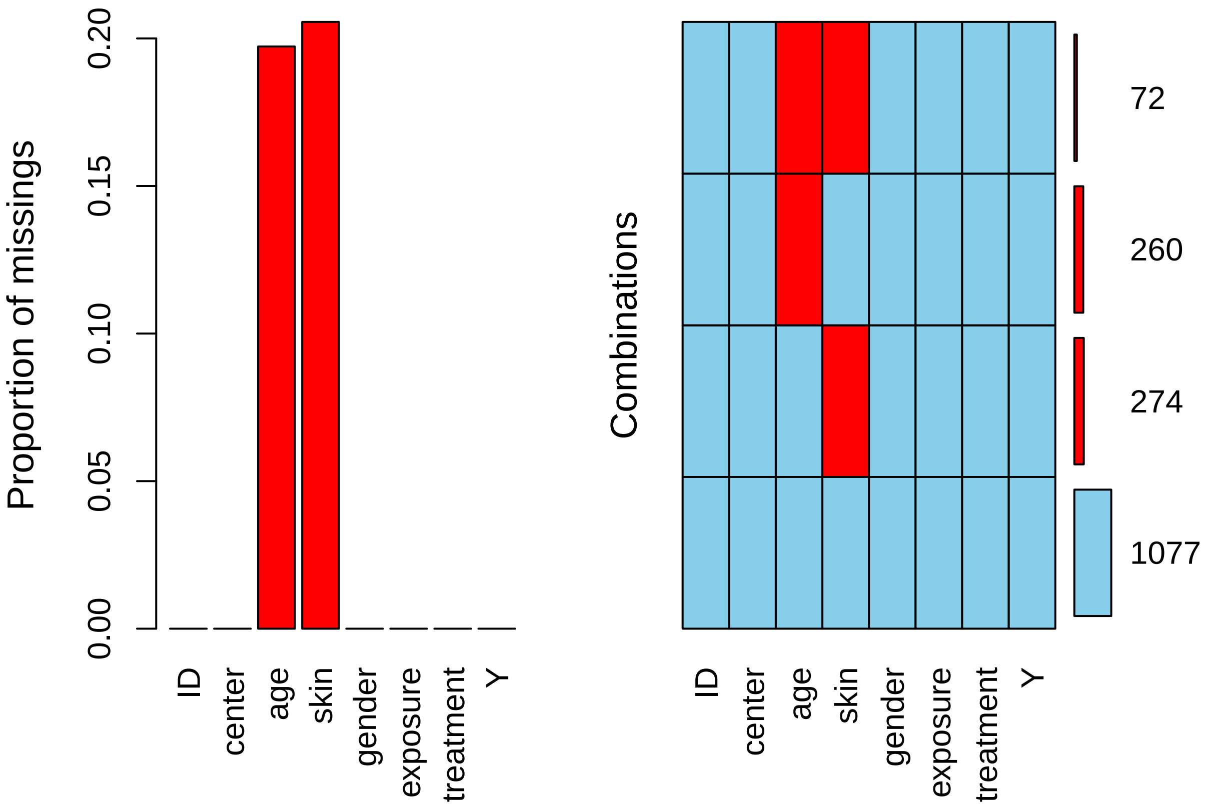
```
# Stratified by gender
naniar::gg_miss_var(skin_mar, show_pct = TRUE, facet = gender)
```



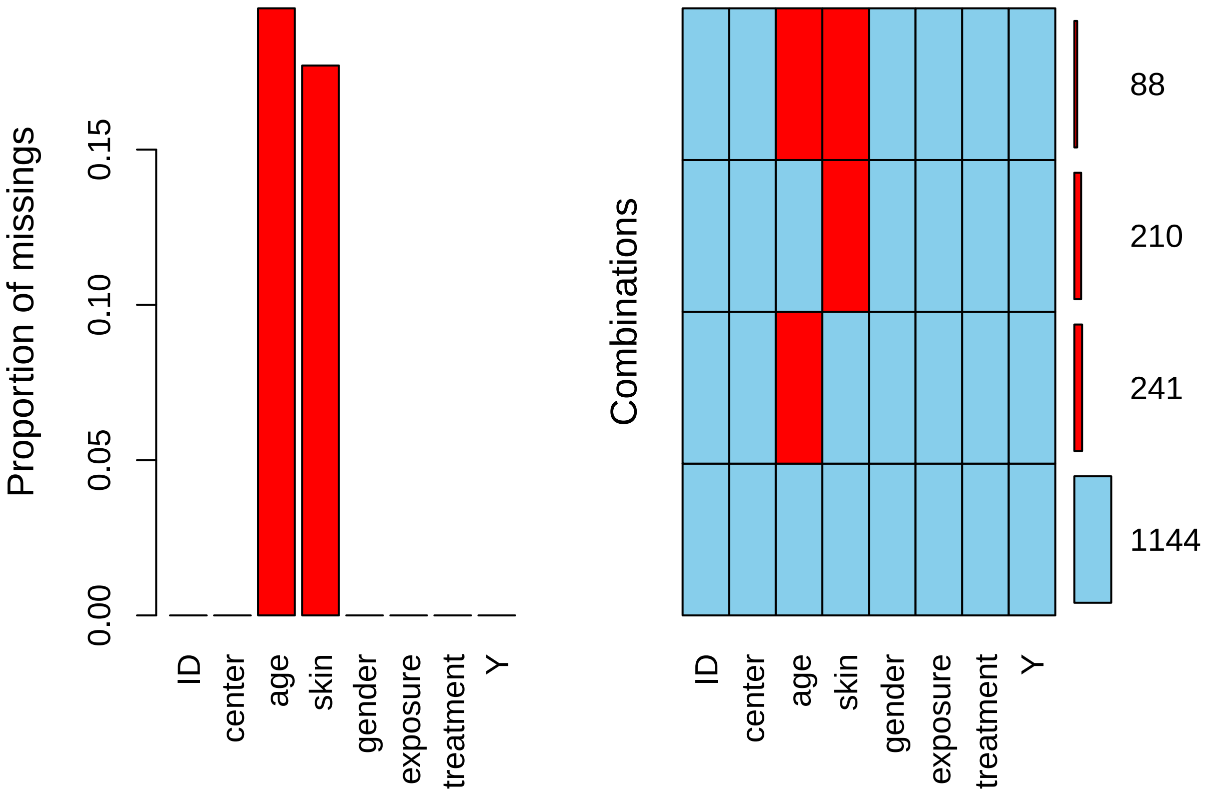
Aggregation plot

The left column is a bar chart of missingness for each variable while the right column depicts how often each **combination** of missing occurs.

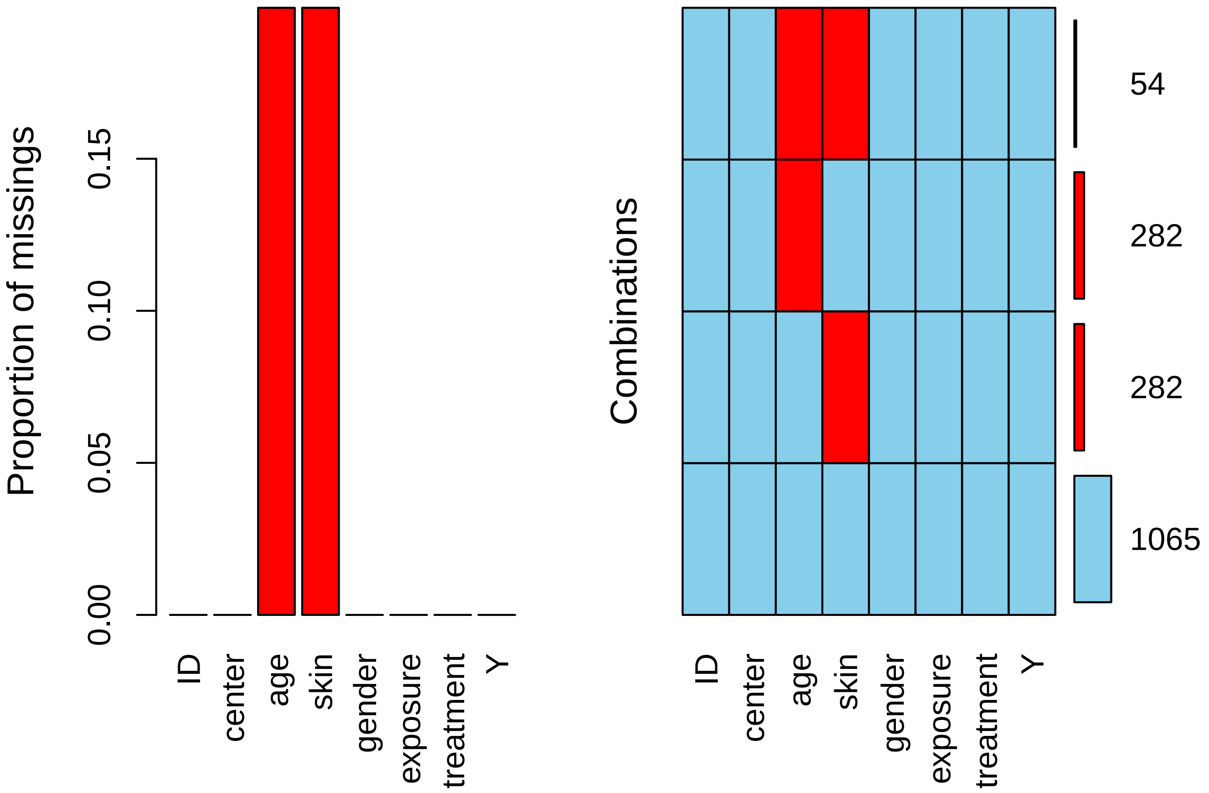
```
VIM::aggr(skin_mcar, numbers = TRUE, prop = c(TRUE, FALSE))
```



```
VIM::aggr(skin_mar, numbers = TRUE, prop = c(TRUE, FALSE))
```



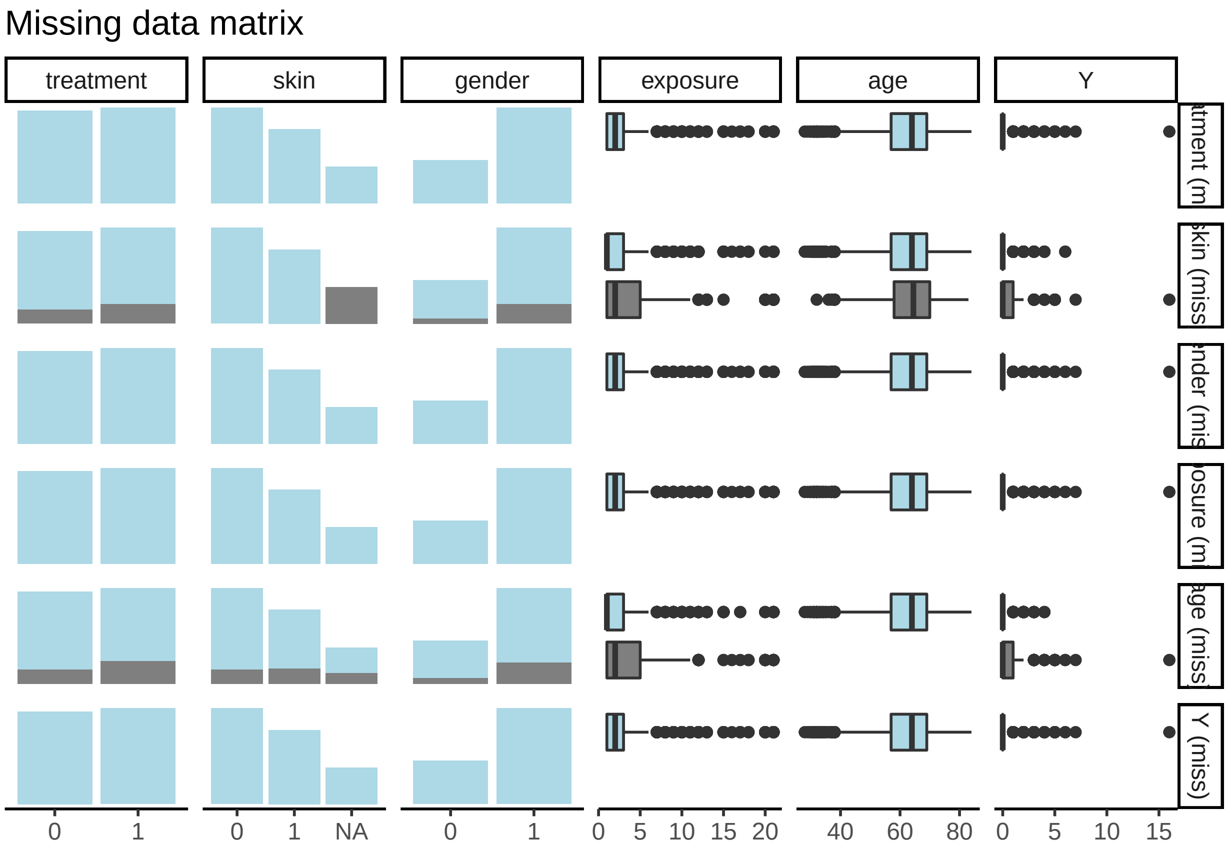
```r
VIM::aggr(skin_mnar, numbers = TRUE, prop = c(TRUE, FALSE))
```



Missing data matrix

For continuous variables, such as age and exposure, the data is presented in box plots. Taking row 2 and column 6 as an example, the distribution of age in patients who have skin data is shown in the blue box plot, and the distribution of age in patients with missing skin data is shown in the grey box plot. For categorical variables, the data is presented as bar plot.

```r
# need to change categorical variables from numeric type to factor type
skin_mar[, c("center", "skin", "gender", "treatment")] <-
  lapply(skin_mar[, c("center", "skin", "gender", "treatment")], factor)
finalfit::missing_pairs(skin_mar, dependent = "treatment",
                        explanatory = c("skin", "gender", "exposure", "age", "Y"))
```



Descriptive table

We can also describe the prevalence of missingness in a table using `missing_compare()`. We show how other variables are associated with the missingness of age in each different missing mechanism

```
skin_mcar %>%
  finalfit::missing_compare(dependent="age",
                            explanatory=c("skin","gender","exposure","treatment","Y")) %>%
  knitr::kable(row.names=FALSE, align = c("l", "l", "r", "r", "r"))
```

Missing data analysis: age		Not missing	Missing	p
skin	0	590 (81.0)	138 (19.0)	0.670
	1	487 (80.0)	122 (20.0)	
gender	0	428 (81.8)	95 (18.2)	0.310
	1	923 (79.6)	237 (20.4)	
exposure	Mean (SD)	2.8 (3.3)	3.2 (3.7)	0.062
treatment	0	668 (80.8)	159 (19.2)	0.656
	1	683 (79.8)	173 (20.2)	
Y	Mean (SD)	0.3 (0.7)	0.3 (1.1)	0.289

```
skin_mar %>%
  finalfit::missing_compare(dependent="age",
                            explanatory=c("skin","gender","exposure","treatment","Y")) %>%
  knitr::kable(row.names=FALSE, align = c("l", "l", "r", "r", "r"))
```

Missing data analysis: age		Not missing	Missing	p
skin	0	666 (85.3)	115 (14.7)	0.004
	1	478 (79.1)	126 (20.9)	
gender	0	451 (86.2)	72 (13.8)	<0.001
	1	903 (77.8)	257 (22.2)	
exposure	Mean (SD)	2.6 (2.9)	4.2 (4.6)	<0.001
treatment	0	700 (84.6)	127 (15.4)	<0.001
	1	654 (76.4)	202 (23.6)	
Y	Mean (SD)	0.2 (0.5)	0.6 (1.5)	<0.001

```r
skin_mnar %>%
  finalfit::missing_compare(dependent="age",
                            explanatory=c("skin","gender","exposure","treatment","Y")) %>%
  knitr::kable(row.names=FALSE, align = c("l", "l", "r", "r", "r"))
```

Missing data analysis: age		Not missing	Missing	p
skin	0	445 (76.9)	134 (23.1)	0.097
	1	620 (80.7)	148 (19.3)	
gender	0	409 (78.2)	114 (21.8)	0.231
	1	938 (80.9)	222 (19.1)	
exposure	Mean (SD)	3.0 (3.5)	2.3 (2.8)	<0.001
treatment	0	663 (80.2)	164 (19.8)	0.941
	1	684 (79.9)	172 (20.1)	
Y	Mean (SD)	0.3 (0.9)	0.2 (0.6)	0.018

Observations

- Note that none of these visualization tools were helpful in distinguishing the missing data mechanism (especially between MAR and MNAR)
- For MNAR, missingness relies on some unknown variables or the missing values themselves and we need subject-matter expertise

Implications of performing complete-case analysis with missing data

- The model of interest is a log-linear model for counts

$$\log(E(Y)) = \beta_0 + \beta_1 \text{treatment} + \beta_2 \text{age} + \beta_3 \text{gender} + \beta_4 \text{exposure} + \beta_5 \text{skin}$$

  where Y is the number of new skin cancers in the first year of treatment.

- Note that we are ignoring some assumptions about the model:
  - Zero inflation
  - Overdispersion
  - These are topics for another workshop!

```r
# we convert variables from numeric type to factor type before running the regression
skinbl[, c("center", "skin", "gender", "treatment")] <-
  lapply(skinbl[, c("center", "skin", "gender", "treatment")], factor)
skin_mcar[, c("center", "skin", "gender", "treatment")] <-
  lapply(skin_mcar[, c("center", "skin", "gender", "treatment")], factor)
skin_mnar[, c("center", "skin", "gender", "treatment")] <-
  lapply(skin_mnar[, c("center", "skin", "gender", "treatment")], factor)
```

```r
fullmod <- glm(Y ~ treatment + age + gender + exposure + skin,
               family = poisson("log"), data = skinbl)
mcarmod <- glm(Y ~ treatment + age + gender + exposure + skin,
               family = poisson("log"), data = skin_mcar)
```

```
marmod <- glm(Y ~ treatment + age + gender + exposure + skin,
              family = poisson("log"), data = skin_mar)
mnarmod <- glm(Y ~ treatment + age + gender + exposure + skin,
               family = poisson("log"), data = skin_mnar)

export_summs(fullmod, mcarmod, marmod, mnarmod, scale = FALSE,
             model.names = c("Full", "MCAR", "MAR", "MNAR"), statistics = c(N = "nobs"))
```

	Full	MCAR	MAR	MNAR
(Intercept)	-3.60 ***	-3.85 ***	-3.02 ***	-4.44 ***
	(0.36)	(0.46)	(0.50)	(0.62)
treatment1	0.06	0.07	0.06	0.42 ***
	(0.09)	(0.12)	(0.14)	(0.12)
age	0.02 ***	0.02 **	0.01	0.02 **
	(0.01)	(0.01)	(0.01)	(0.01)
gender1	0.47 ***	0.59 ***	0.28	0.43 **
	(0.12)	(0.15)	(0.16)	(0.14)
exposure	0.13 ***	0.13 ***	0.13 ***	0.12 ***
	(0.01)	(0.01)	(0.01)	(0.01)
skin1	0.36 ***	0.41 ***	0.30 *	0.73 ***
	(0.09)	(0.12)	(0.14)	(0.13)
N	1683	1077	1144	1065

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Observations

- Coefficient estimates from MCAR are close to Full, but the standard error estimates are larger
- Coefficient estimates from MAR and MNAR are quite different and the standard estimates are larger compared to Full
- Coefficient estimate of `treatment` is much larger in MNAR even though `treatment` was not related to missingness of `age` or `skin`