

EP16: Missing Values in Clinical Research: Multiple Imputation

5. Know Your Data


Nicole S. Erler

Department of Biostatistics, Erasmus Medical Center

✉ n.erler@erasmusmc.nl

Missing Data Pattern

To demonstrate the **work flow** when performing **multiple imputation** with the **mice** package, we use data from the [National Health and Nutrition Examination Survey \(NHANES\)](#).

There are several packages in  that provide functions to investigate **the missing data pattern**.

Examples are:

mice, JointAI, VIM, Amelia, visdat, naniar, ...

Missing Data Pattern

```
mdp <- mice::md.pattern(NHANES, plot = FALSE)
head(mdp[, -c(7:14)]) # omit some columns to fit it on the slide
```

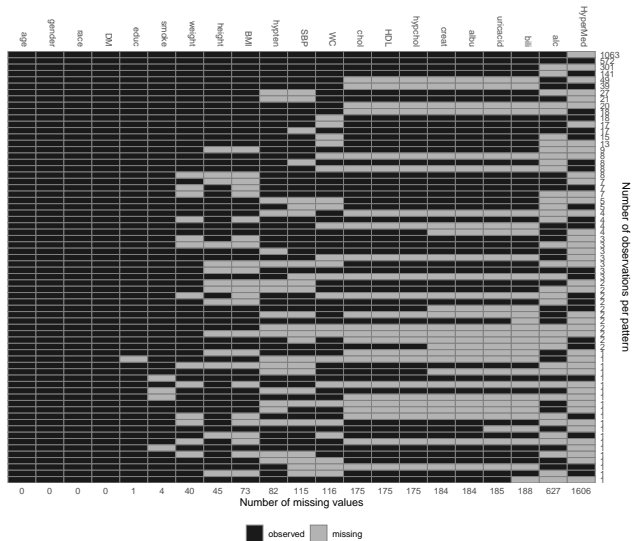
##	age	gender	race	DM	educ	smoke	hypchol	creat	albu	uricacid	bili	alc	HyperMed
## 572	1	1	1	1	1	1	1	1	1	1	1	1	0
## 1063	1	1	1	1	1	1	1	1	1	1	1	1	0 1
## 141	1	1	1	1	1	1	1	1	1	1	1	0	1 1
## 301	1	1	1	1	1	1	1	1	1	1	1	0	0 2
## 2	1	1	1	1	1	1	1	1	1	1	0	1	0 2
## 1	1	1	1	1	1	1	1	1	1	1	0	0	0 3

```
tail(mdp[, -c(7:14)])
```

##	age	gender	race	DM	educ	smoke	hypchol	creat	albu	uricacid	bili	alc	HyperMed
## 1	1	1	1	1	1	0	1	1	1	1	1	1	1
## 1	1	1	1	1	1	0	1	1	1	1	1	1	0 2
## 1	1	1	1	1	1	0	0	0	0	0	0	0	0 10
## 1	1	1	1	1	1	0	1	1	1	1	1	1	0 4
## 1	1	1	1	1	0	1	0	0	0	0	0	1	0 12
##	0	0	0	0	1	4	175	184	184	185	188	627	1606 3975

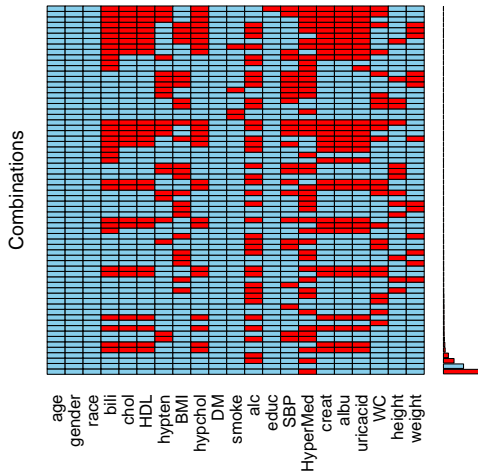
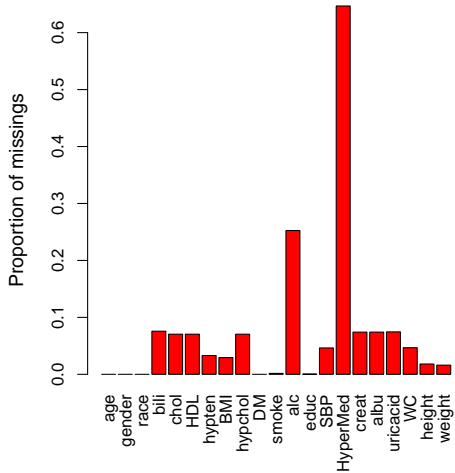
Missing Data Pattern

JointAI::md_pattern(NHANES)



Missing Data Pattern

```
VIM::aggr(NHANES, prop = TRUE)
```



Missing Data Pattern

We are also interested in the number and proportion of (in)complete cases ...

```
cctab <- table(complete.cases(NHANES))  
cbind(  
  "#" = setNames(cctab, c('incomplete', 'complete')),  
  "%" = round(100 * cctab/nrow(NHANES), 2)  
)
```

##		#	%
##	incomplete	1911	76.96
##	complete	572	23.04

Missing Data Pattern

... and the proportion of missing values per variable:

```
cbind("# NA" = sort(colSums(is.na(NHANES))),  
      "% NA" = round(sort(colMeans(is.na(NHANES))) * 100, 2))
```

##	# NA	% NA	##	# NA	% NA	##	# NA	% NA
## age	0	0.00	## height	45	1.81	## hypchol	175	7.05
## gender	0	0.00	## BMI	73	2.94	## creat	184	7.41
## race	0	0.00	## hypten	82	3.30	## albu	184	7.41
## DM	0	0.00	## SBP	115	4.63	## uricacid	185	7.45
## educ	1	0.04	## WC	116	4.67	## bili	188	7.57
## smoke	4	0.16	## chol	175	7.05	## alc	627	25.25
## weight	40	1.61	## HDL	175	7.05	## HyperMed	1606	64.68

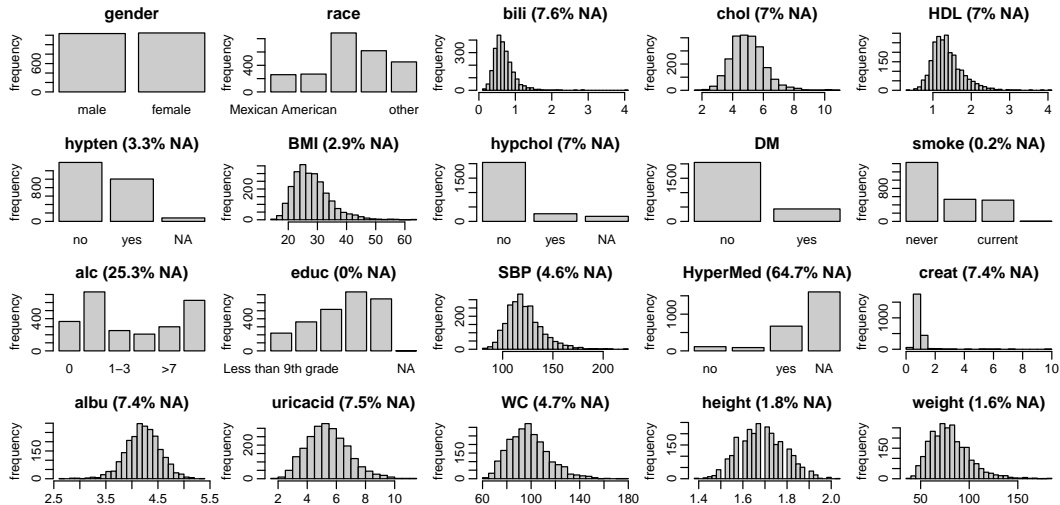
Missing Data Pattern

See also:

- ▶ `mice::md.pattern()`
- ▶ `visdat::vis_miss()`
- ▶ `visdat::vis_dat()`
- ▶ ...
- ▶ `naniar::prop_miss_case()`,
`naniar::pct_miss_case()`
- ▶ `naniar::prop_complete_case()`,
`naniar::pct_complete_case()`
- ▶ `naniar::miss_var_summary()`
- ▶ `mice::md.pairs()`
- ▶ ...

Data Distribution

```
JointAI::plot_all(NHANES[, -1])
```

exclude 1st column to fit on slide

Correlations & Patterns

A quick (and dirty) way to check for strong correlations between variables is:

```
# re-code all variables as numeric and calculate spearman correlation
Corr <- cor(sapply(NHANES, as.numeric),
            use = "pairwise.complete.obs", method = "spearman")
```

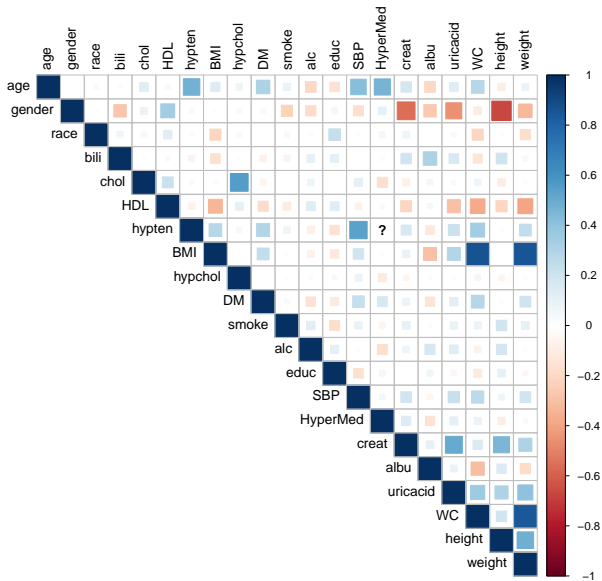
```
## Warning in cor(sapply(NHANES, as.numeric), use =
## "pairwise.complete.obs", : the standard deviation is zero
```

```
corrplot::corrplot(Corr, method = "square", type = "upper",
                    tl.col = "black")
```

Note:

We only use the correlation coefficient for categorical variables for visualization, not as a statistical result!

Correlations & Patterns



Correlations & Patterns

Check out what the problem is with `hypertension` and `HyperMed`:

```
table(hypertension = NHANES$hypten,  
      HyperMed = NHANES$HyperMed, exclude = NULL)
```

```
##           HyperMed  
## hypertension  no previous  yes <NA>  
##           no      0         0    0 1397  
##           yes    114        90  673  127  
##           <NA>    0         0    0   82
```

Why are values missing?

Knowing your data also means being able to answer these questions:

- ▶ Do missing values in multiple variables always **occur together**? (e.g. blood measurements)
- ▶ Are there **structural missing values**? (e.g. pregnancy status in men)
- ▶ Are there **patterns** in the missing values? (e.g. only patients with hypertension have observations of HyperMed)
- ▶ Are values **missing by design**?
- ▶ Is the **assumption of ignorable missingness** (MAR or MCAR) justifiable?

Auxiliary Variables

Auxiliary variables are variables that are not part of the analysis but **can help during imputation**.

Good auxiliary variables

- ▶ are **related to the probability of missingness** in a variable, or
- ▶ are **related to the incomplete variable** itself,
- ▶ do **not have many missing values** themselves and
- ▶ are (mostly) **observed** when the incomplete variable of interest is missing.

Your Turn!

Practical

Getting to know your data [html](#)