Data visualization

for incomplete datasets in R

Hanne Oberman

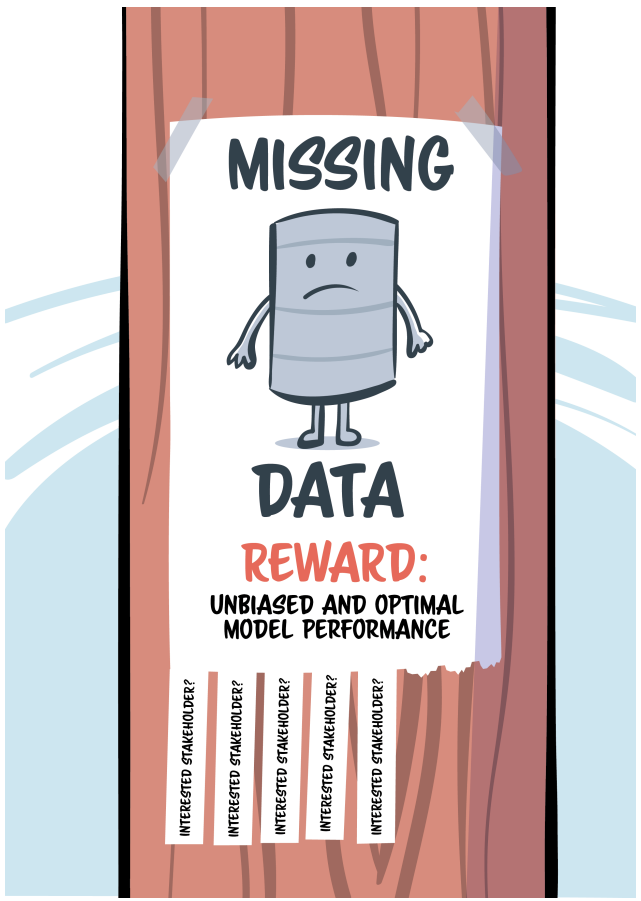
Utrecht University

Take aways

Missing data are

- a pervasive problem
- visualizable & analyzable
- informative!

Missingness



A problem to fix

- unit non-response

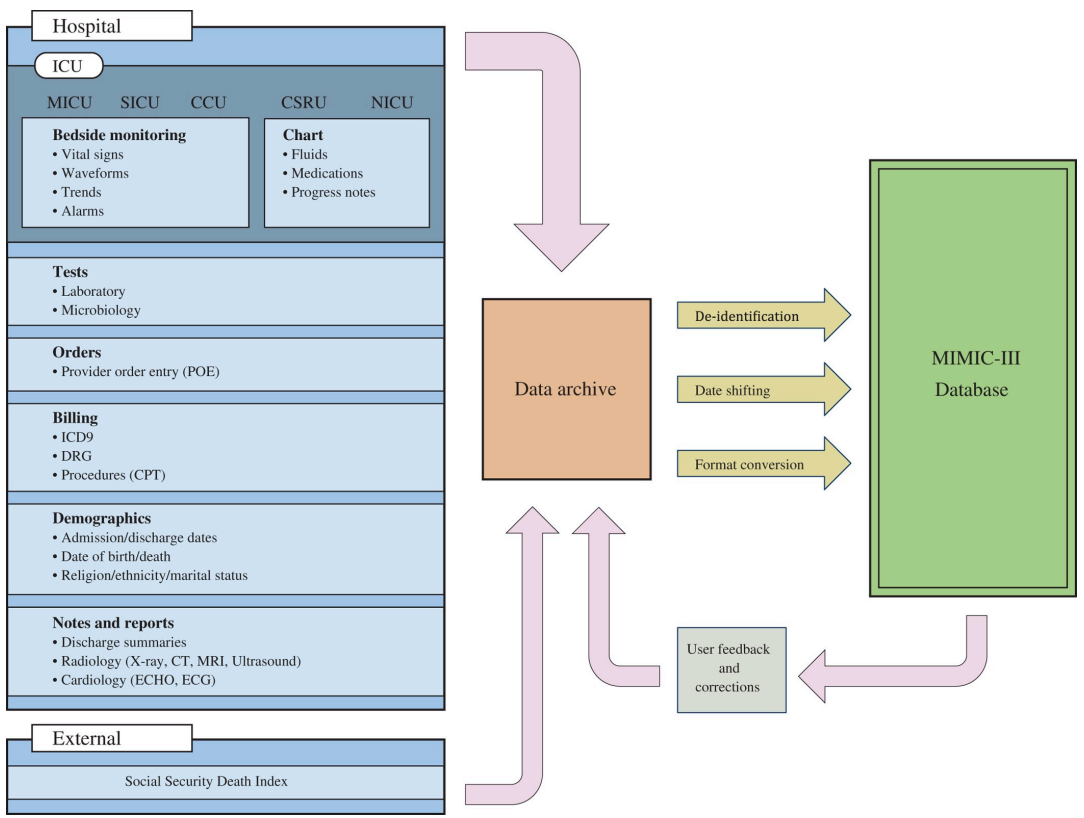
⟶ weighting etc.

- item non-response

⟶ imputation etc.

Case study

```
set.seed(123)
library(ricu)
library(mice)
library(ggmice)
library(ggplot2)
dat <- mimic_demo |> clean_mimic_demo()
```



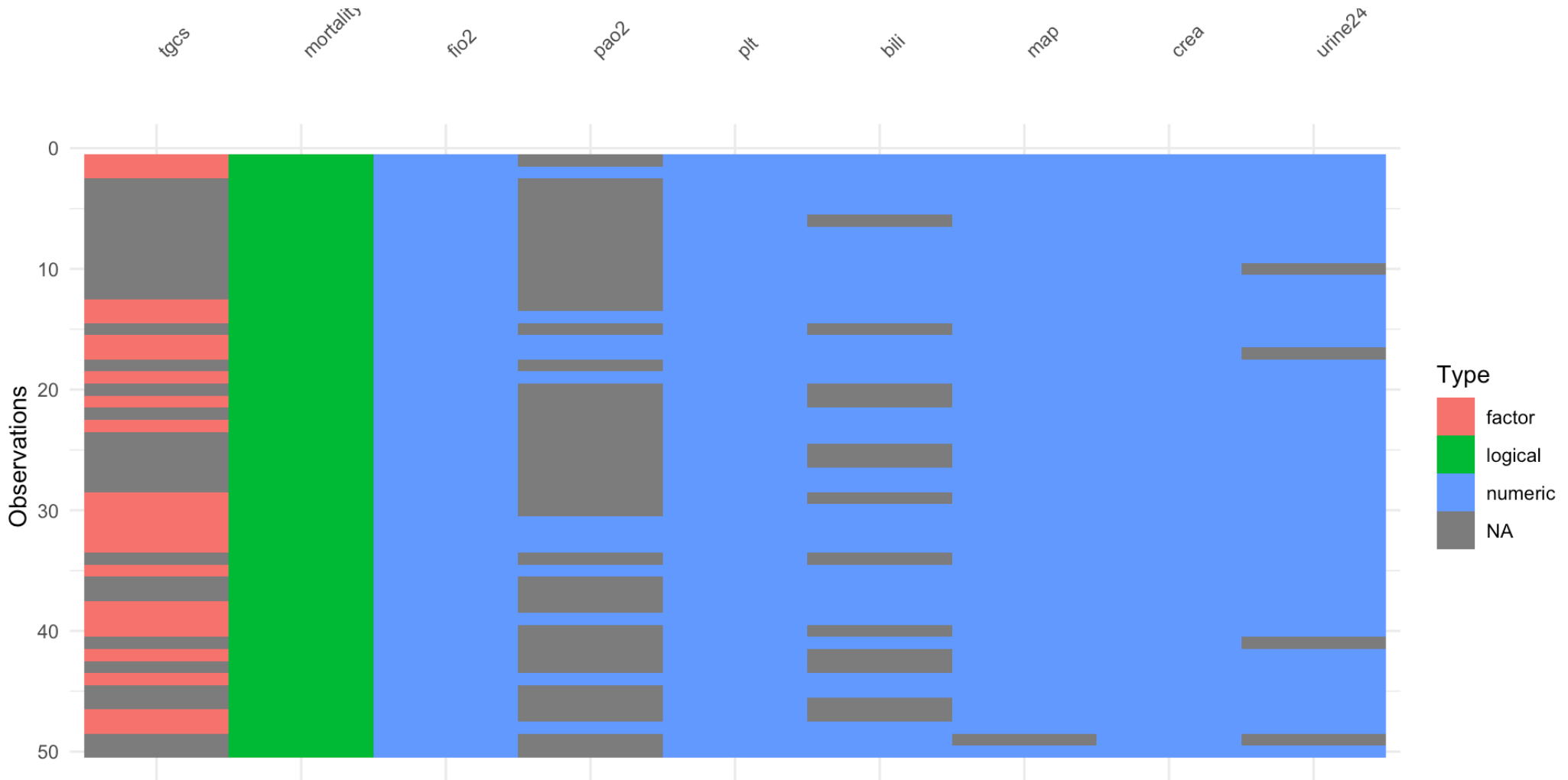
Incomplete data

```
1 str(dat)
```

```
'data.frame':   50 obs. of  9 variables:
 $ fio2     : num  100 100 100 100 40 50 100 50 100 100 ...
 $ pao2     : num  NA 20 NA NA NA NA NA NA NA NA ...
 $ plt      : num  80 16 189 189 175 127 11 156 87 80 ...
 $ bili     : num  0.9 2.3 0.3 0.3 0.8 NA 1.2 0.6 1.1 2.1 ...
 $ tgcs     : Factor w/ 8 levels "3","4","5","6",..: 1 1 NA NA NA NA NA NA NA
NA ...
 $ map      : num  101 94 146 105 94 ...
 $ crea     : num  3.7 1.8 0.5 1.2 2 1.2 1.6 1.6 1 0.9 ...
 $ urine24  : num  296.6 38.1 721.9 1257.6 1516.8 ...
 $ mortality: logi  TRUE TRUE TRUE FALSE FALSE TRUE ...
```


Incomplete data

```
1 visdat::vis_dat(dat)
```



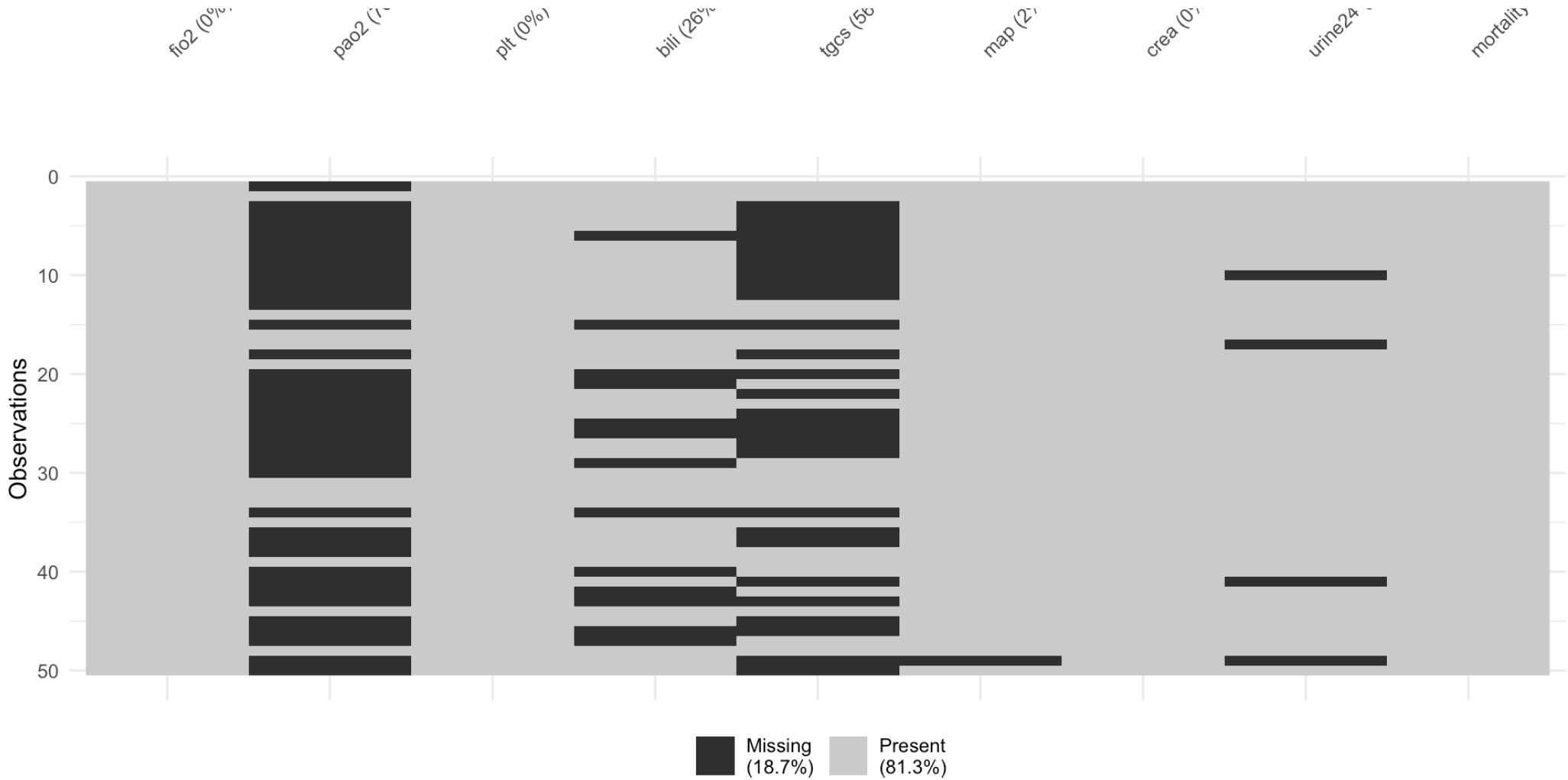
Response indicator

```
1 is.na(dat)
```

```
    fio2  pao2   plt  bili  tgcs   map  crea urine24 mortality
1  FALSE  TRUE FALSE FALSE FALSE FALSE FALSE   FALSE     FALSE
2  FALSE FALSE FALSE FALSE FALSE FALSE FALSE   FALSE     FALSE
3  FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE   FALSE     FALSE
4  FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE   FALSE     FALSE
5  FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE   FALSE     FALSE
6  FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE   FALSE     FALSE
7  FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE   FALSE     FALSE
8  FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE   FALSE     FALSE
9  FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE   FALSE     FALSE
10 FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE    TRUE     FALSE
11 FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE   FALSE     FALSE
12 FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE   FALSE     FALSE
13 FALSE  TRUE FALSE FALSE FALSE FALSE FALSE   FALSE     FALSE
```


Response indicator

```
1 naniar::vis_miss(dat)
```



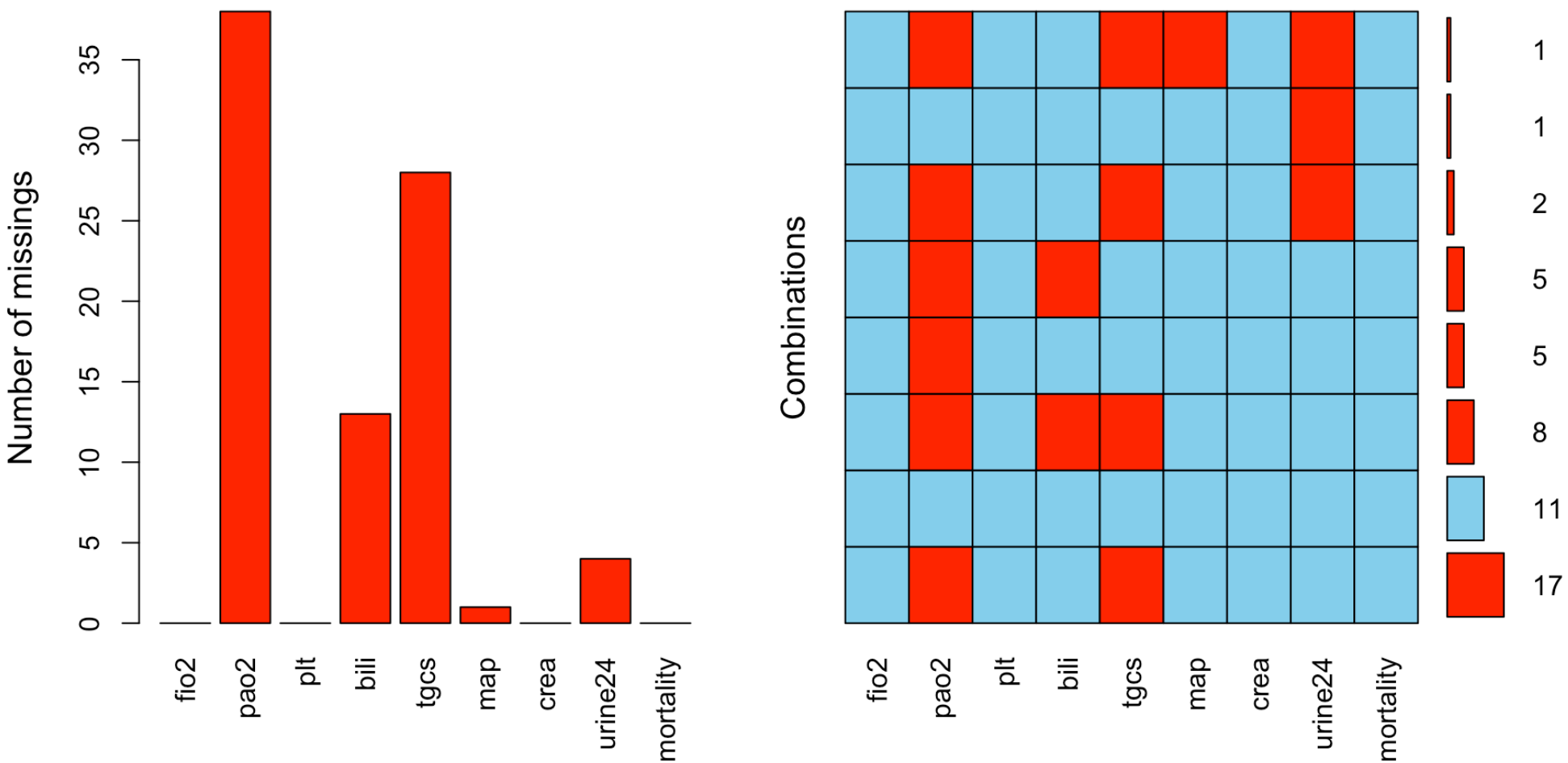
Missingness rate

```
1 colSums(is.na(dat))
```

```
     fio2       pao2        plt       bili       tgcs        map       crea
urine24
        0         38          0         13         28          1          0
4
mortality
        0
```


Missingness rate

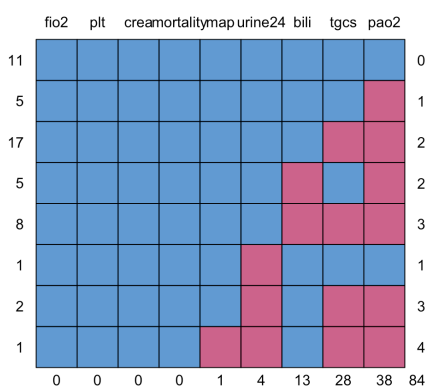
```
1 VIM::aggr(dat, numbers = TRUE, prop = FALSE)
```



Missing data pattern

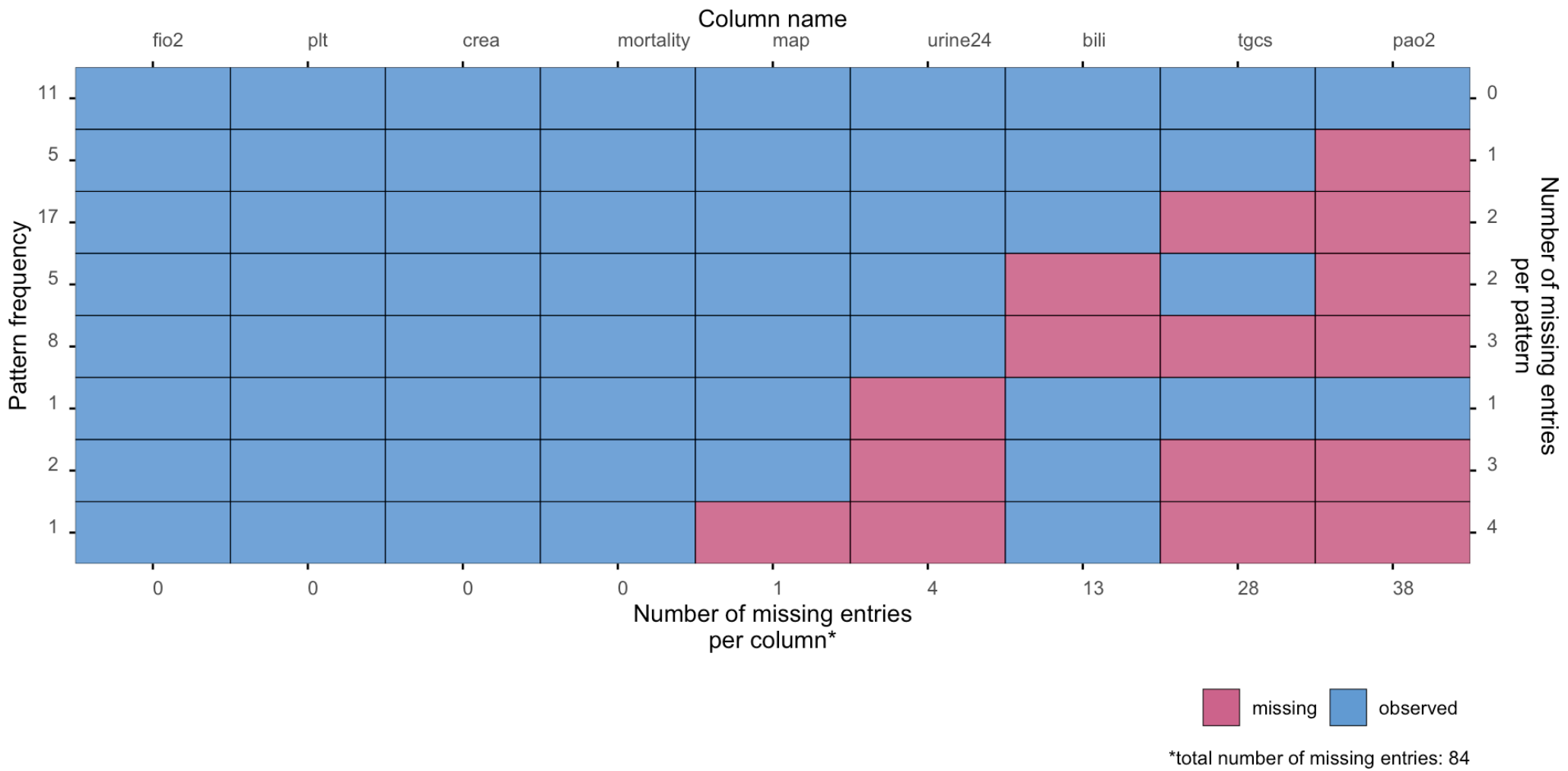
```
1 mice::md.pattern(dat)
```

```
    fio2 plt crea mortality map urine24 bili tgcs pao2
11     1   1    1         1   1        1    1    1    1  0
5      1   1    1         1   1        1    1    1    0  1
17     1   1    1         1   1        1    1    0    0  2
5      1   1    1         1   1        1    0    1    0  2
8      1   1    1         1   1        1    0    0    0  3
1      1   1    1         1   1        0    1    1    1  1
2      1   1    1         1   1        0    1    0    0  3
1      1   1    1         1   0        0    1    0    0  4
       0   0    0         0   1        4   13   28   38 84
```



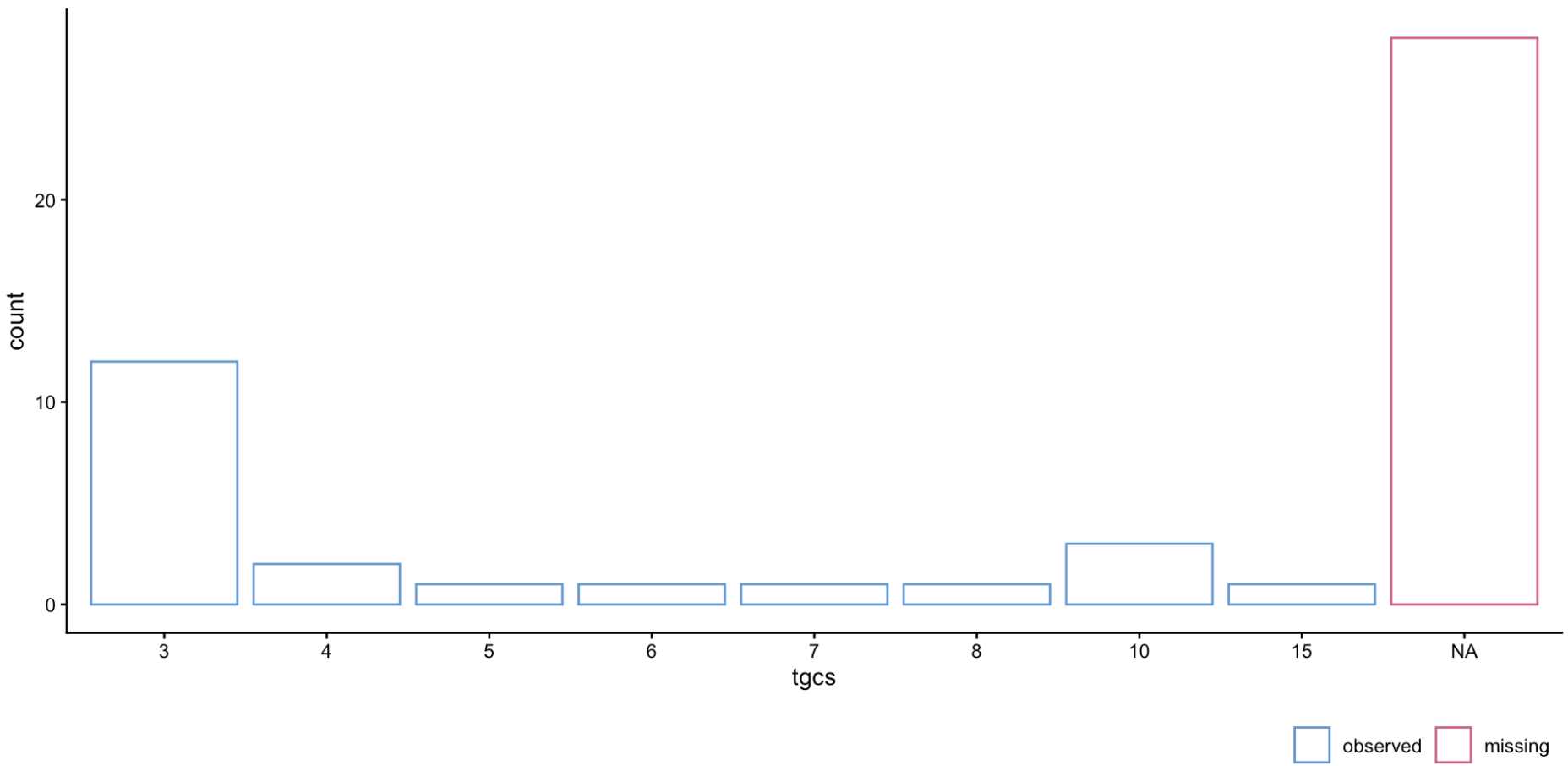
Missing data pattern

```
1 ggmice::plot_pattern(dat, square = FALSE)
```



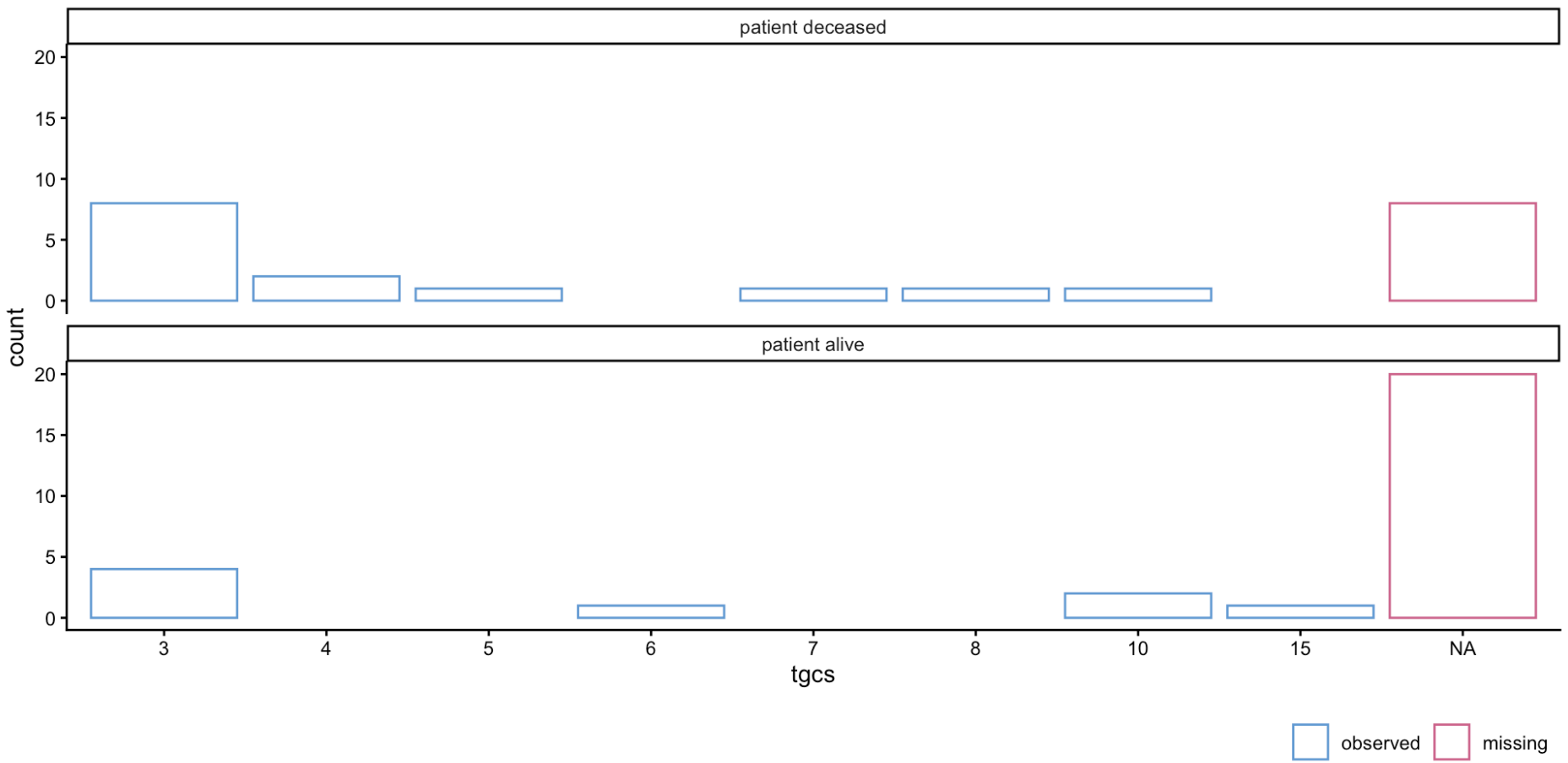
Coma symptoms

▶ Code



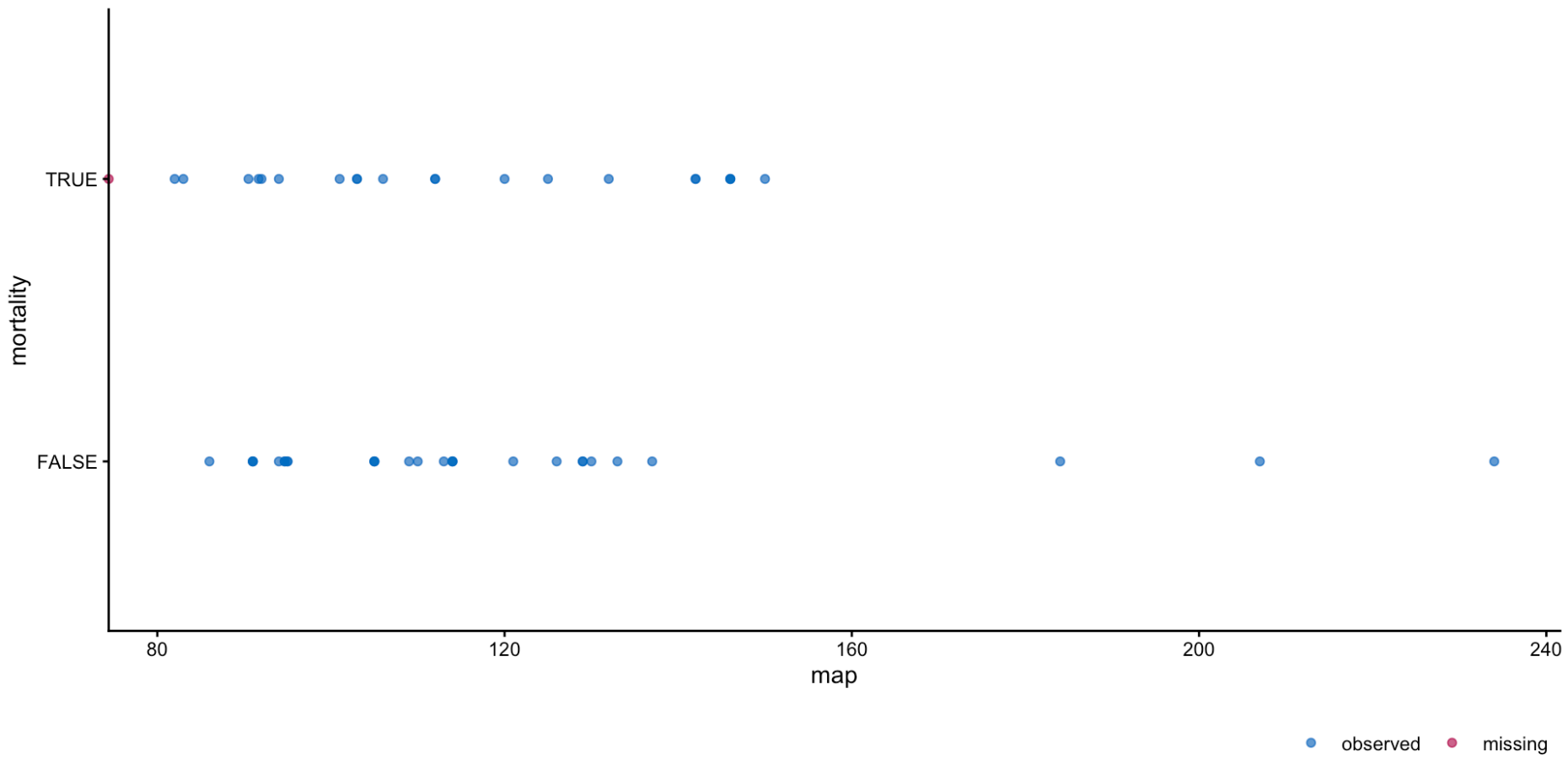
Coma symptoms by mortality

- Code



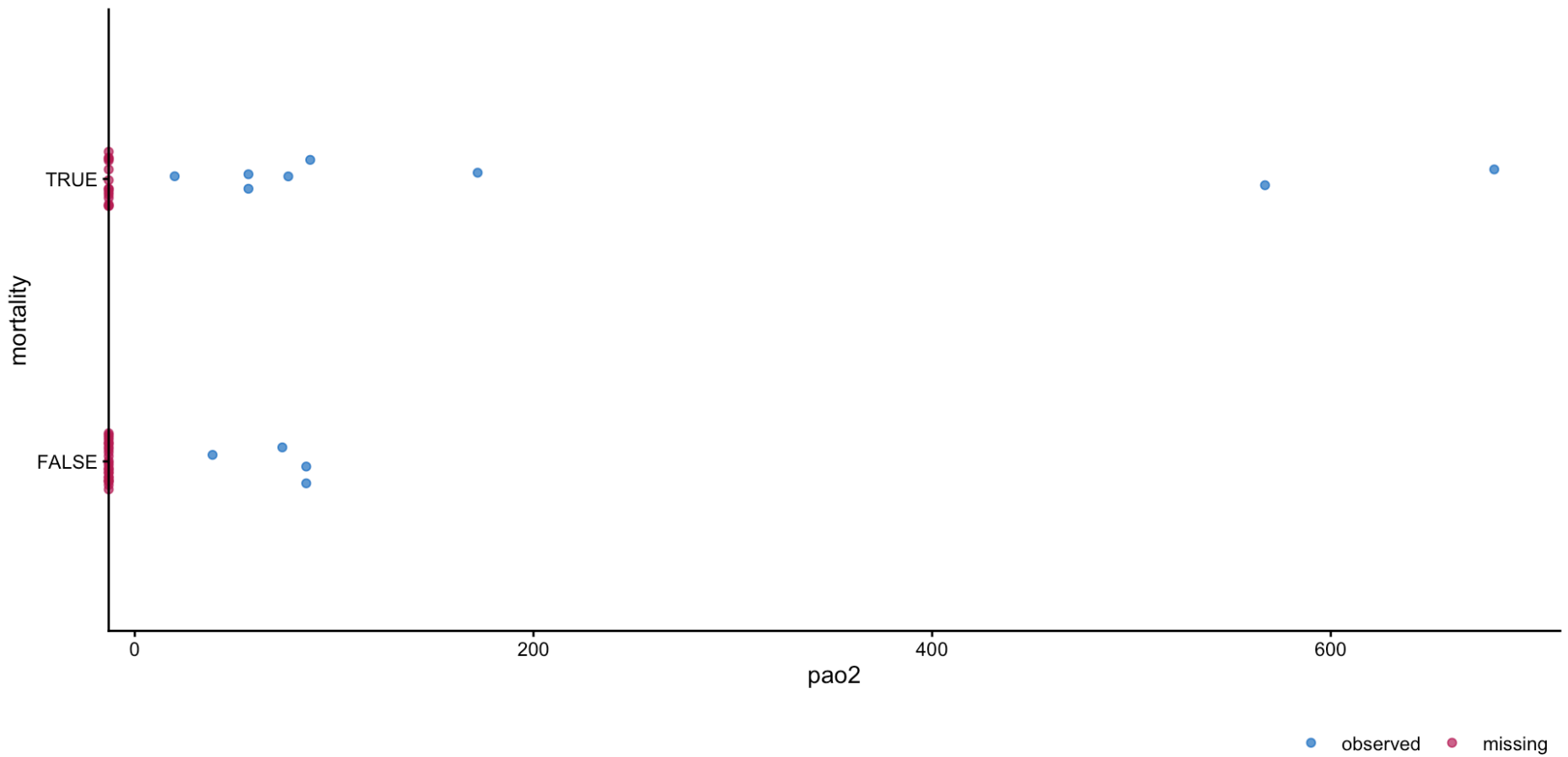
Blood pressure by mortality

▶ Code



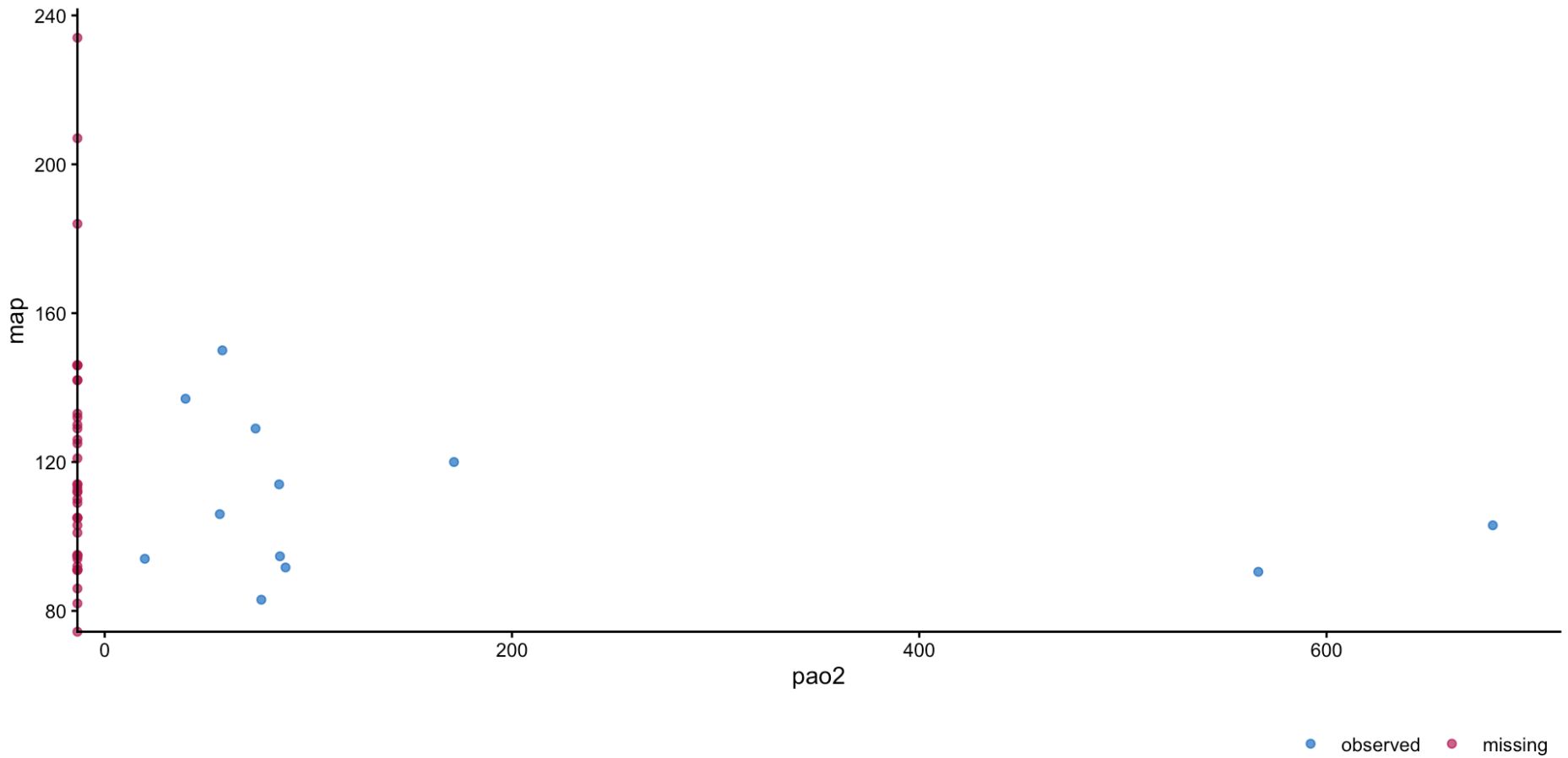
Blood oxygenation by mortality

▶ Code



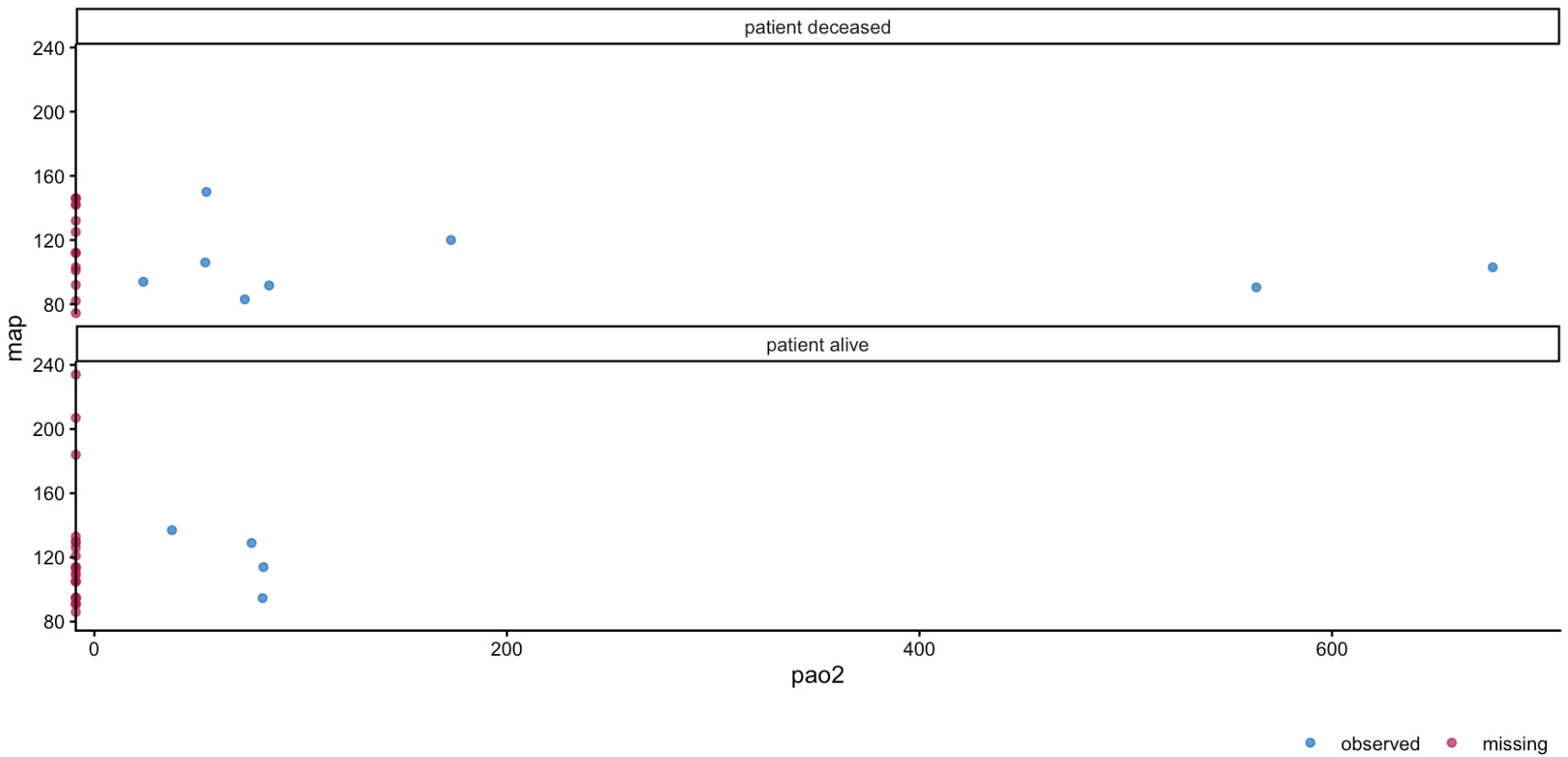
Scatter plot

▶ Code

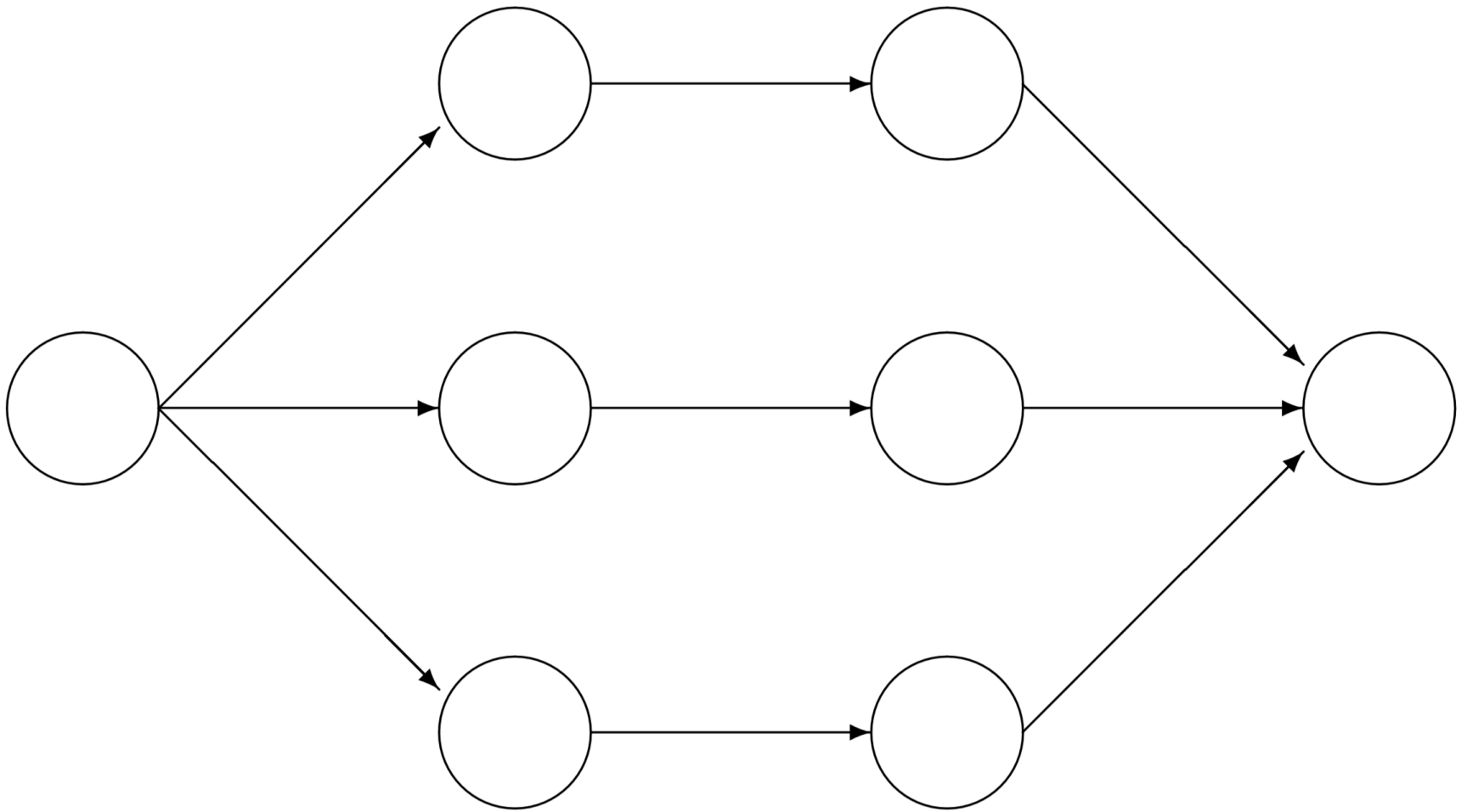


Faceted scatter plot

- Code



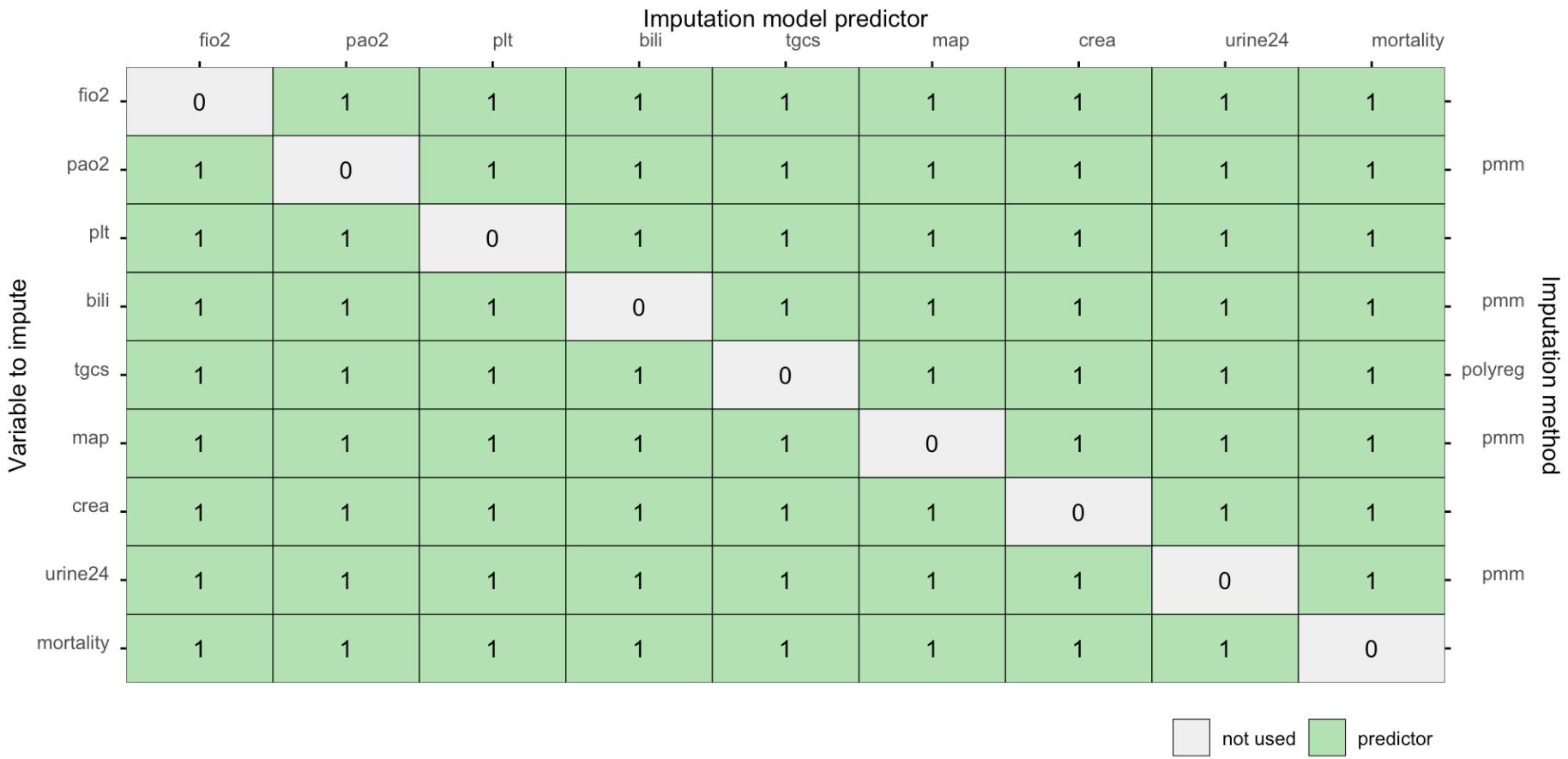
Imputation workflow



Incomplete data Imputed data Analysis results Pooled result

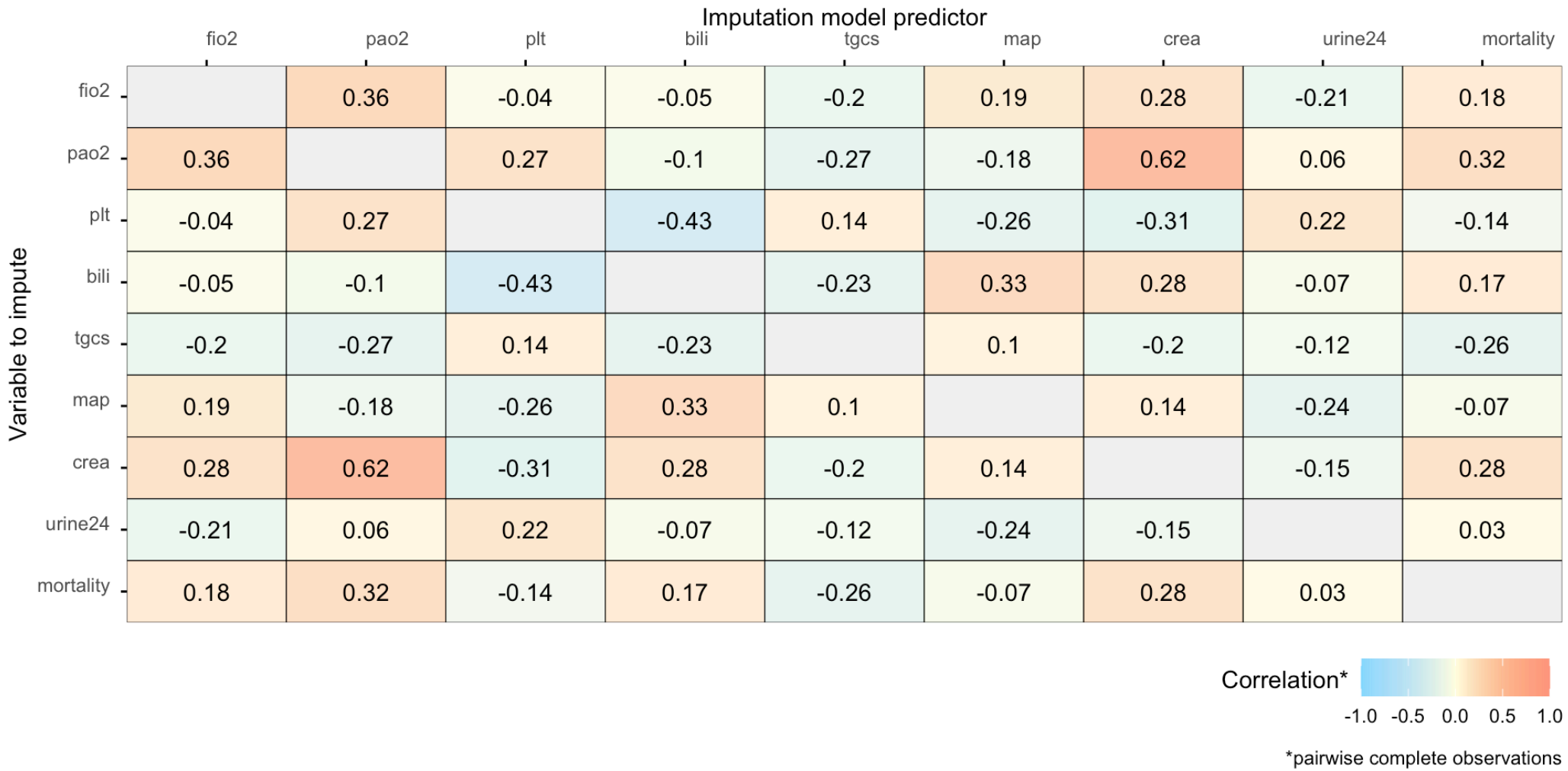
Imputation models

```
1 pred <- make.predictorMatrix(dat)
2 meth <- make.method(dat)
3 plot_pred(pred, method = meth, square = FALSE)
```



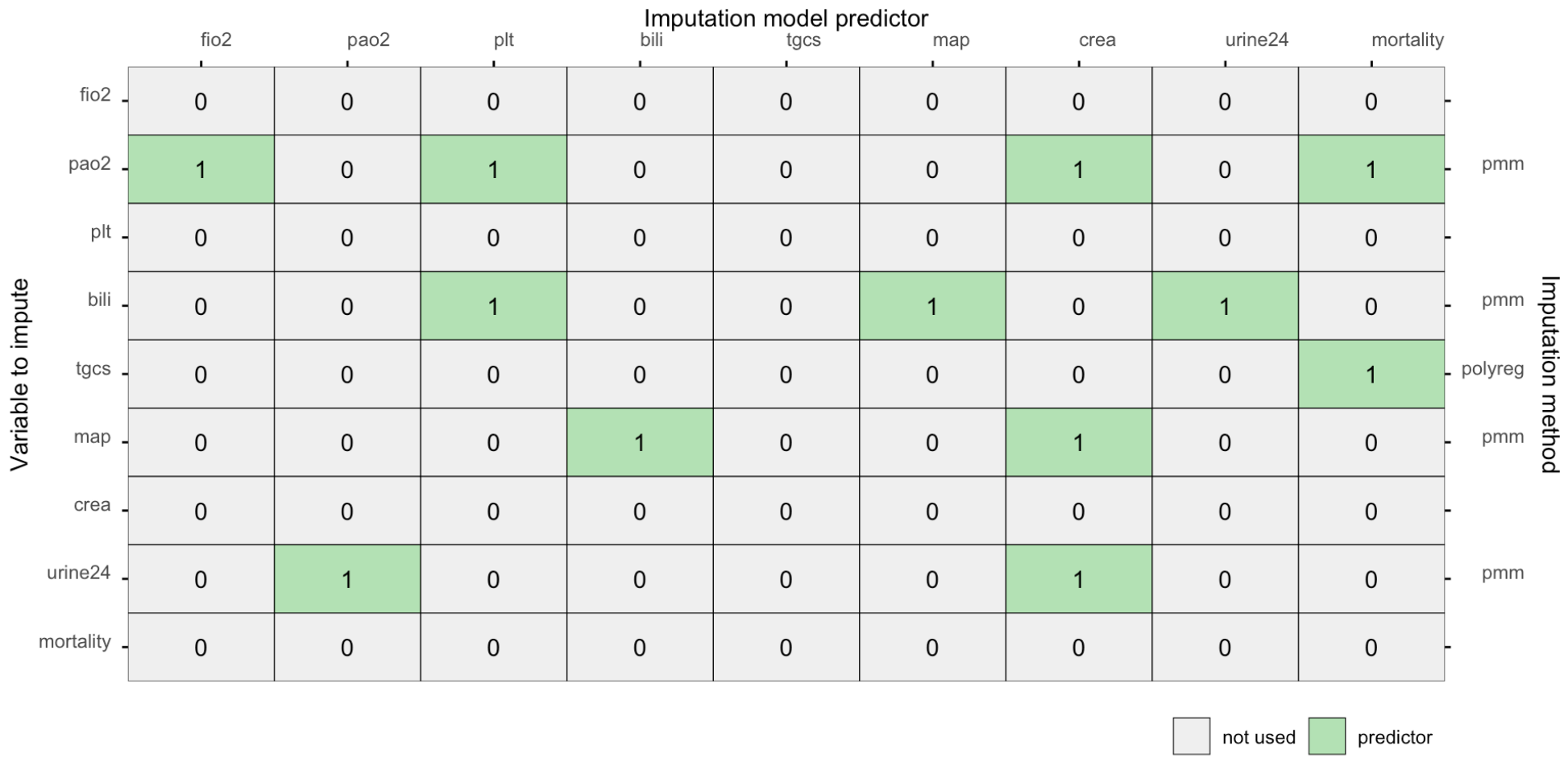
Correlation

```
1 plot_corr(dat, square = FALSE, label = TRUE)
```



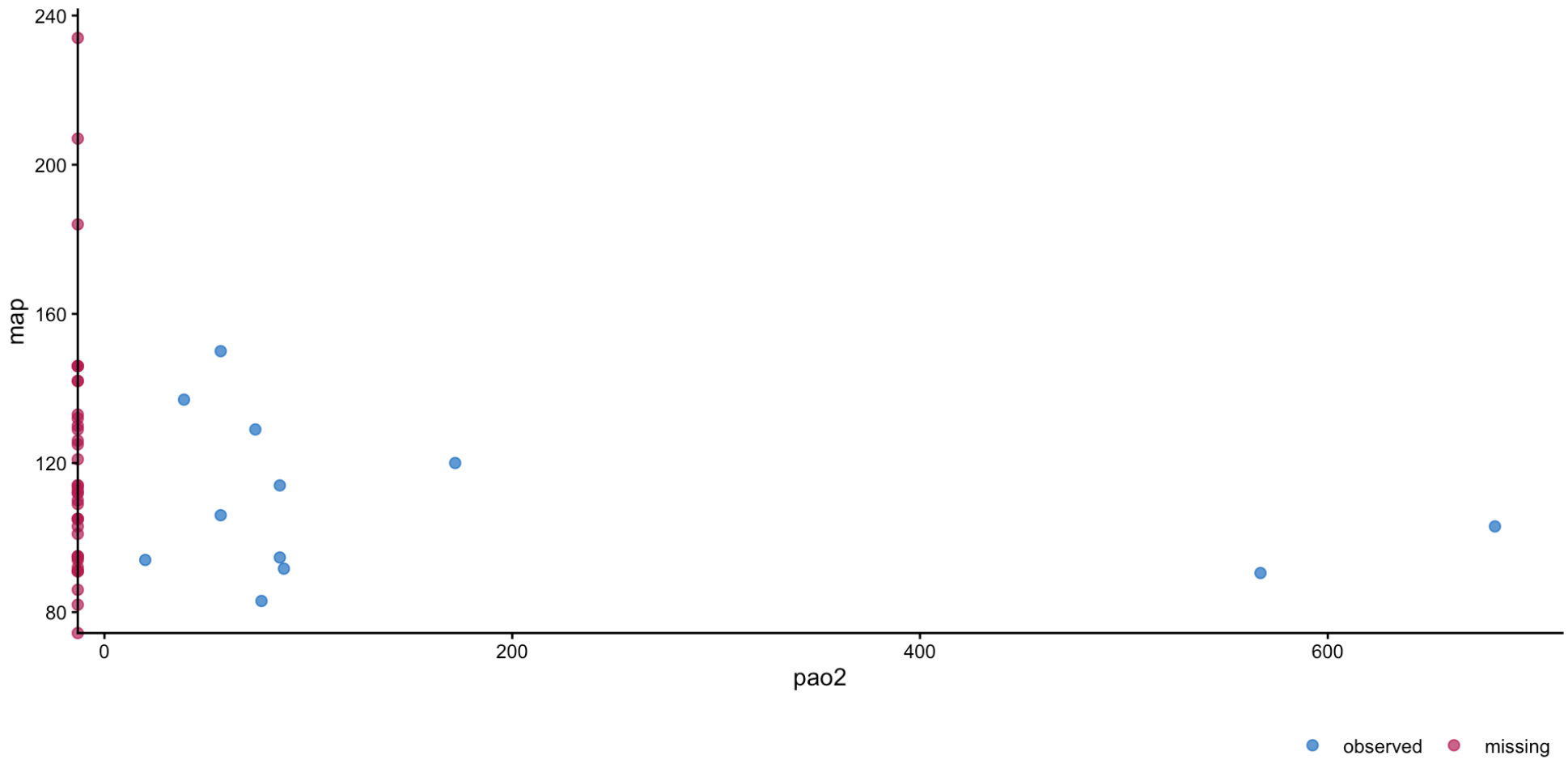
Imputation models

```
1 pred <- quickpred(dat, mincor = 0.3)
2 plot_pred(pred, method = meth, square = FALSE)
```



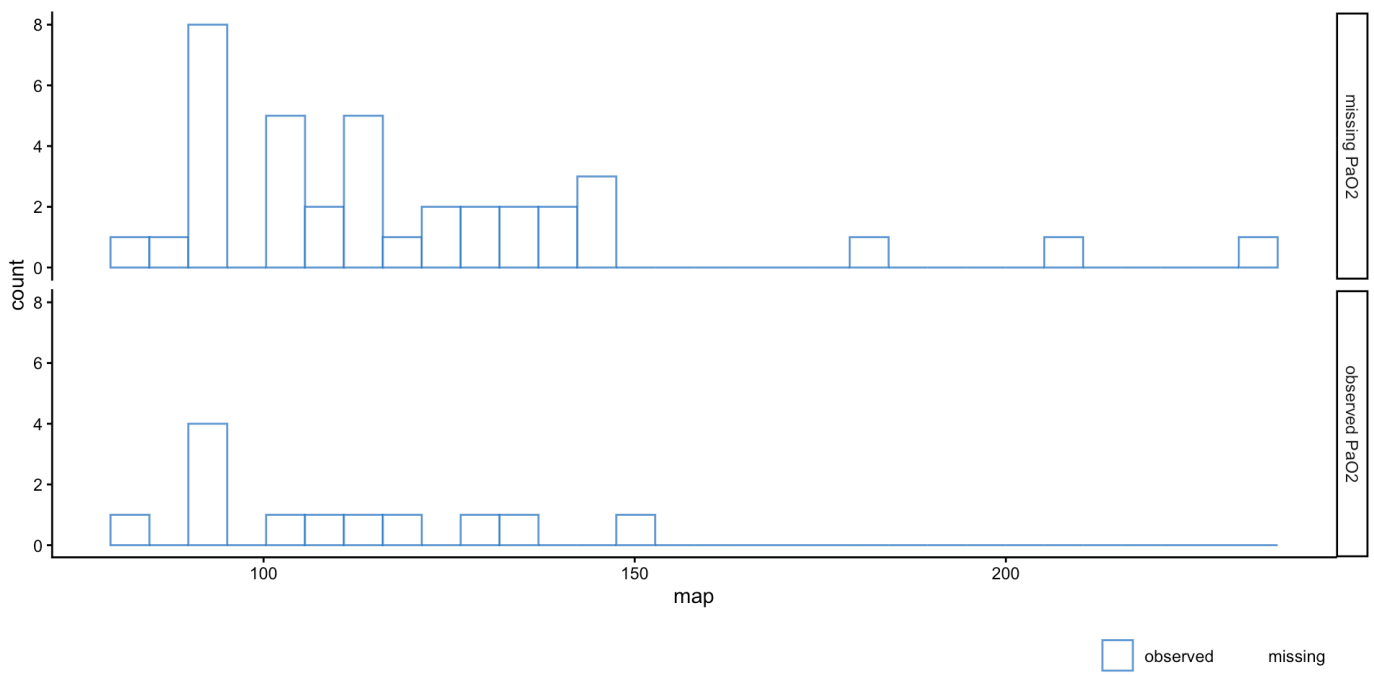
Scatter plot

```
1 ggmice(dat, aes(pao2, map)) +
2   geom_point(size = 2)
```



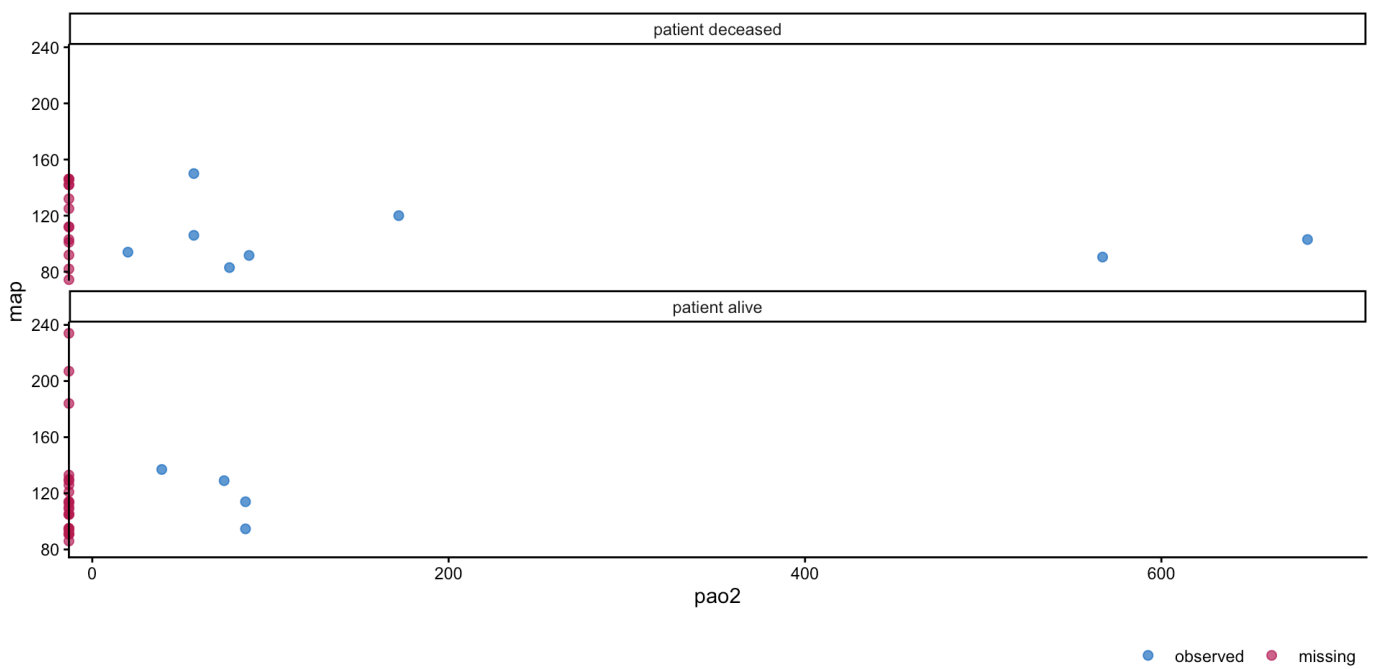
Faceted distribution

```
ggmice(dat, aes(map)) +
  geom_histogram(fill = "white") +
  facet_grid(factor(
    is.na(pao2),
    levels = c(TRUE, FALSE),
    labels = c("missing PaO2", "observed PaO2")
    ) ~ .)
```



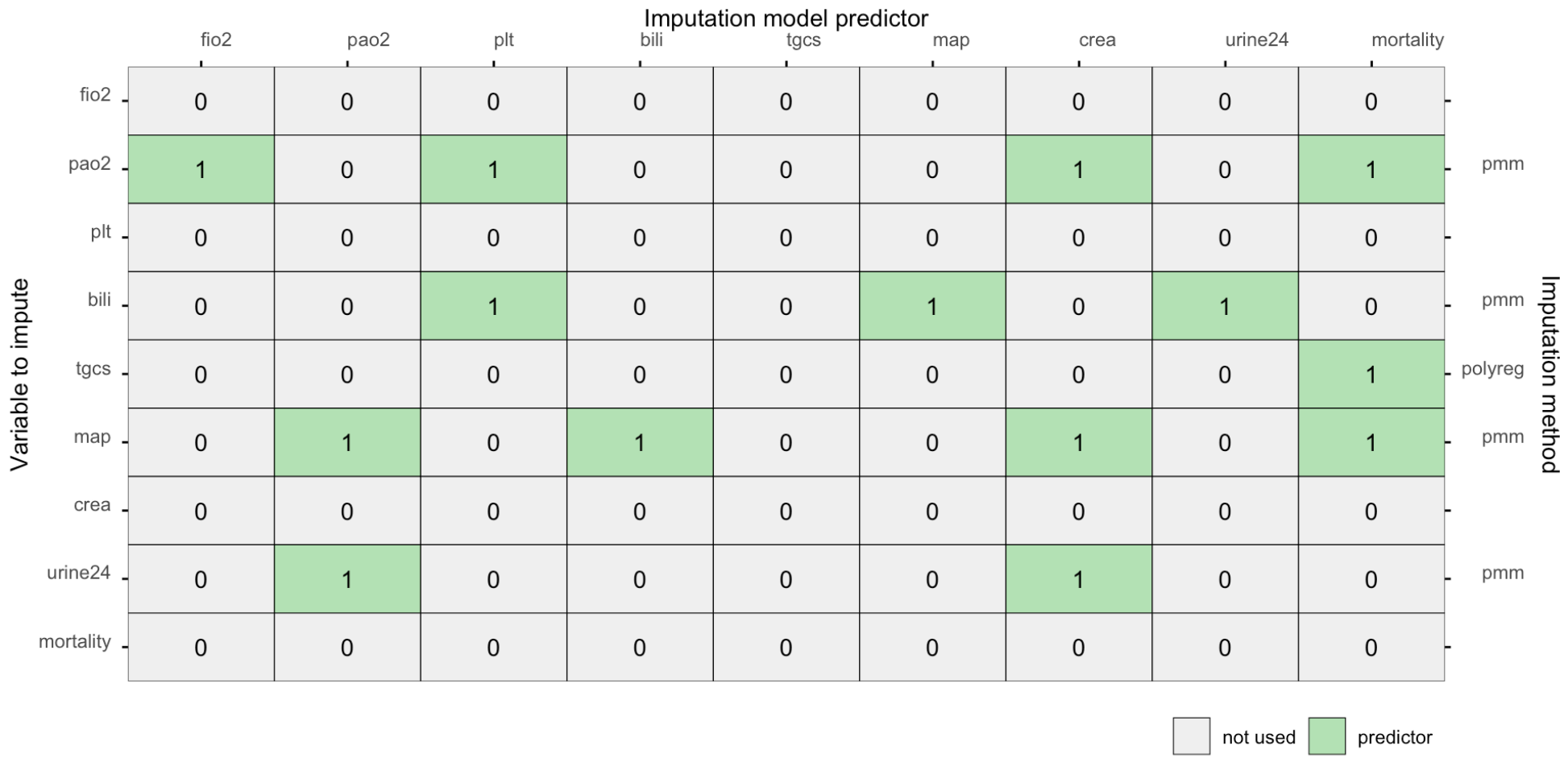
Faceted scatter plot

```
ggmice(dat, aes(pao2, map)) +
  geom_point(size = 2) +
  facet_wrap(~factor(
    mortality,
    levels = c(TRUE, FALSE),
    labels = c("patient deceased", "patient alive")
    ), ncol = 1)
```



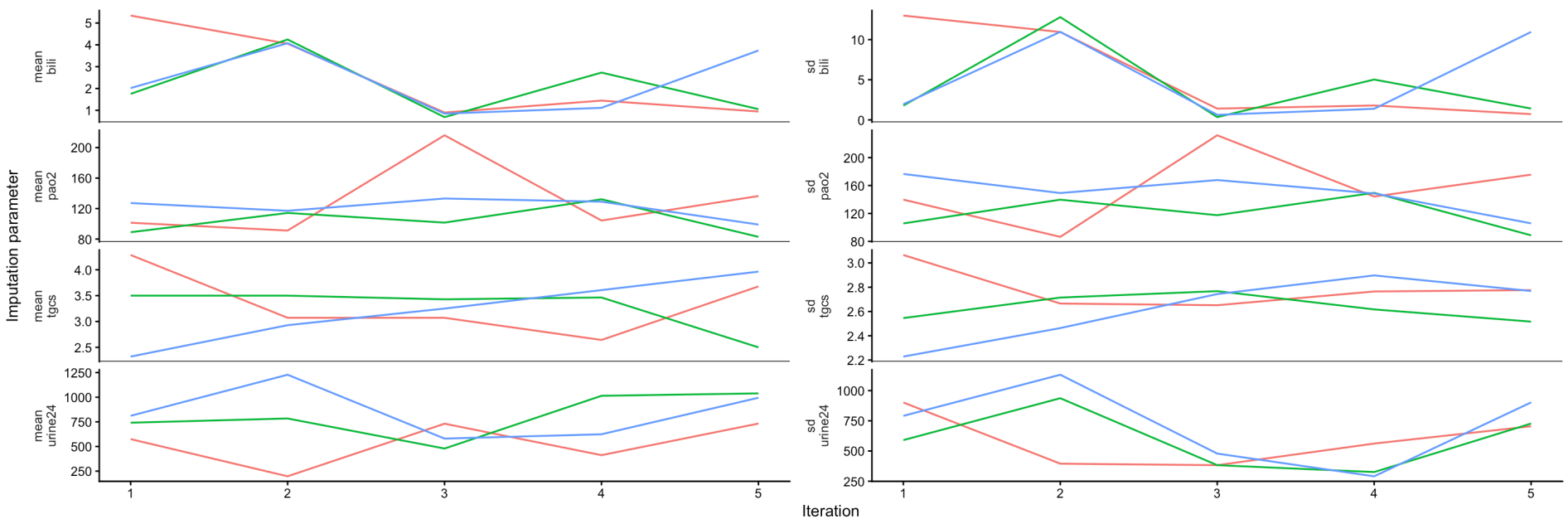
Adjust imputation models

```
1 pred["map", c("pao2", "mortality")] <- 1
2 plot_pred(pred, method = meth, square = FALSE)
```



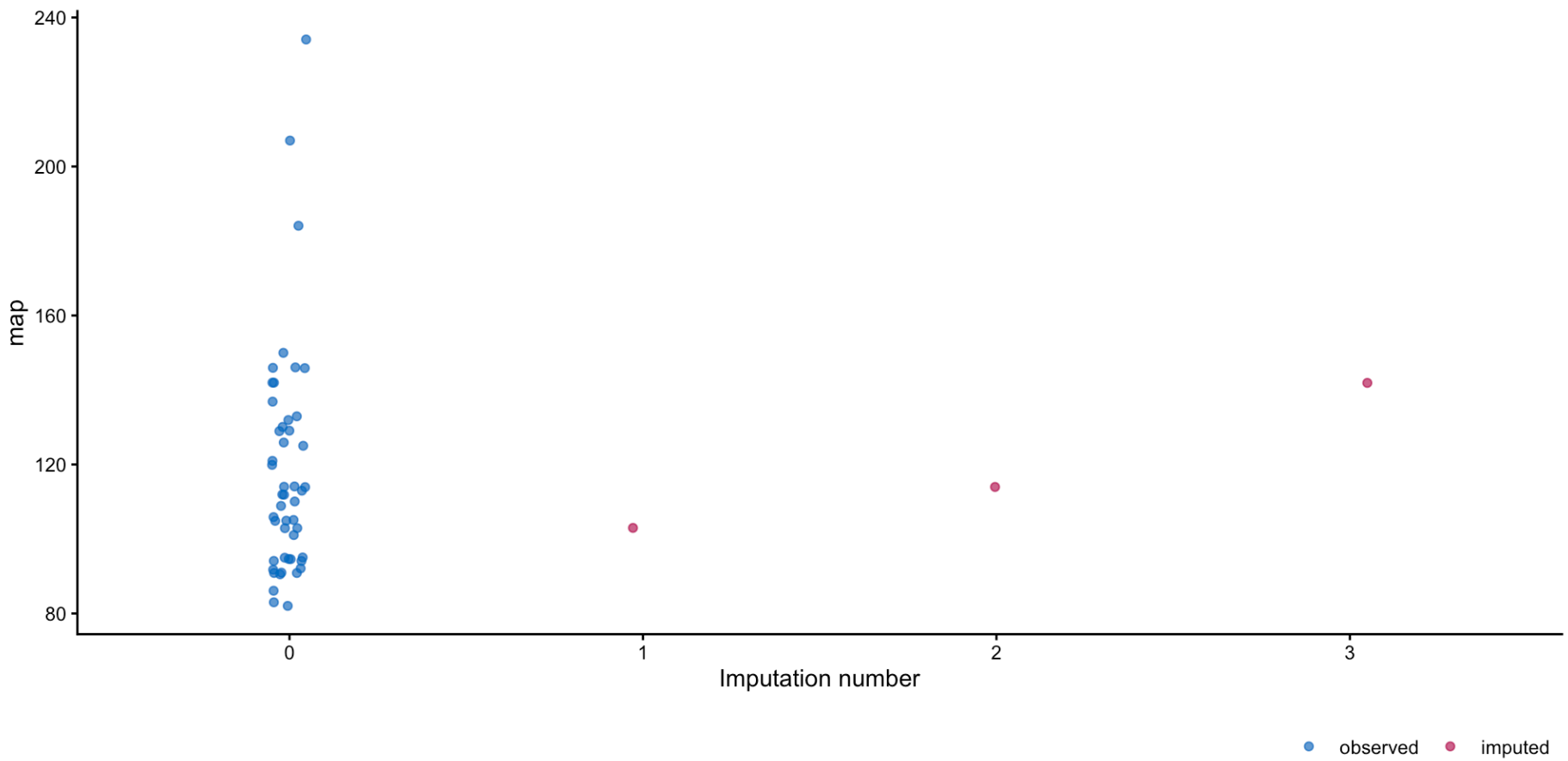
Impute

```
imp <- mice(
  dat,
  pred = pred,
  method = meth,
  m = 3,
  seed = 11,
  print = FALSE)
plot_trace(imp, legend = FALSE)
```



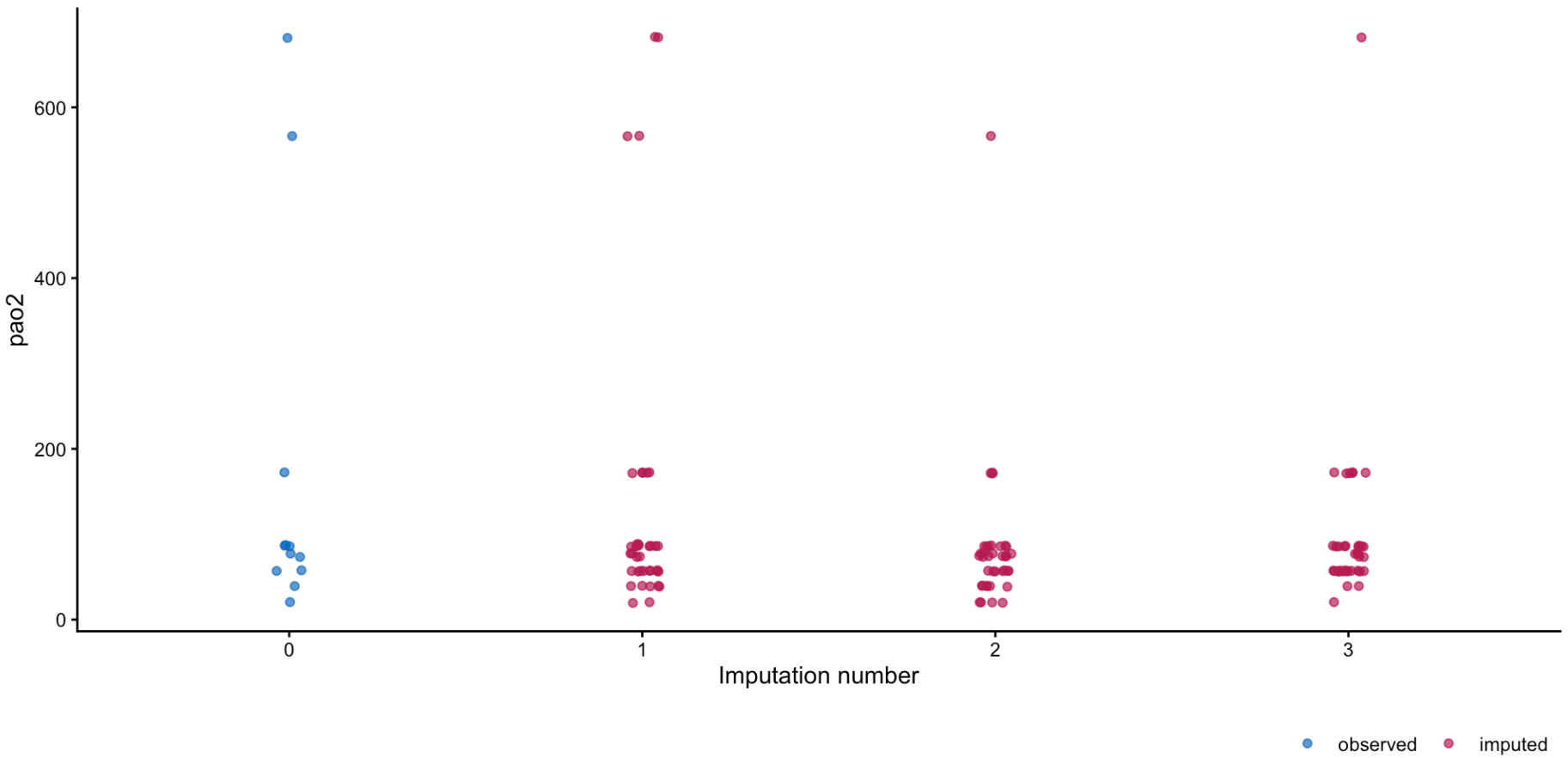
Stripplot of blood pressure

```
ggmice(imp, aes(x = .imp, y = map)) +
  geom_jitter(width = 0.05) +
  labs(x = "Imputation number")
```

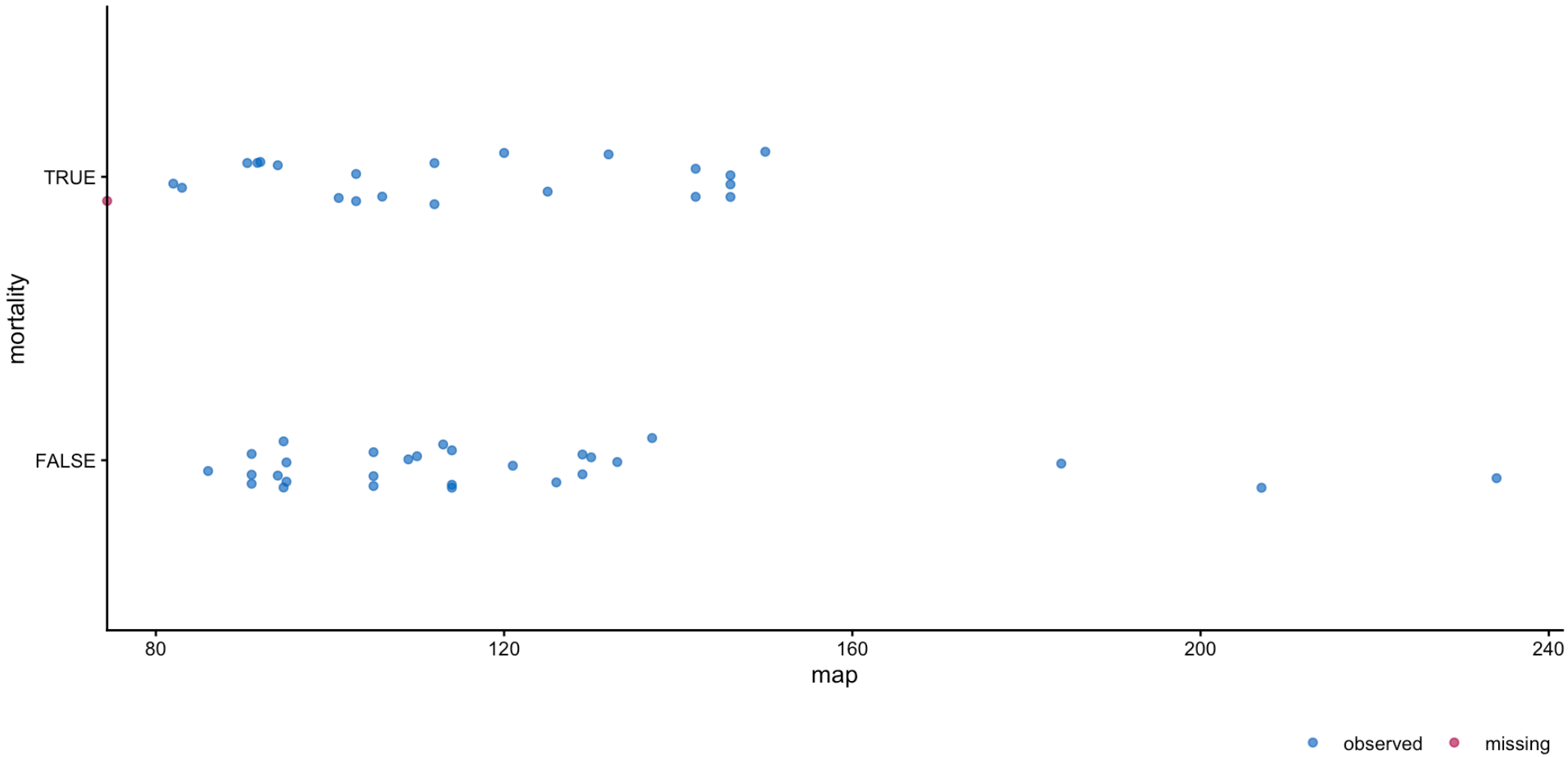


Stripplot of oxygenation

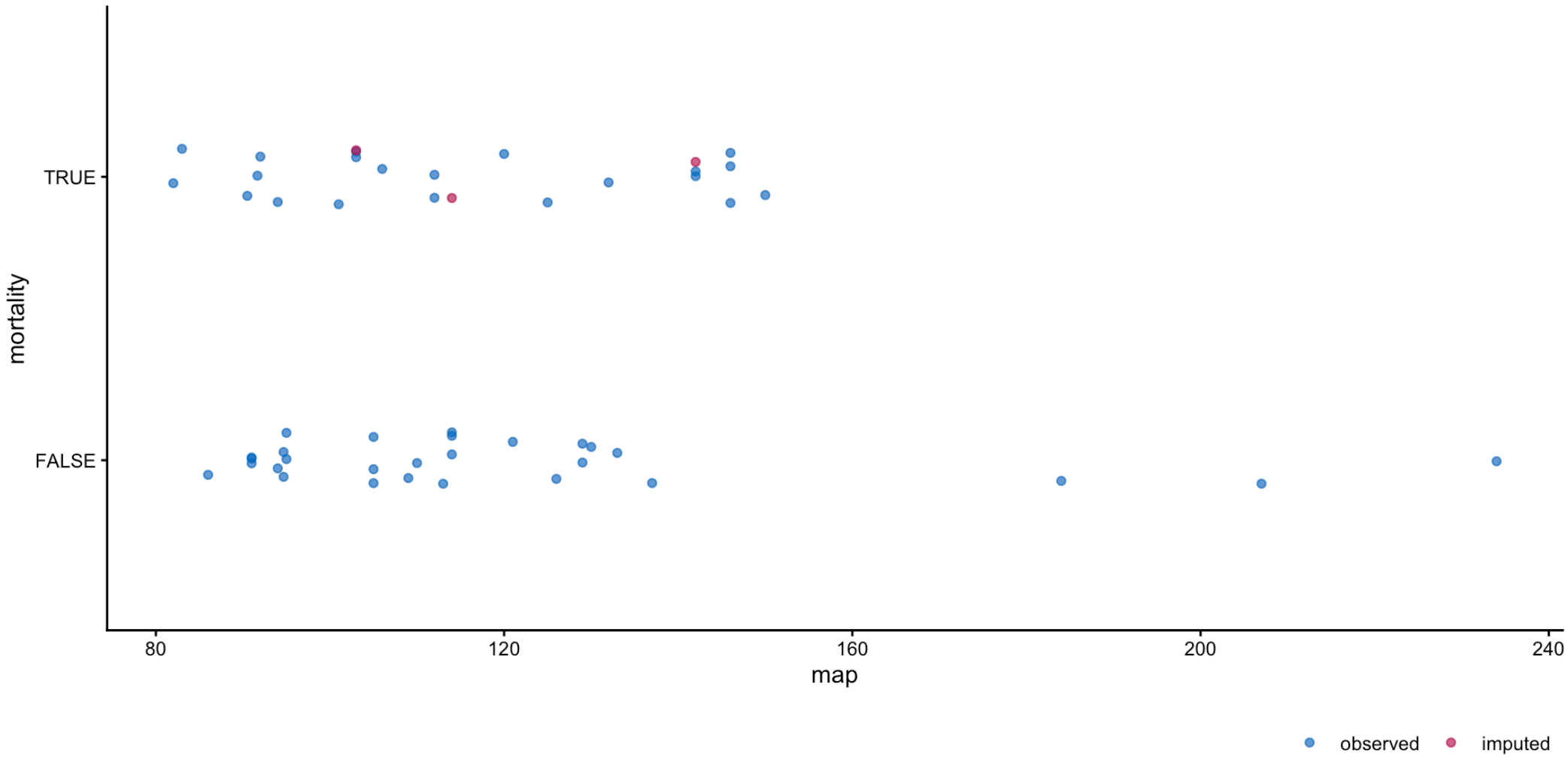
```
ggmice(imp, aes(x = .imp, y = pao2)) +
  geom_jitter(width = 0.05) +
  labs(x = "Imputation number")
```



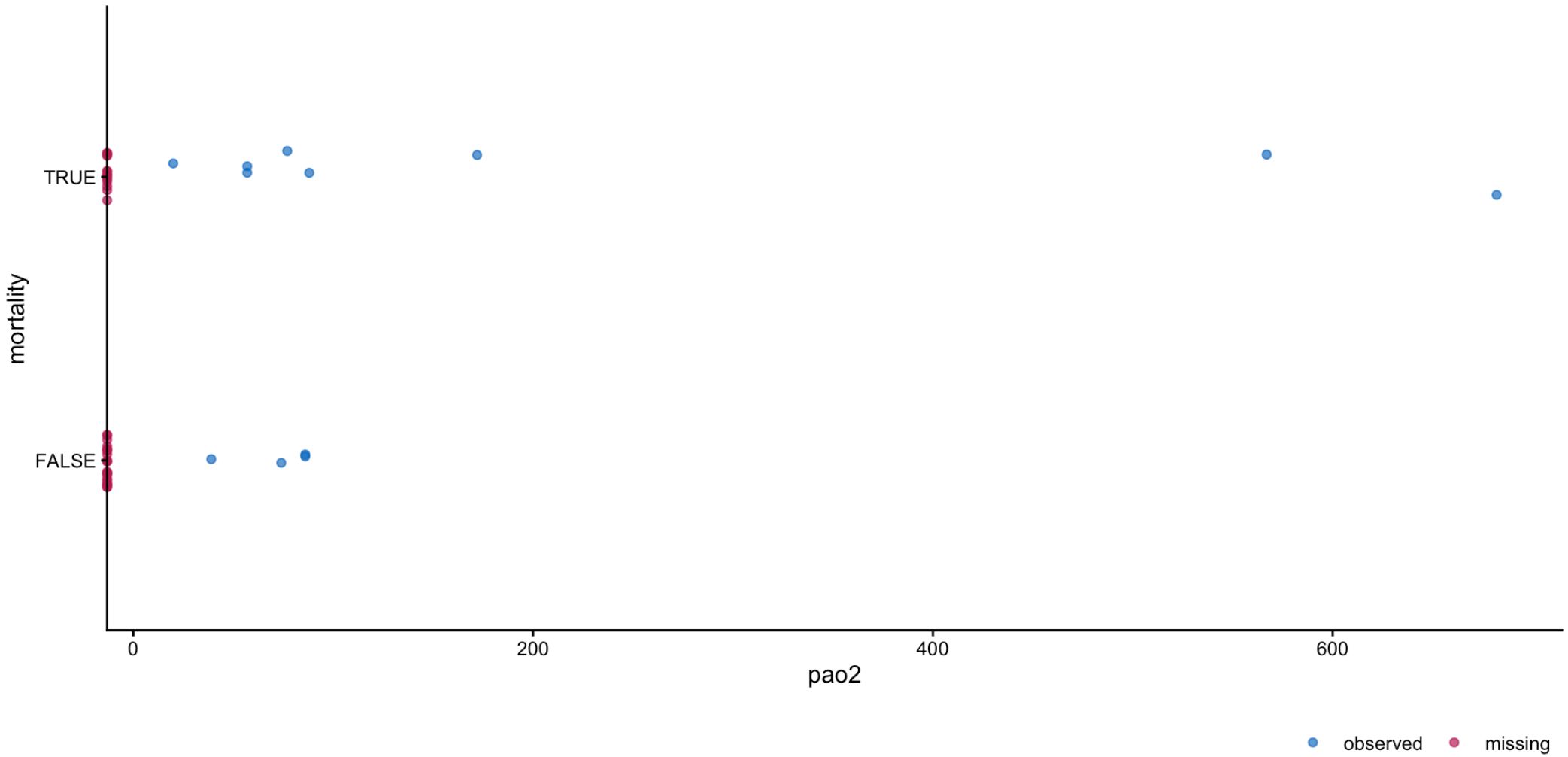
Blood pressure by mortality



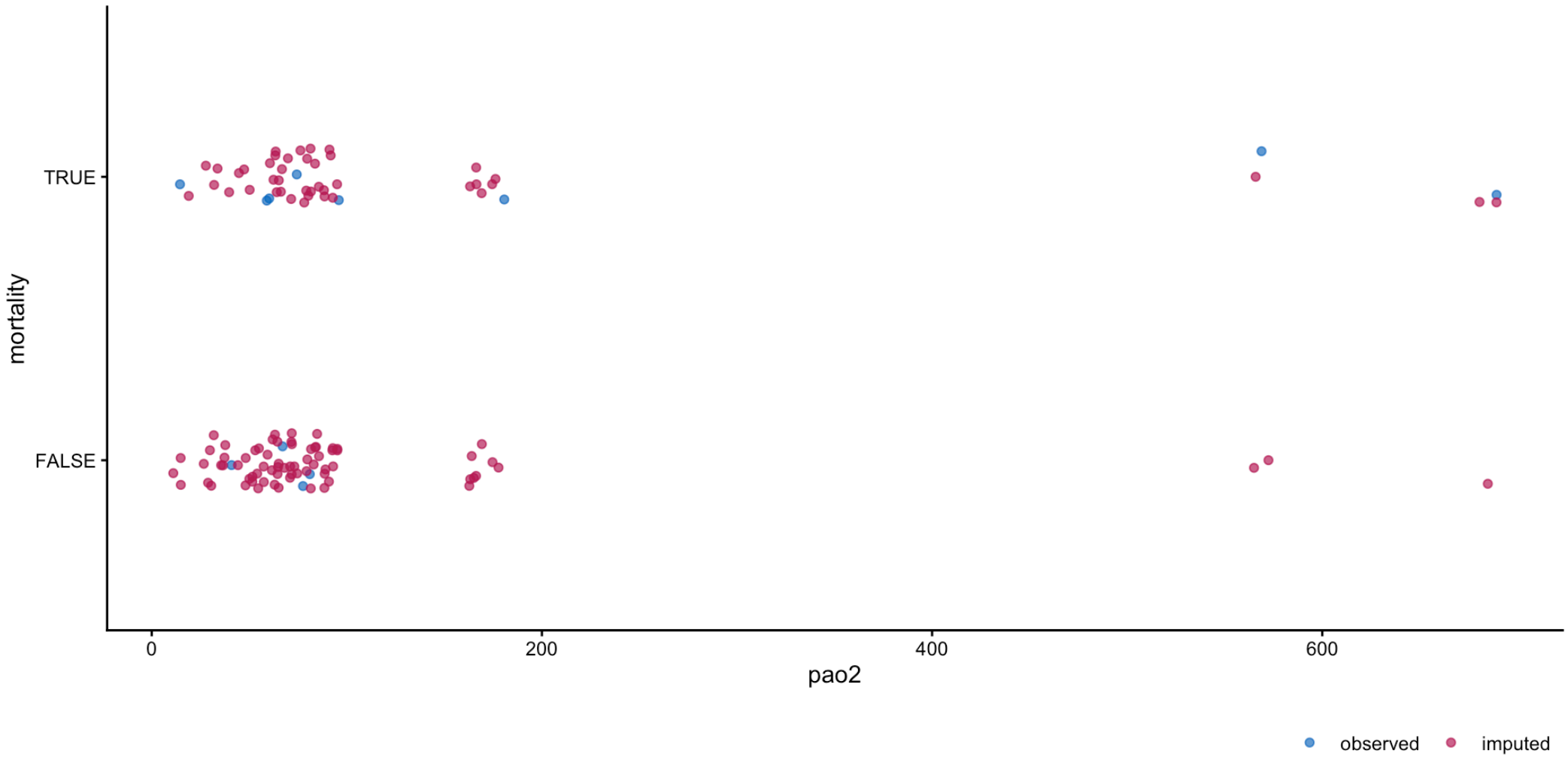
Blood pressure by mortality



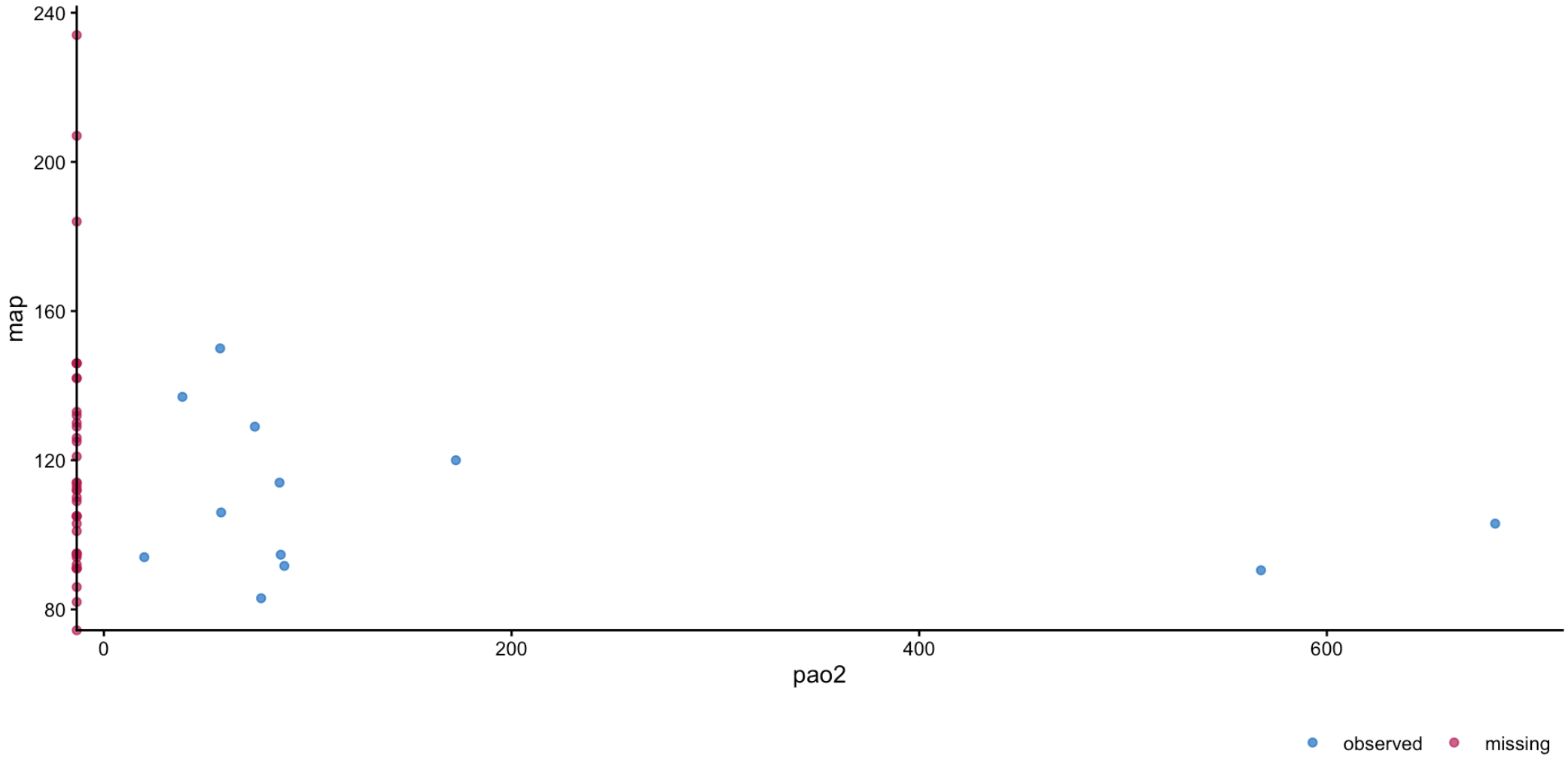
Oxygenation by mortality



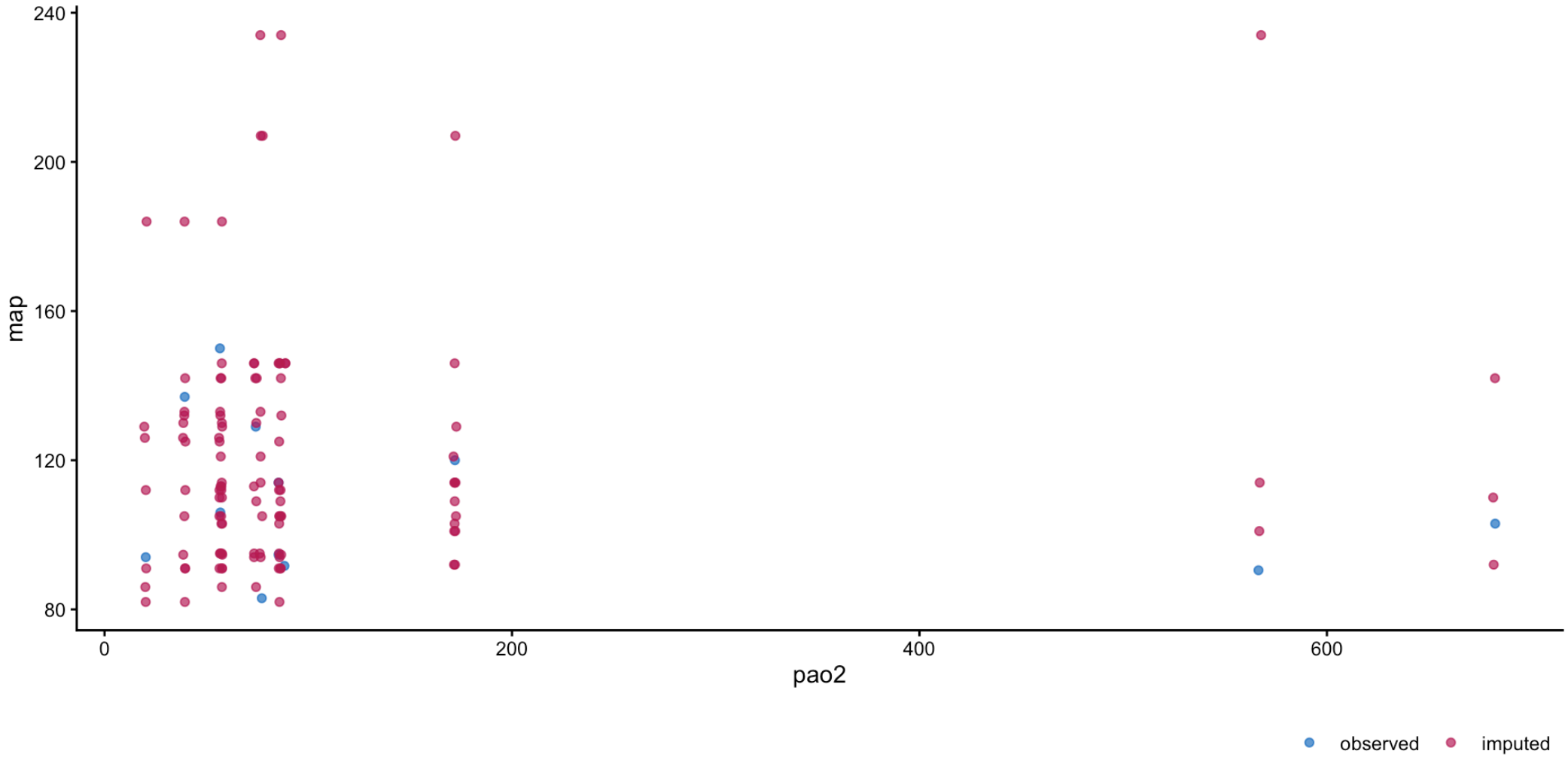
Oxygenation by mortality



Scatter plot

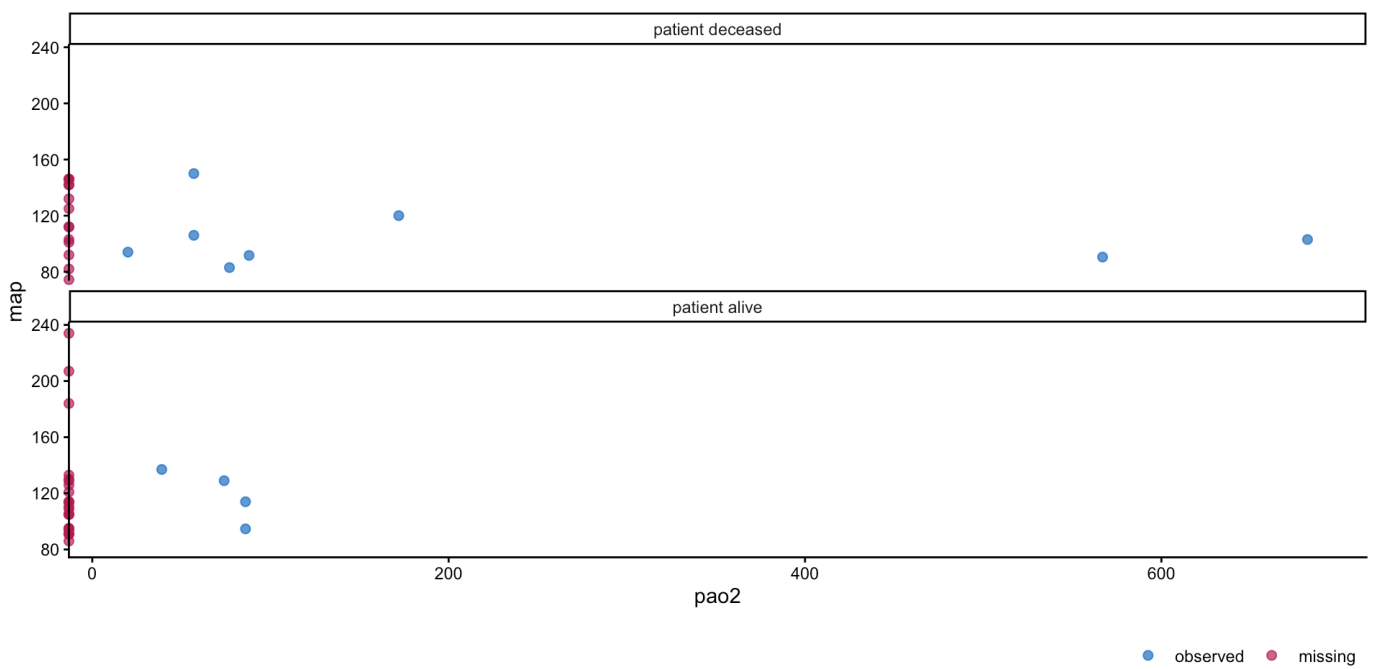


Scatter plot



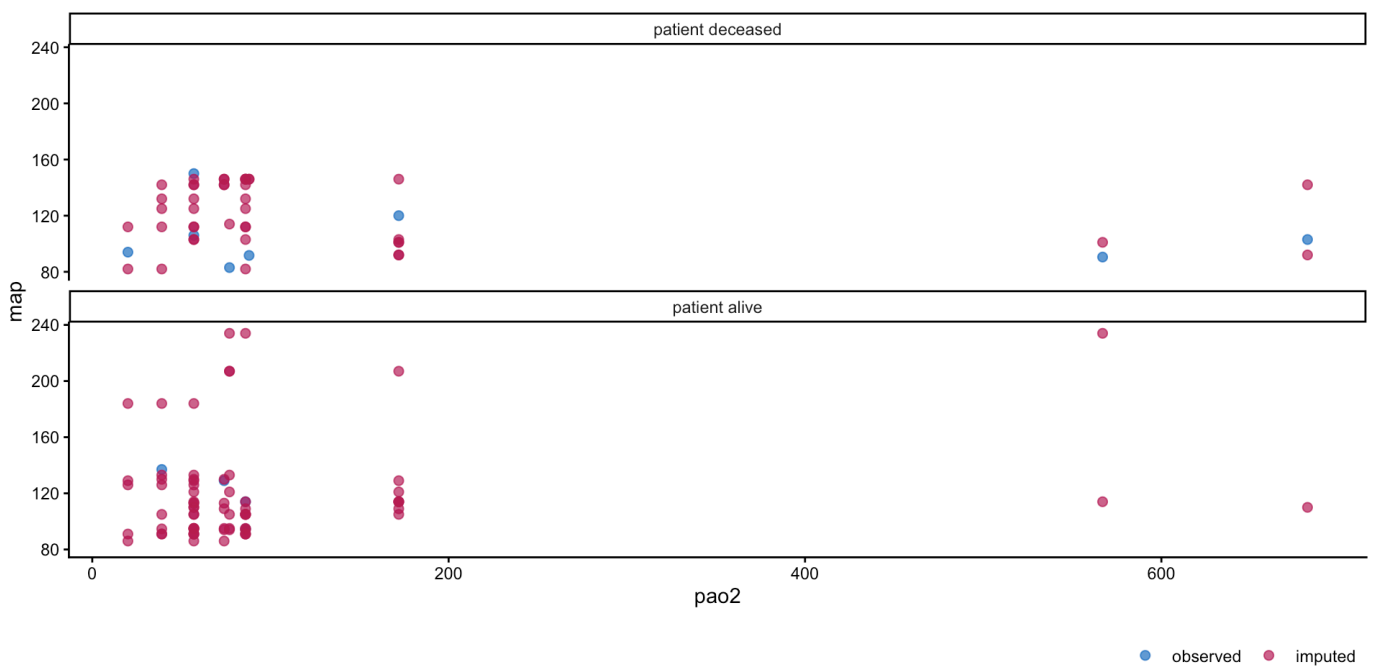
Faceted scatter plot

```
ggmice(dat, aes(pao2, map)) +
  geom_point(size = 2) +
  facet_wrap(~factor(
    mortality,
    levels = c(TRUE, FALSE),
    labels = c("patient deceased", "patient alive")
    ), ncol = 1)
```



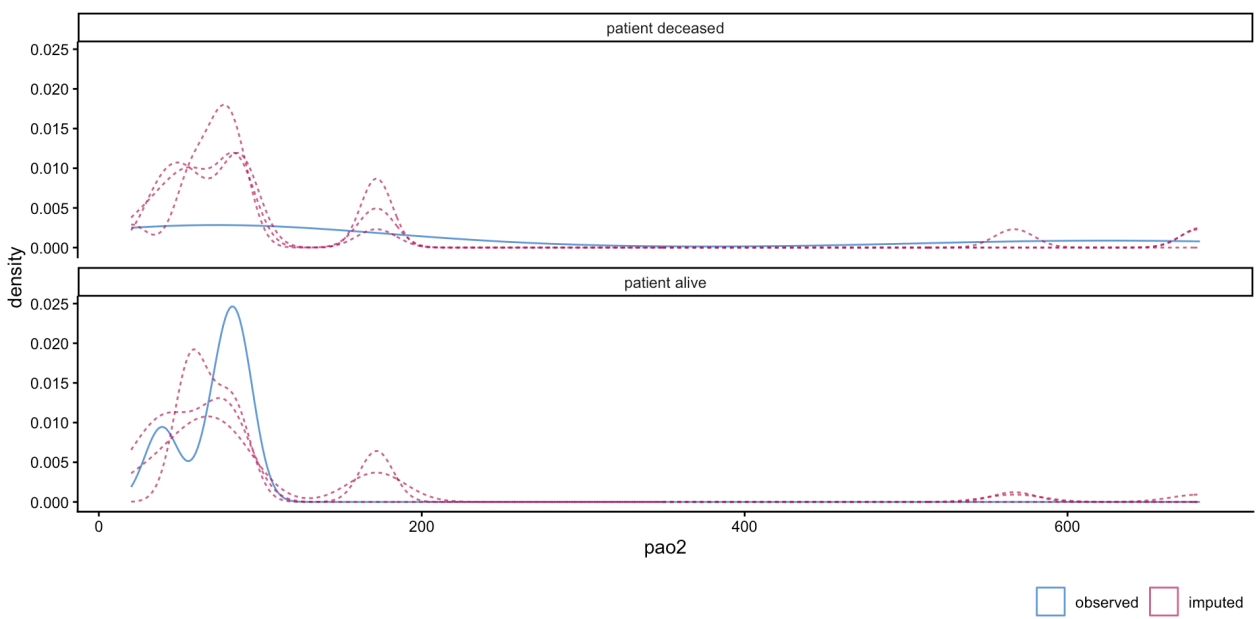
Faceted scatter plot

```
ggmice(imp, aes(pao2, map)) +
  geom_point(size = 2) +
  facet_wrap(~factor(
    mortality,
    levels = c(TRUE, FALSE),
    labels = c("patient deceased", "patient alive")
    ), ncol = 1)
```



Faceted distribution

```
ggmice(imp, aes(pao2, group = .imp, linetype = !(.imp == 0))) +
  geom_density() +
  facet_wrap(~factor(
    mortality,
    levels = c(TRUE, FALSE),
    labels = c("patient deceased", "patient alive")
    ), ncol = 1) +
  scale_linetype(guide = "none")
```

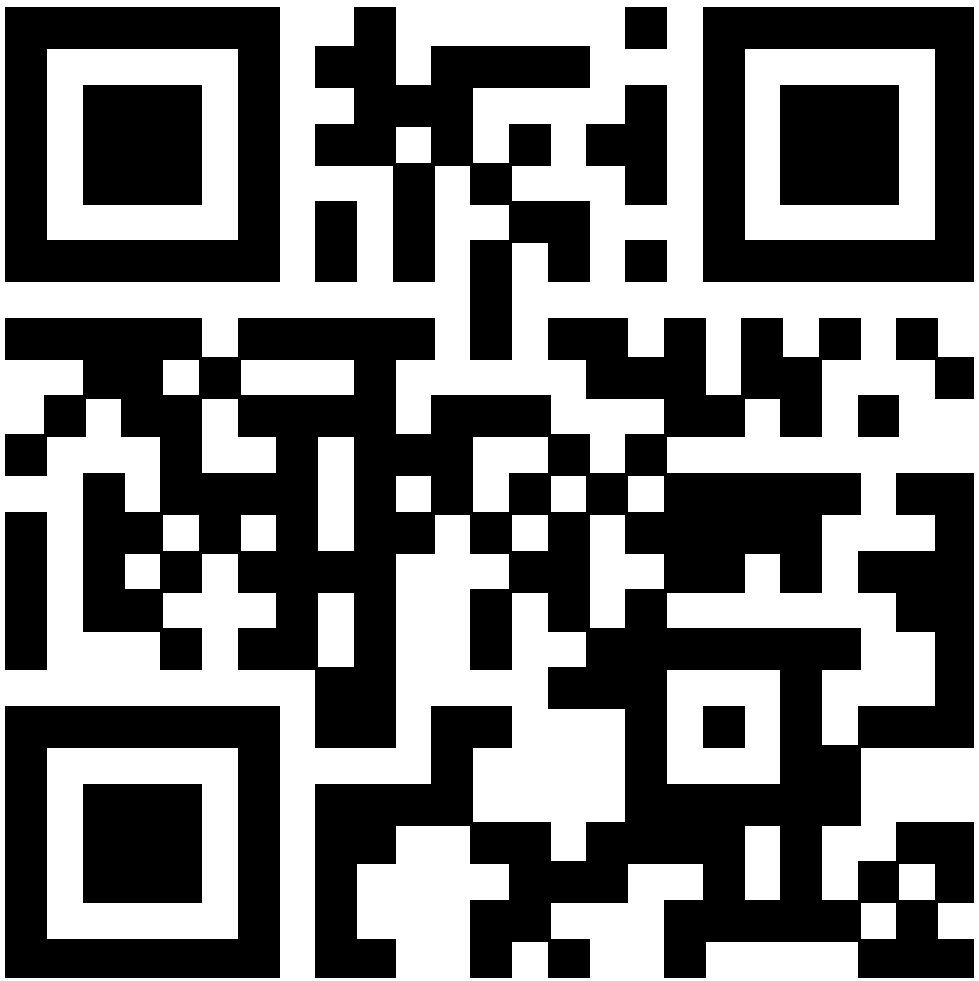


Take aways

Missing data are

- a pervasive problem
- visualizable & analyzable
- informative!

Thank you!



edu.nl/nbd3q