

Data visualization

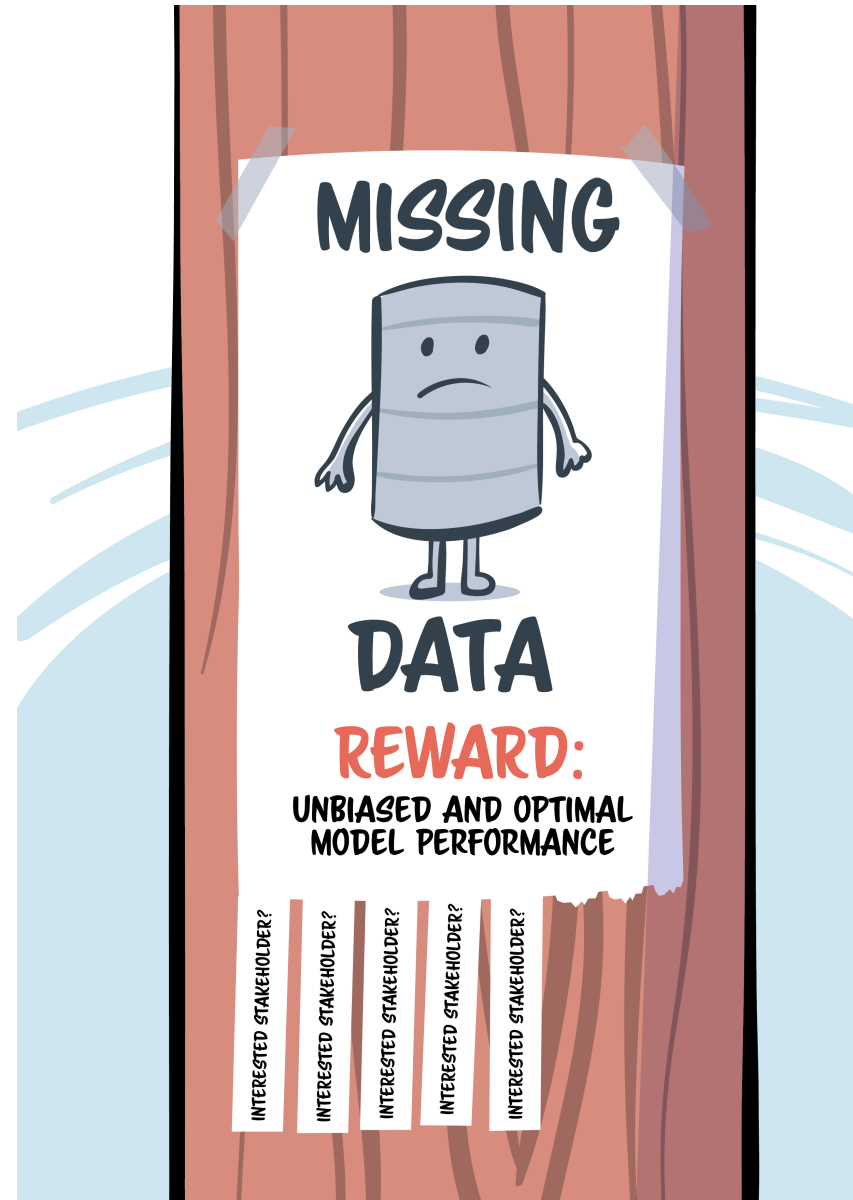
for incomplete datasets in R

Hanne Oberman 

Utrecht University



Missingness



A problem to fix

- unit non-response

→ weighting etc.

- item non-response

→ imputation etc.



Take aways

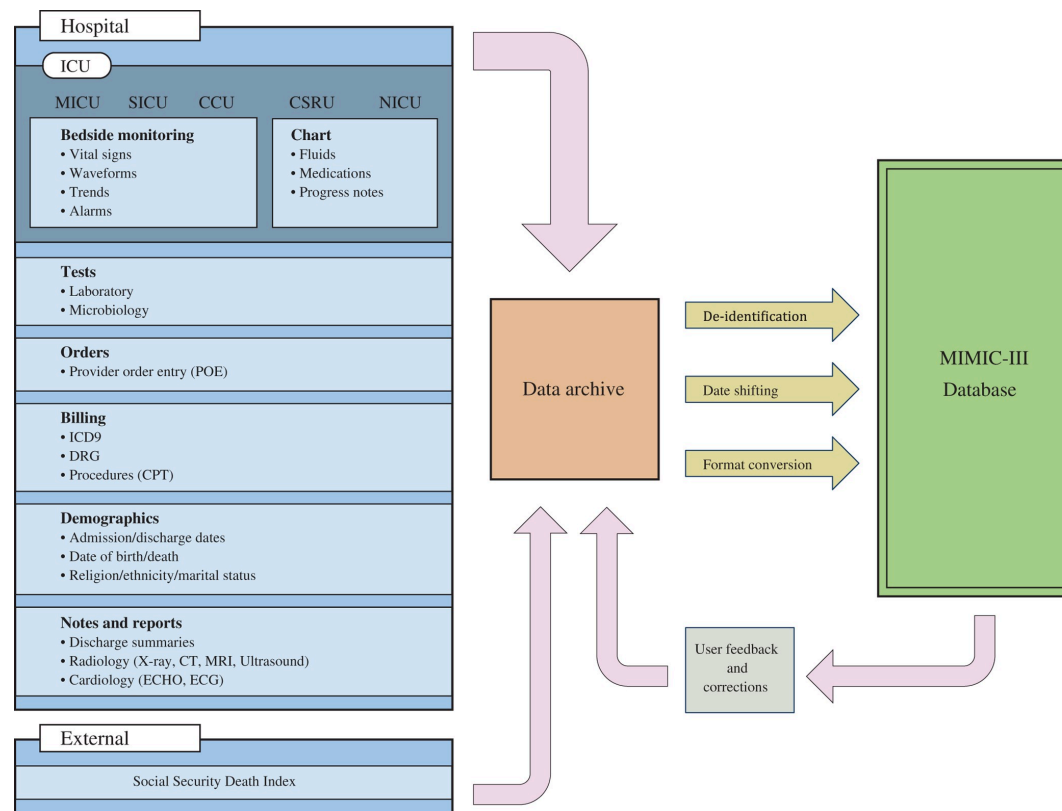
Missing data are

- a pervasive problem
- visualizable & analyzable
- informative!



Case study

```
1 set.seed(123)
2 library(ricu)
3 library(mice)
4 library(ggmice)
5 library(ggplot2)
6 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 ...
 $ 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

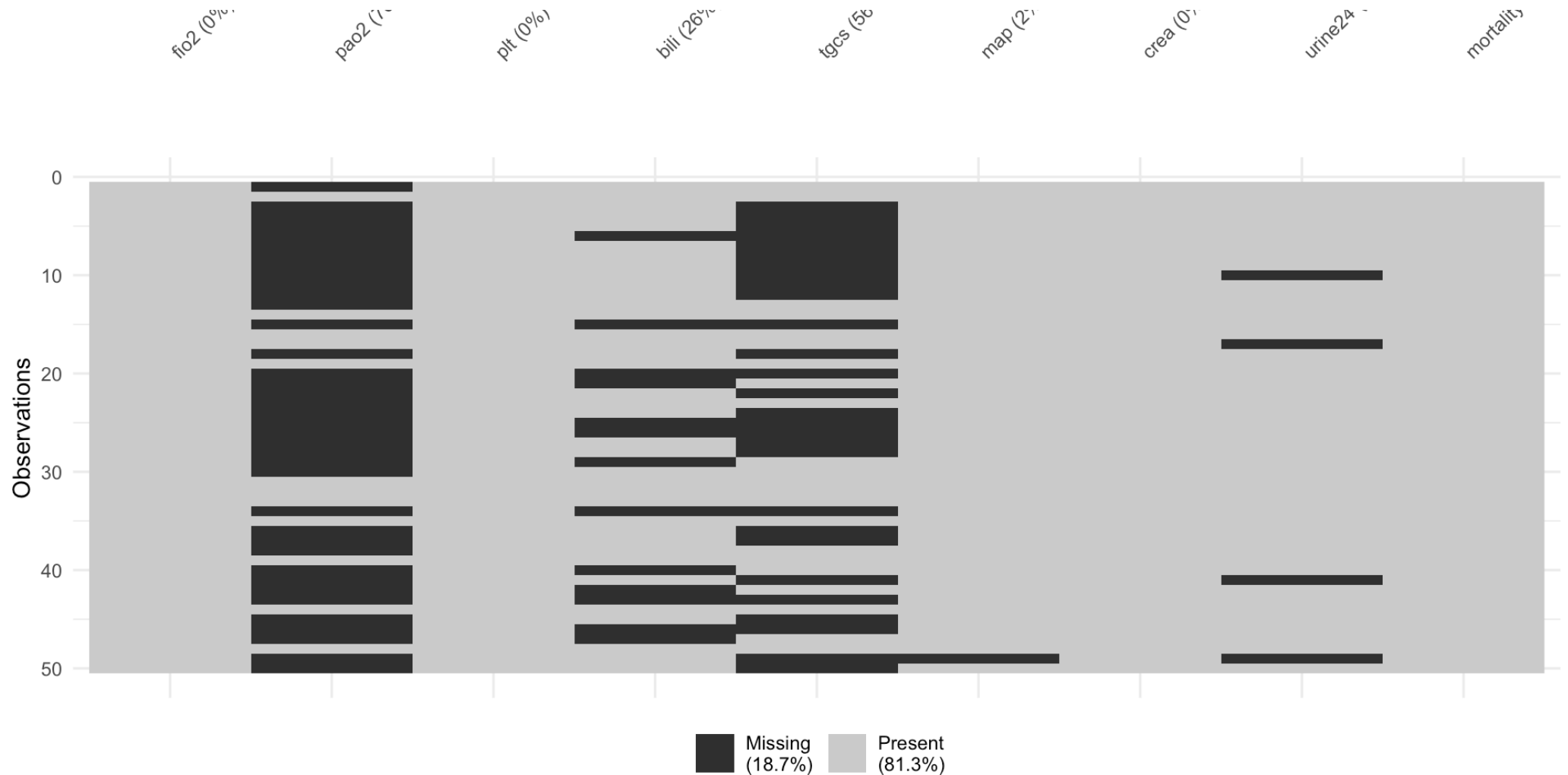
```
1 is.na(dat)
```

	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

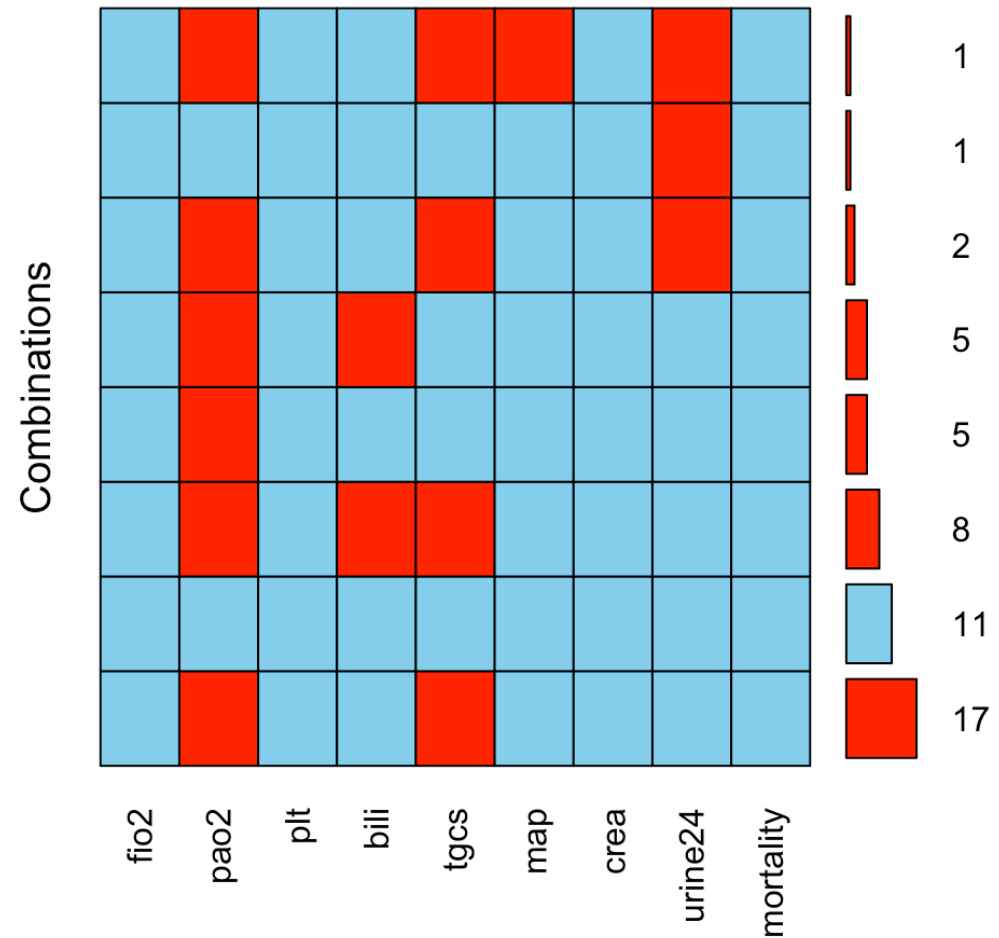
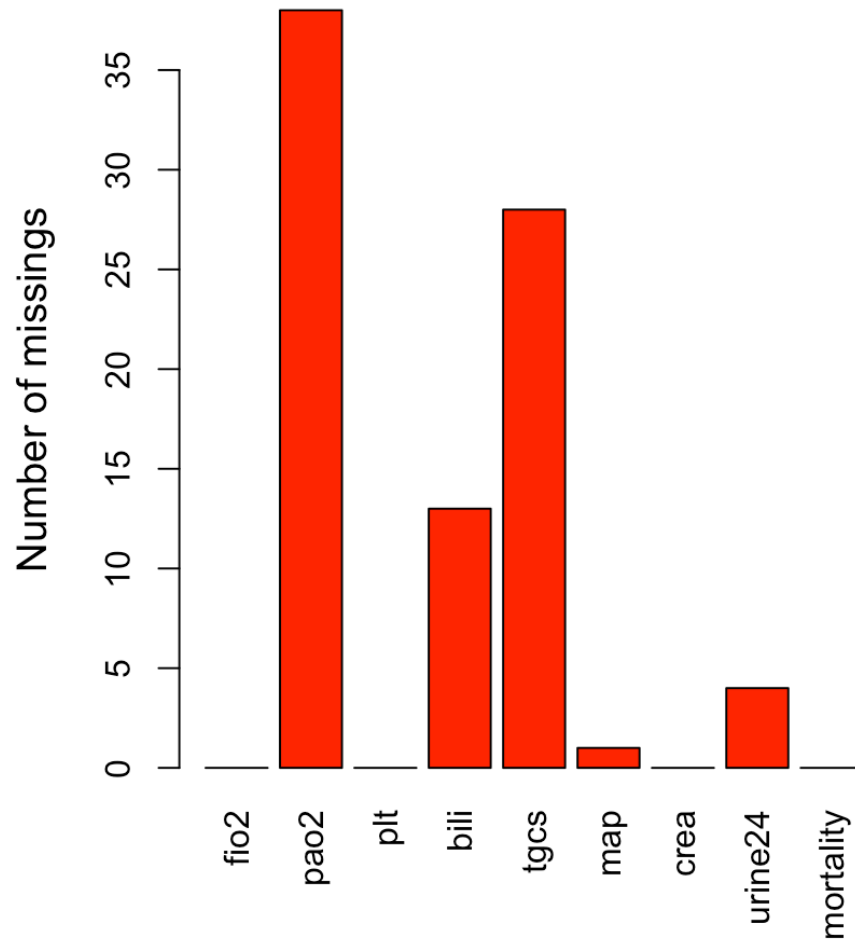
```
1 colSums(is.na(dat))
```

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



Missingness rate

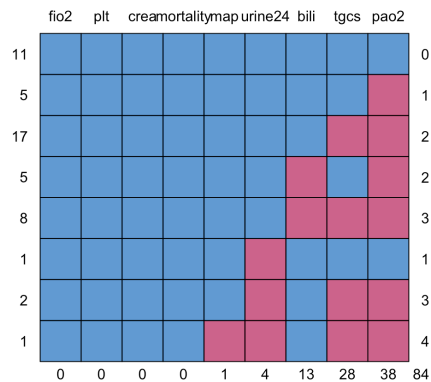
```
1 VIM::aggr(dat, numbers = TRUE, prop = FALSE)
```



Missing data pattern

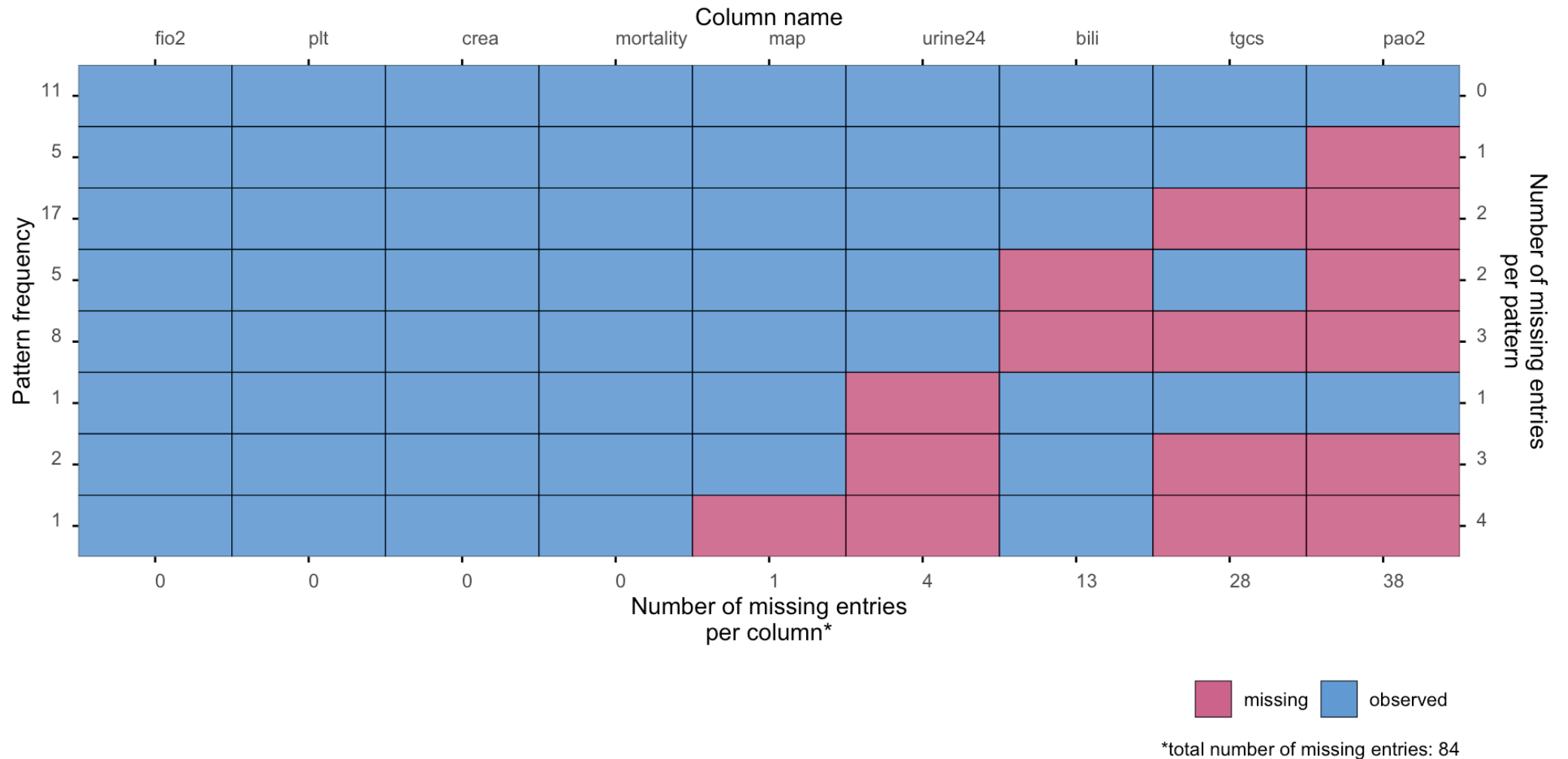
```
1 mice::md.pattern(dat)
```

	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

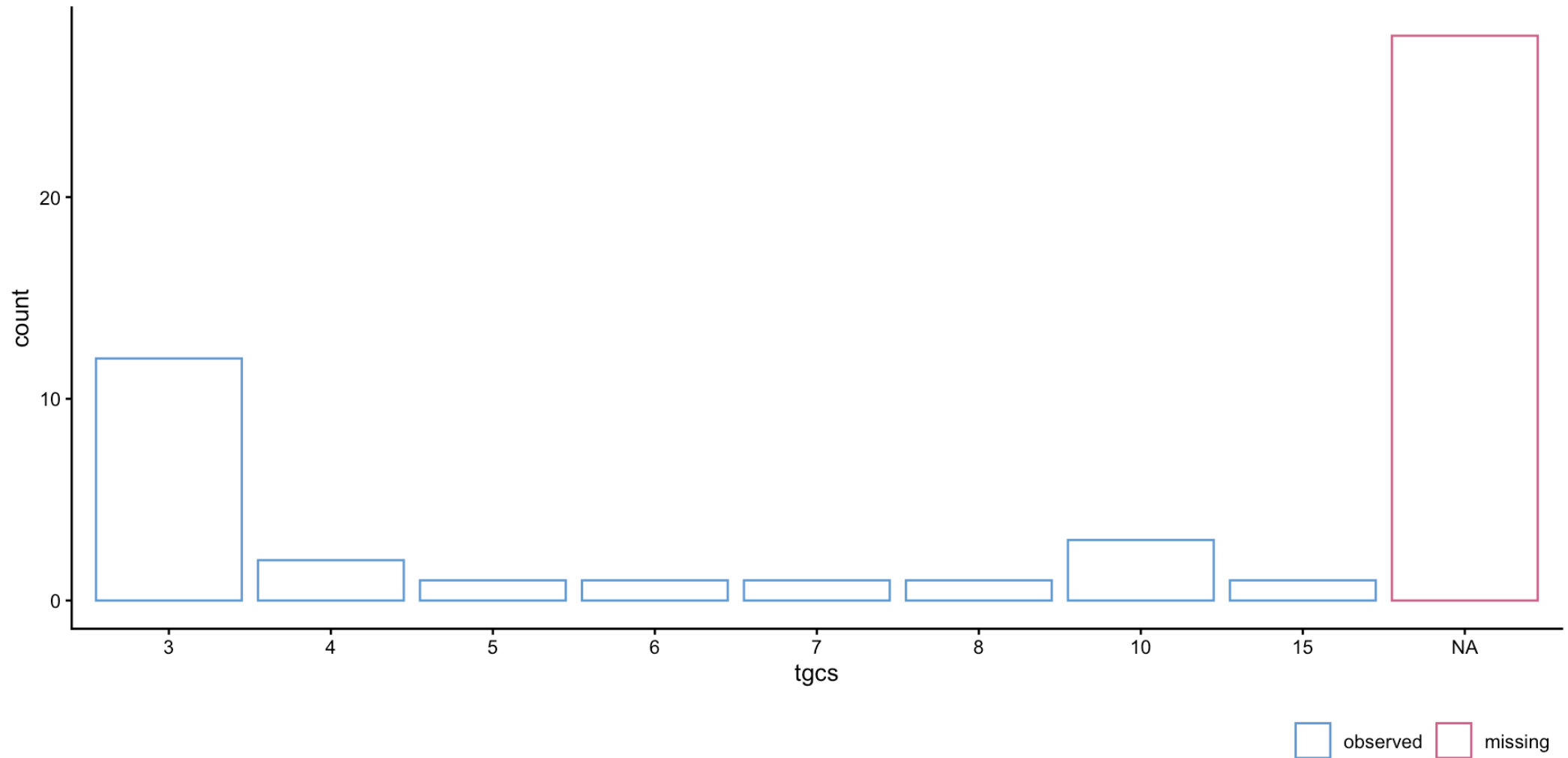


Missing data pattern

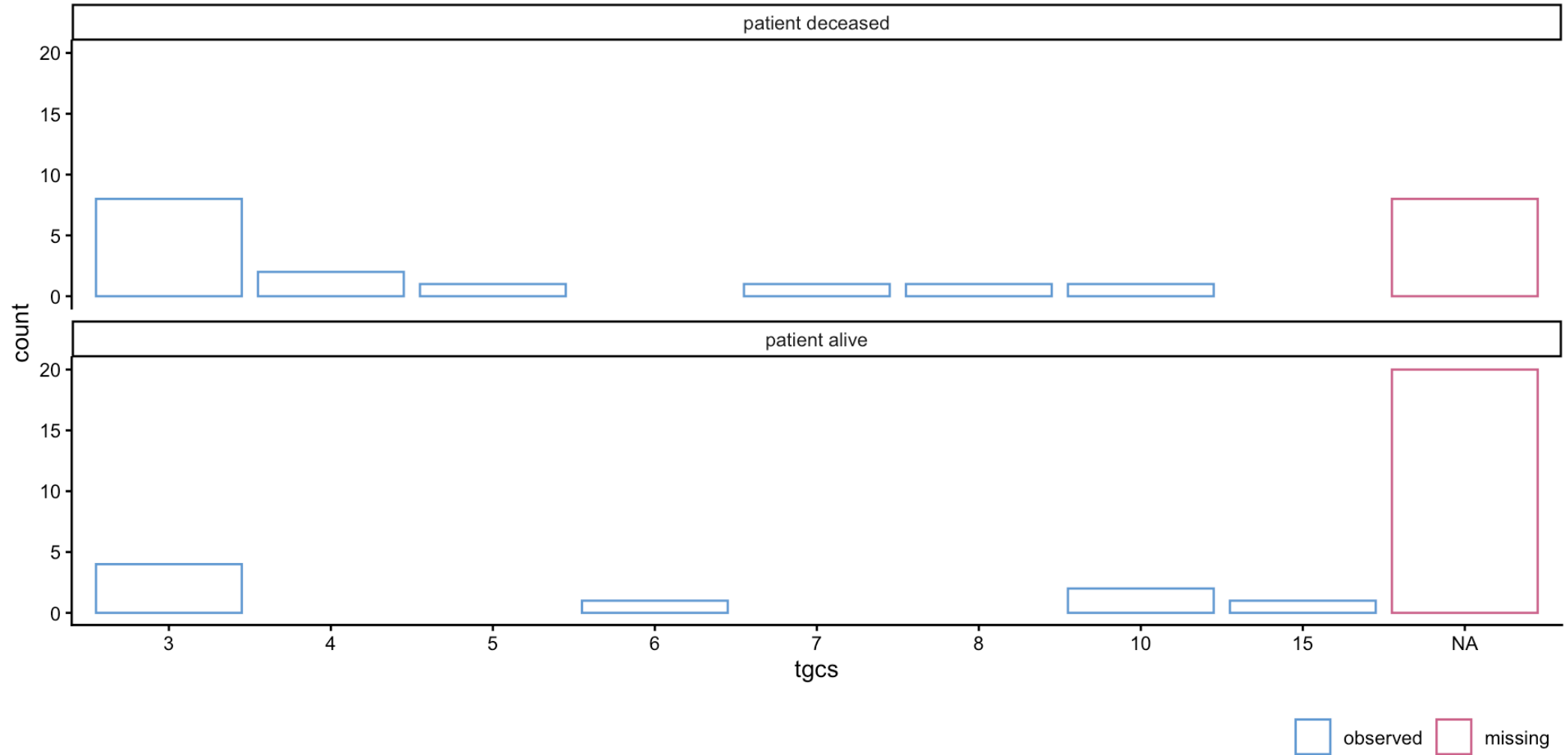
```
1 ggmlce::plot_pattern(dat, square = FALSE)
```



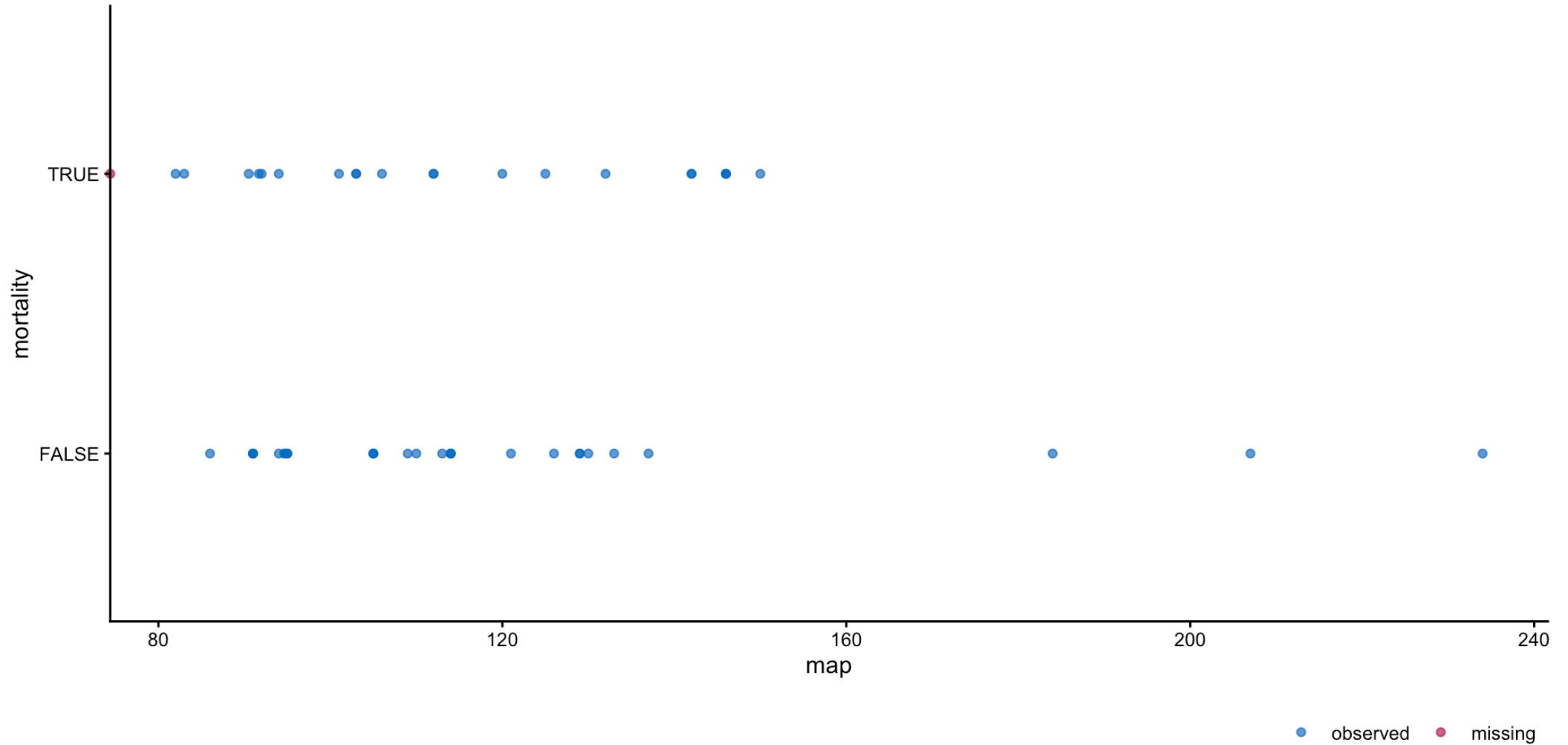
Coma symptoms



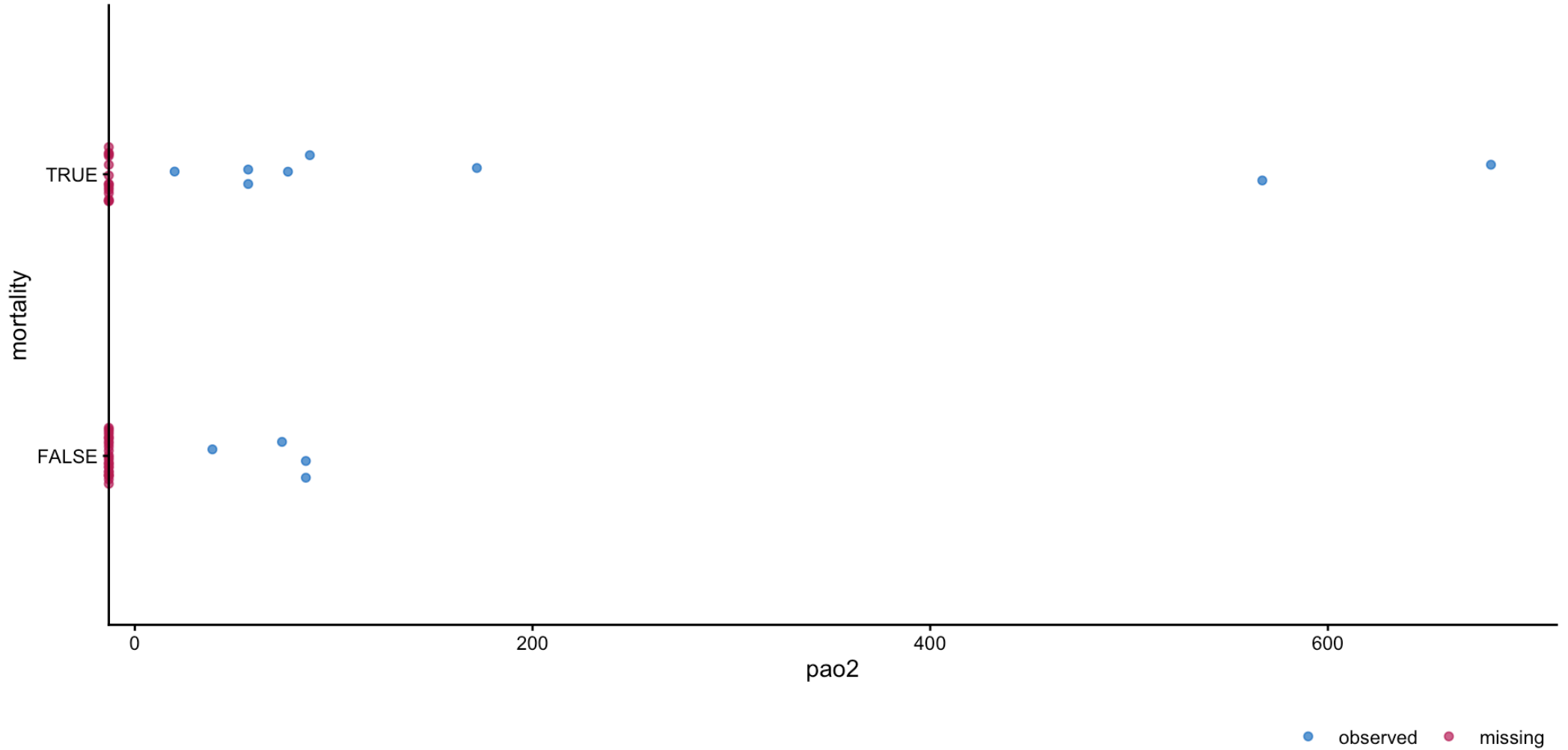
Coma symptoms



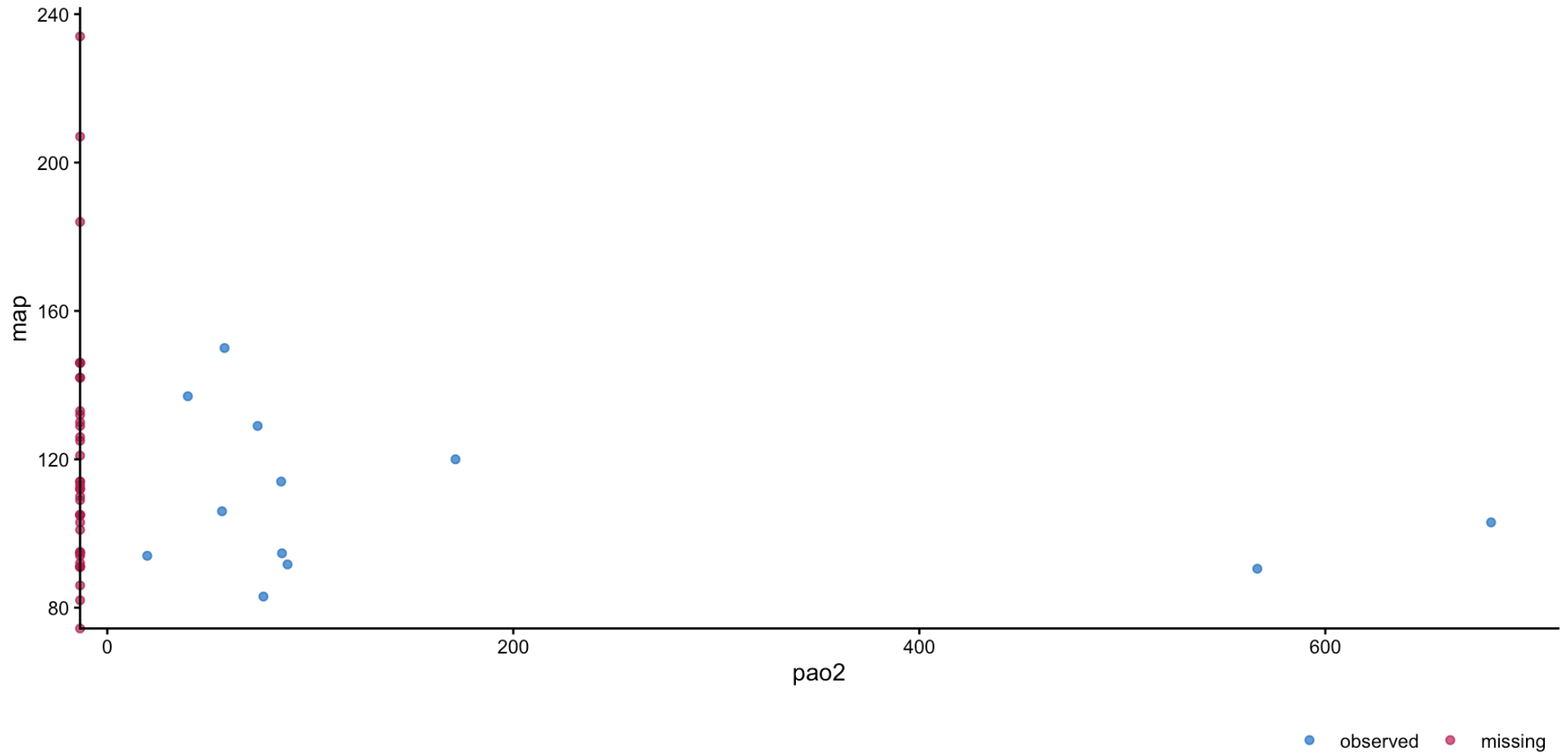
Mortality by blood pressure



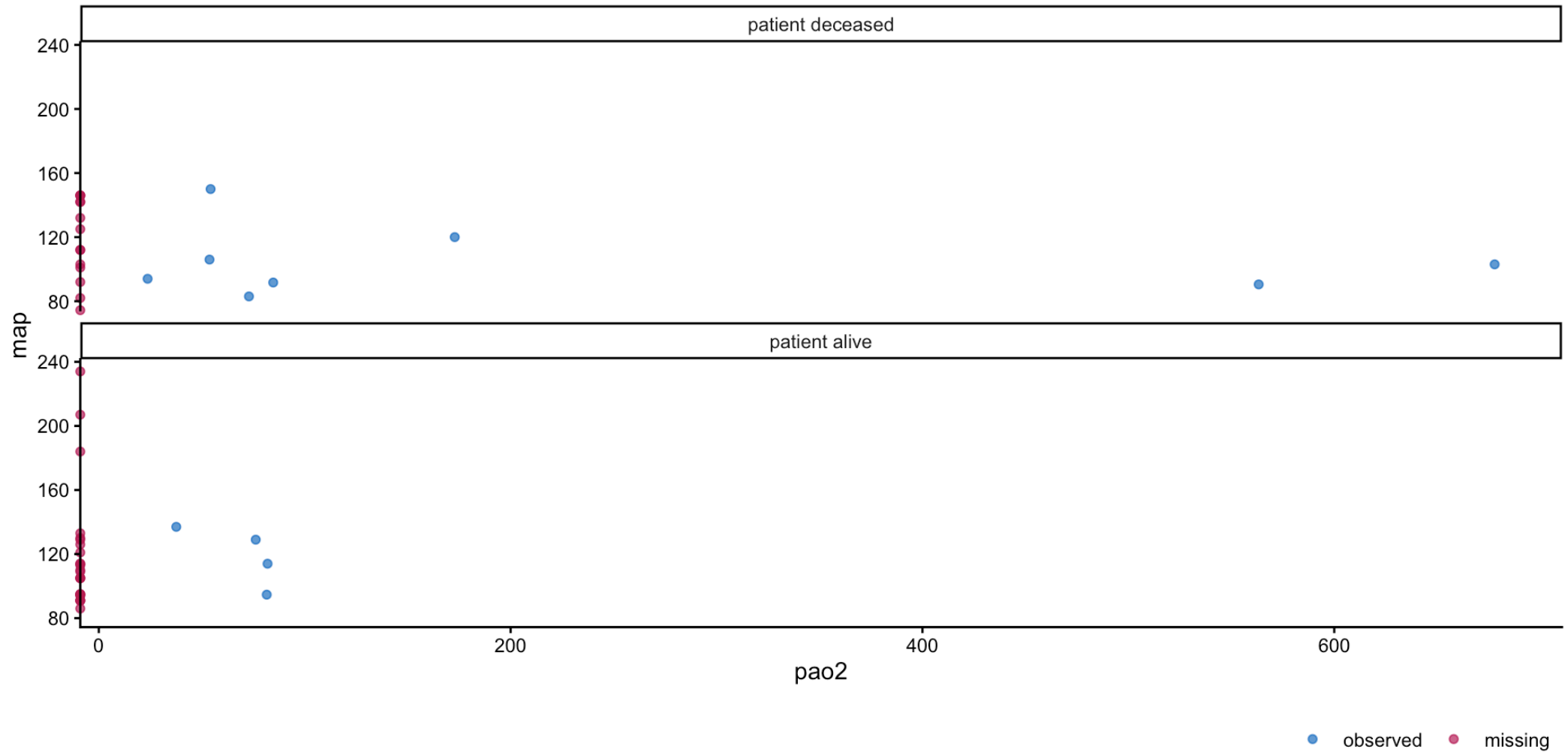
Mortality by oxygenation



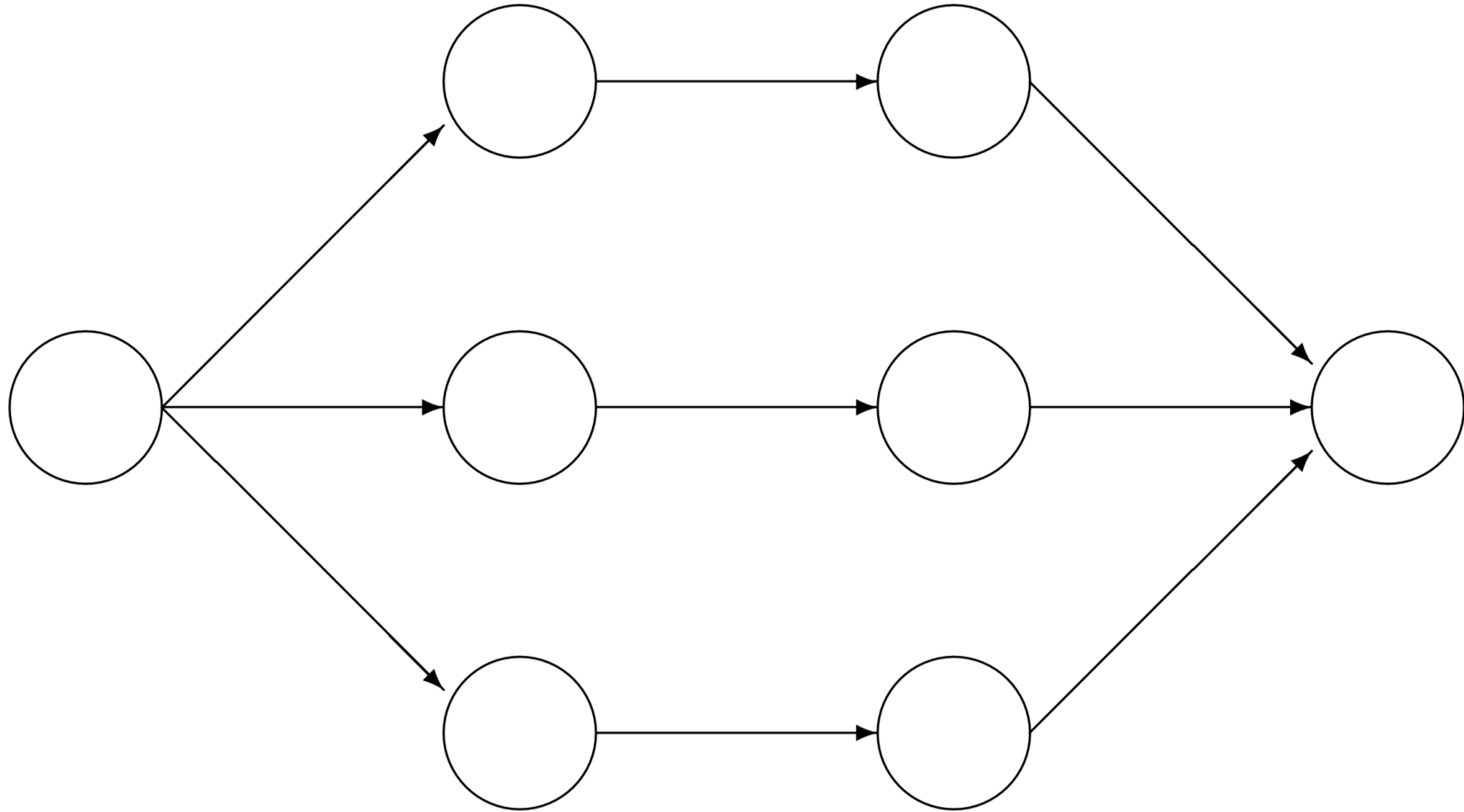
Scatter plot



Faceted scatter plot



Imputation workflow



Incomplete data

Imputed data

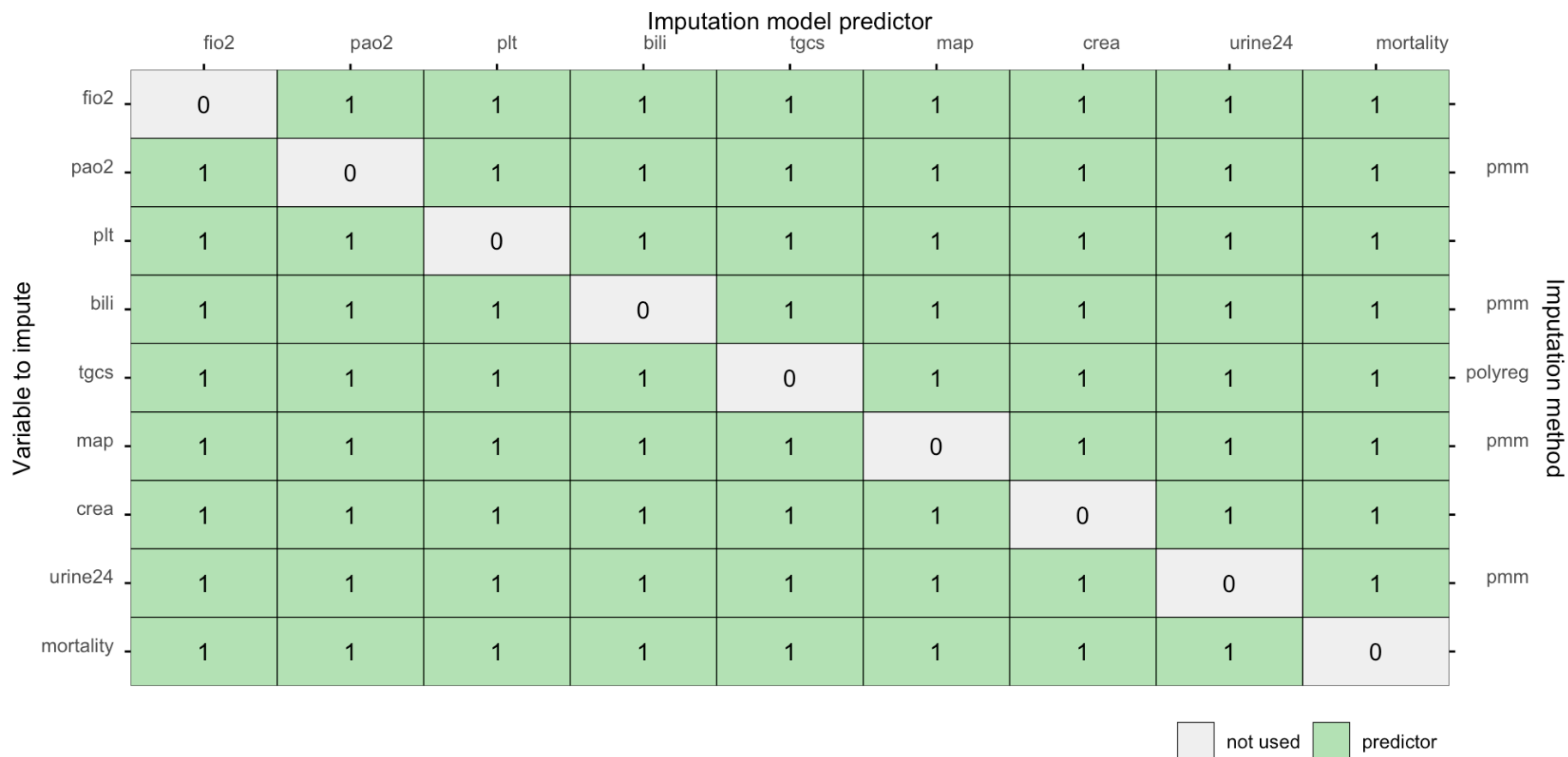
Analysis results

Pooled result



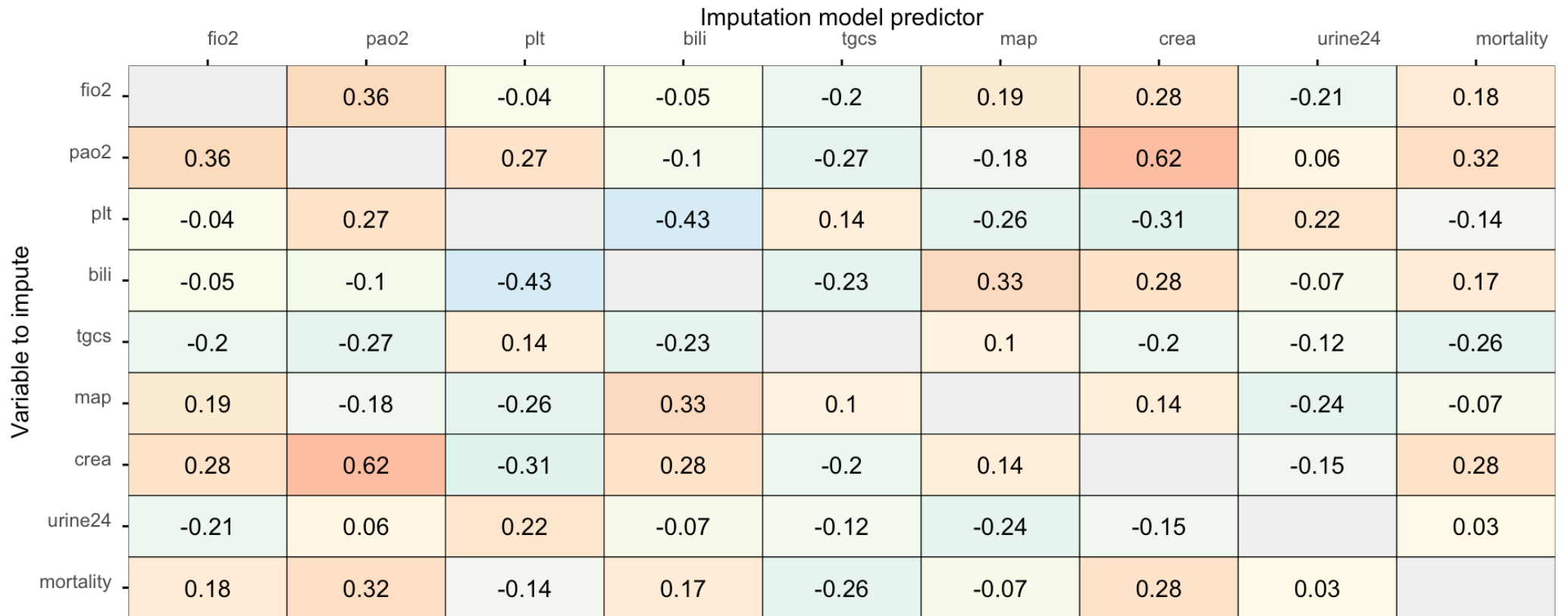
Imputation models


```
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)
```



Correlation* 
-1.0 -0.5 0.0 0.5 1.0

*pairwise complete observations



Imputation models

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

		Imputation model predictor										
		fio2	pao2	plt	bili	tgcs	map	crea	urine24	mortality		
fio2		0	0	0	0	0	0	0	0	0		
pao2		1	0	1	0	0	0	1	0	1	pmm	
plt		0	0	0	0	0	0	0	0	0		
bili		0	0	1	0	0	1	0	1	0	pmm	
tgcs		0	0	0	0	0	0	0	0	1	polyreg	
map		0	0	0	1	0	0	1	0	0	pmm	
crea		0	0	0	0	0	0	0	0	0		
urine24		0	1	0	0	0	0	1	0	0	pmm	
mortality		0	0	0	0	0	0	0	0	0		

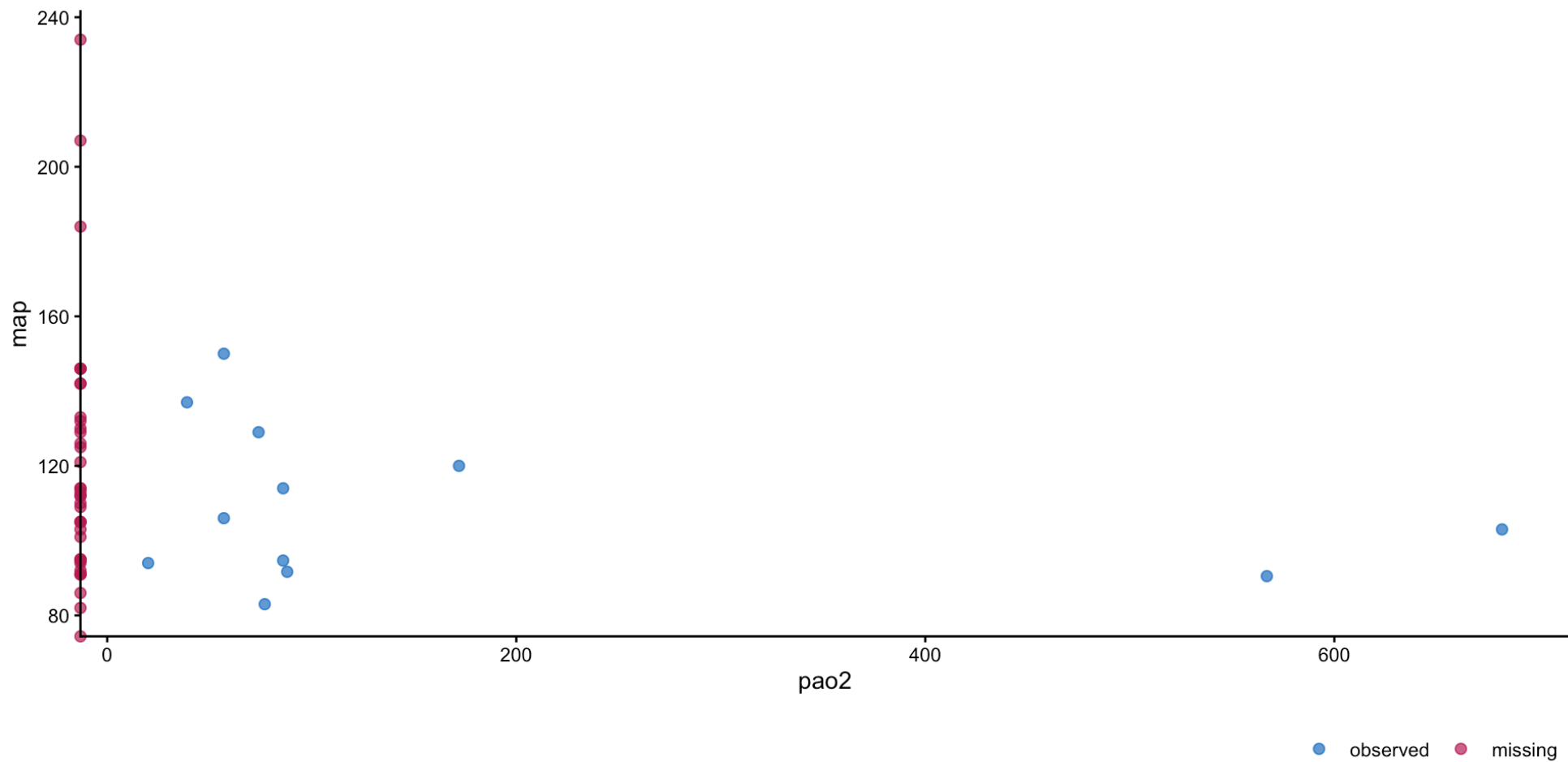
Variable to impute
Imputation method

not used
predictor



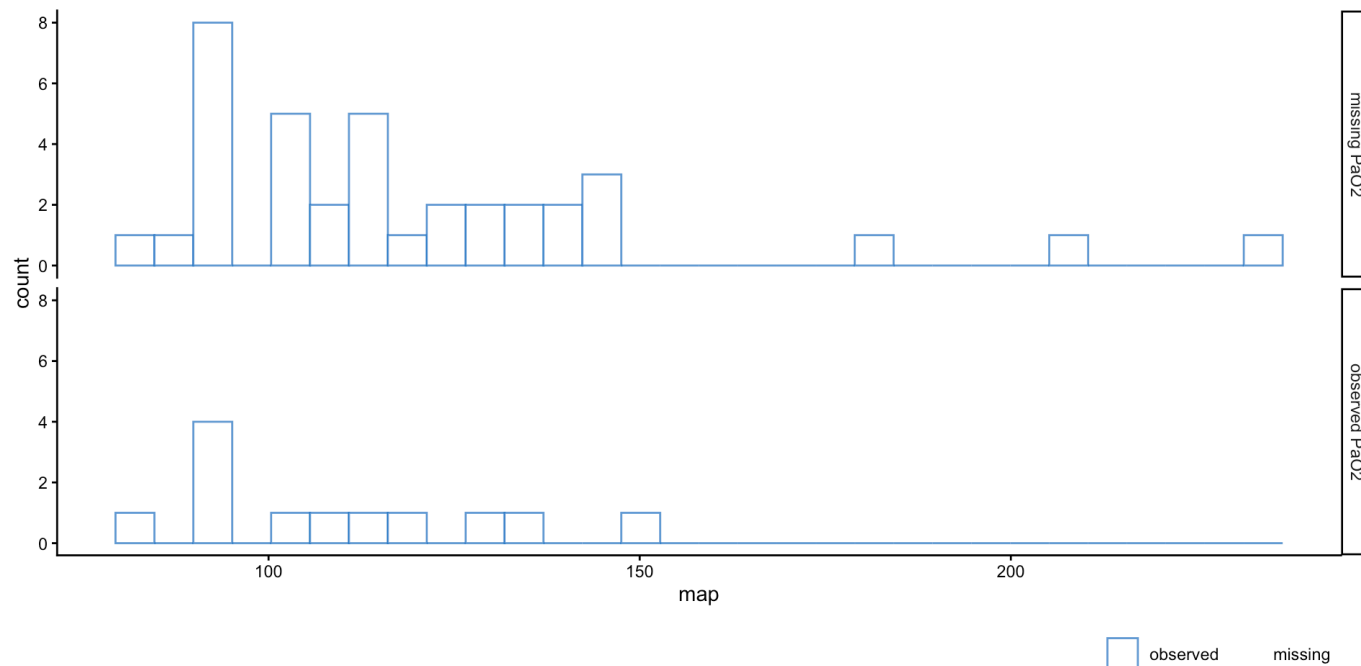
Scatter plot

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



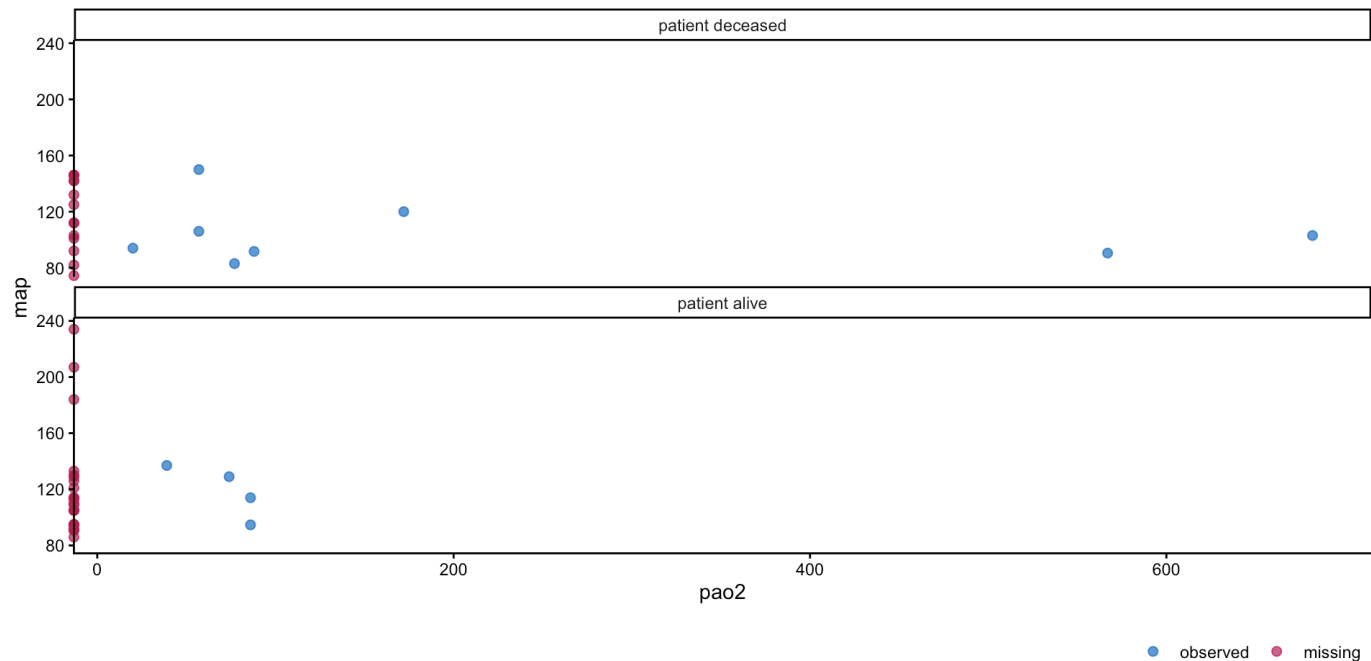
Faceted distribution

```
1 ggmlce(dat, aes(map)) +  
2   geom_histogram(fill = "white") +  
3   facet_grid(factor(  
4     is.na(pao2),  
5     levels = c(TRUE, FALSE),  
6     labels = c("missing PaO2", "observed PaO2")  
7     ) ~ .)
```



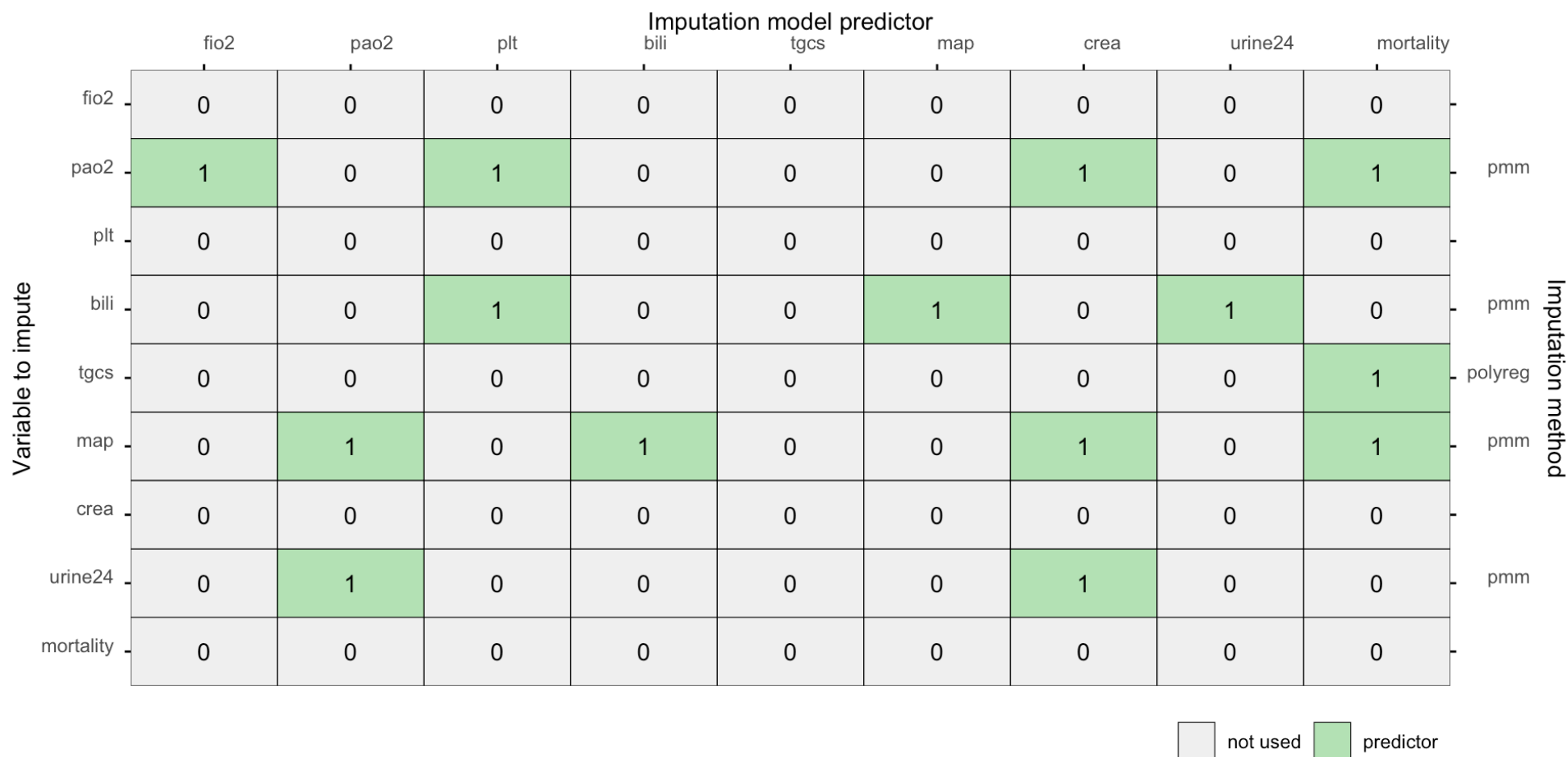
Faceted scatter plot

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



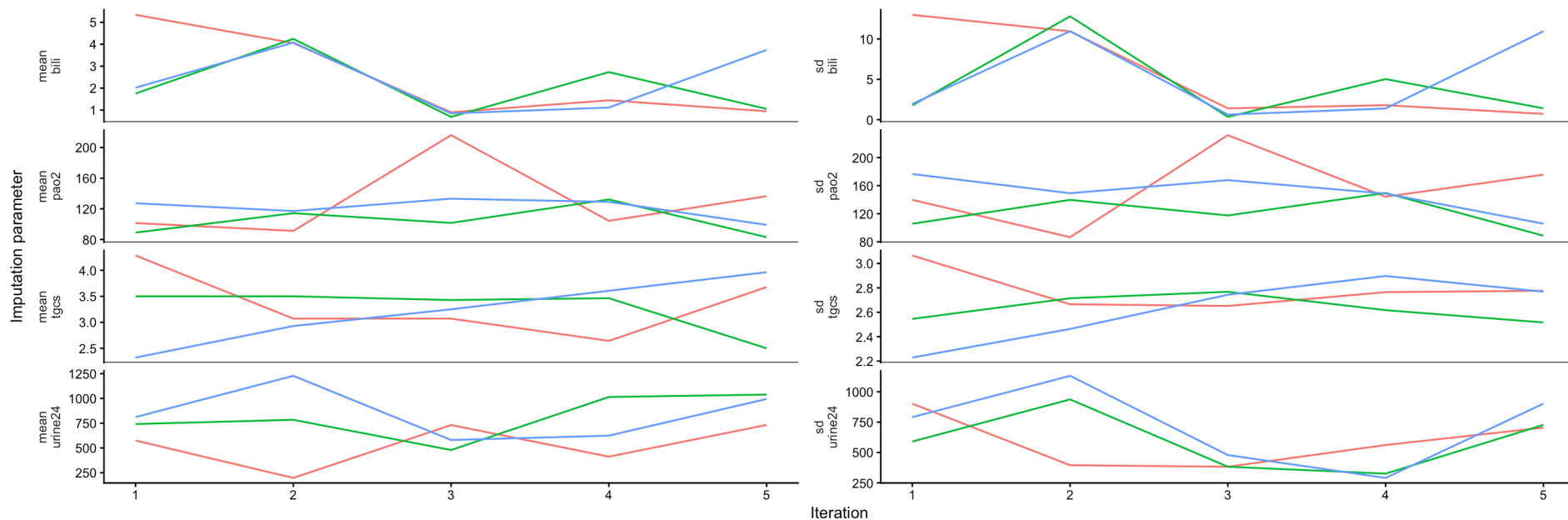
Adjust imputation models

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



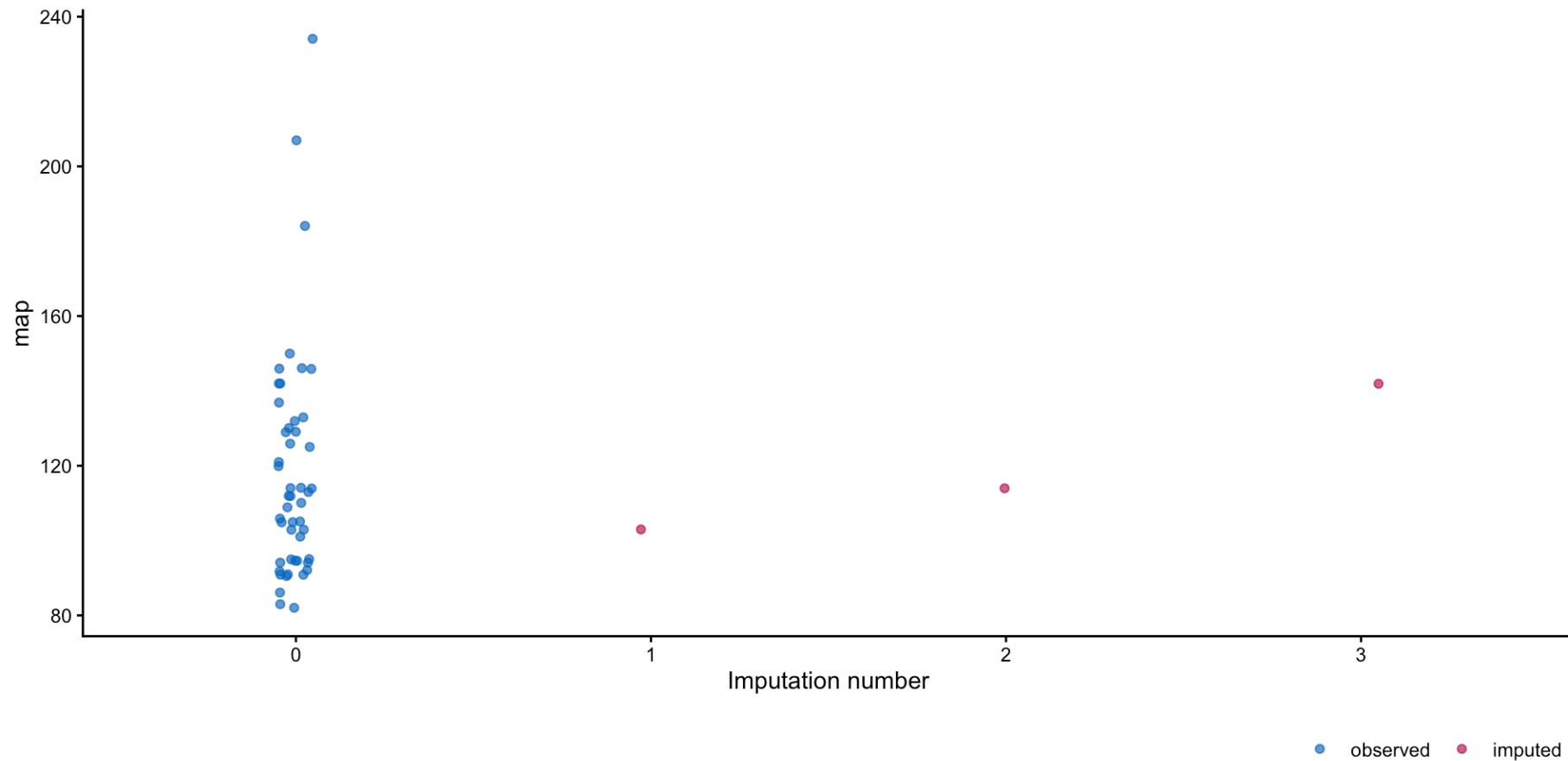
Impute

```
1 imp <- mice(  
2   dat,  
3   pred = pred,  
4   method = meth,  
5   m = 3,  
6   seed = 11,  
7   print = FALSE)  
8 plot_trace(imp, legend = FALSE)
```



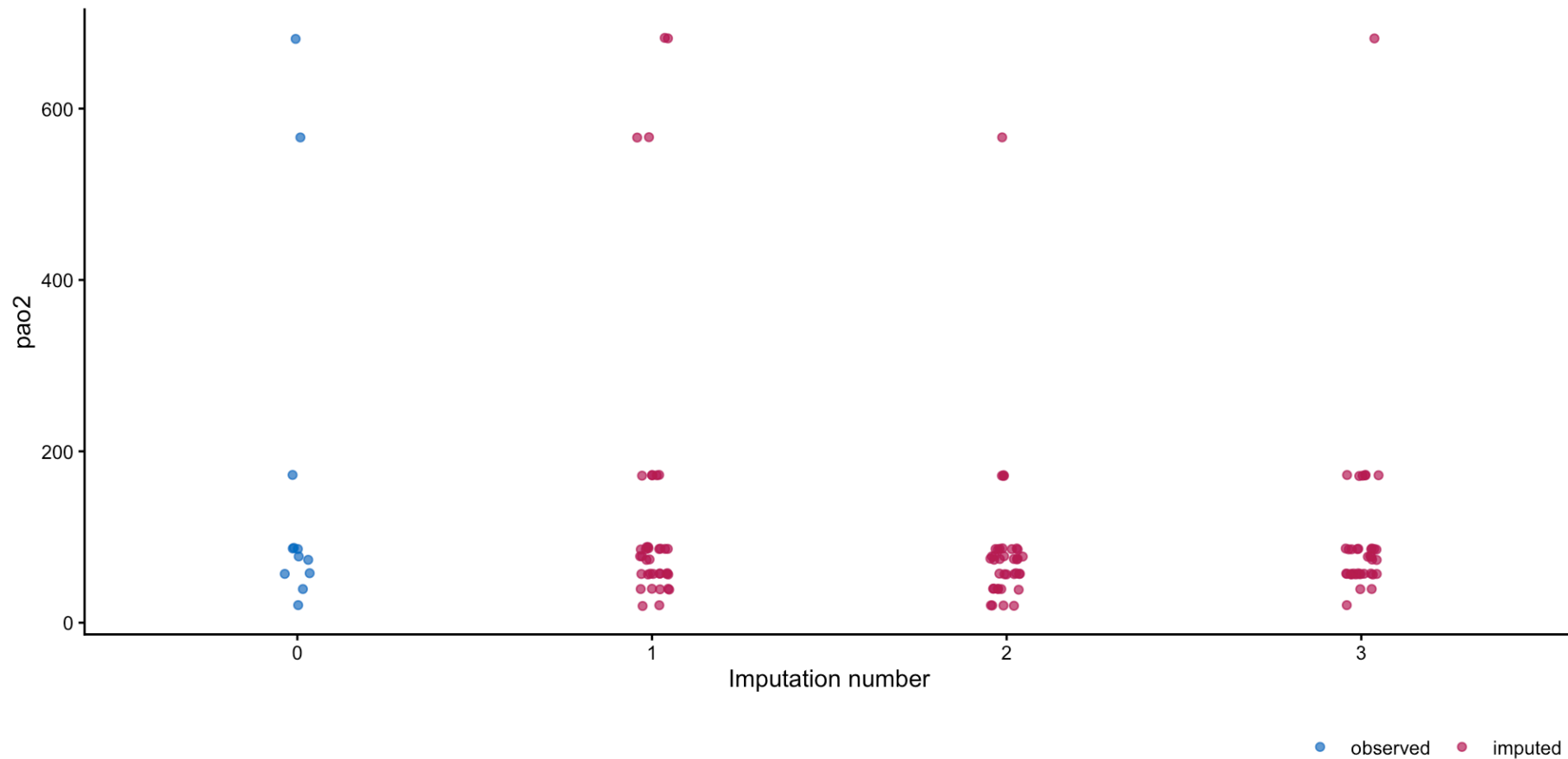
Stripplot of blood pressure

```
1 ggmise(imp, aes(x = .imp, y = map)) +  
2   geom_jitter(width = 0.05) +  
3   labs(x = "Imputation number")
```

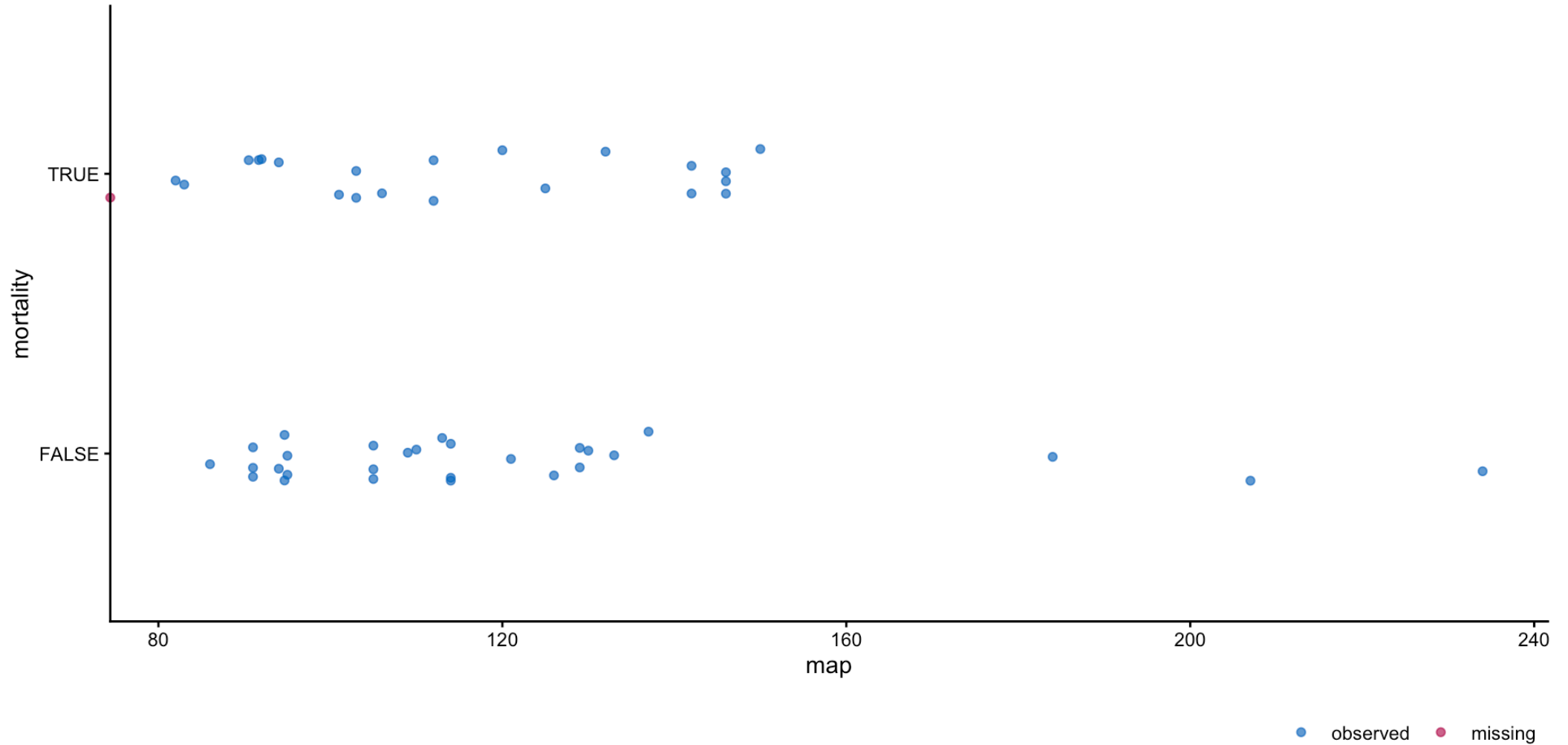


Stripplot of oxygenation

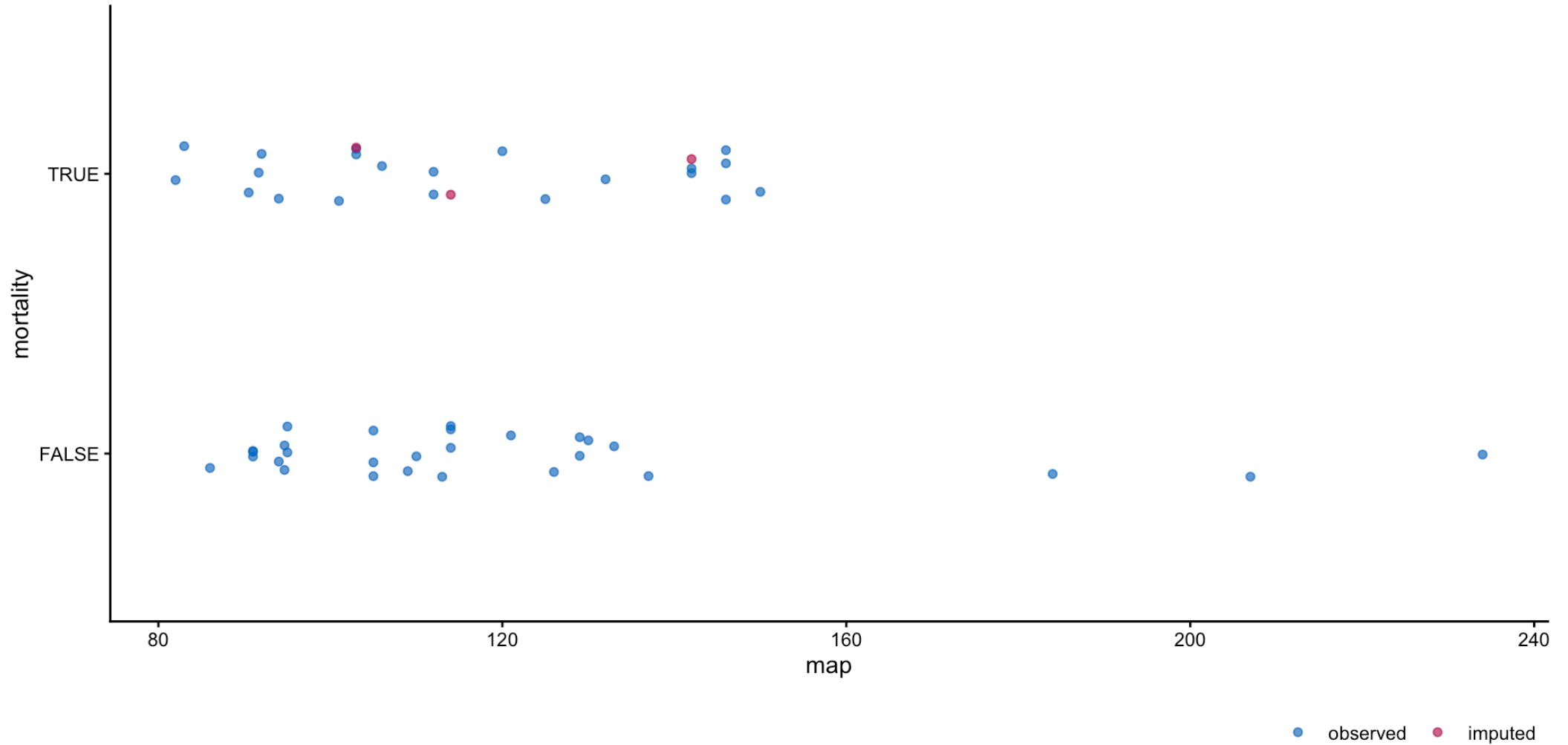
```
1 ggmise(imp, aes(x = .imp, y = pao2)) +  
2   geom_jitter(width = 0.05) +  
3   labs(x = "Imputation number")
```



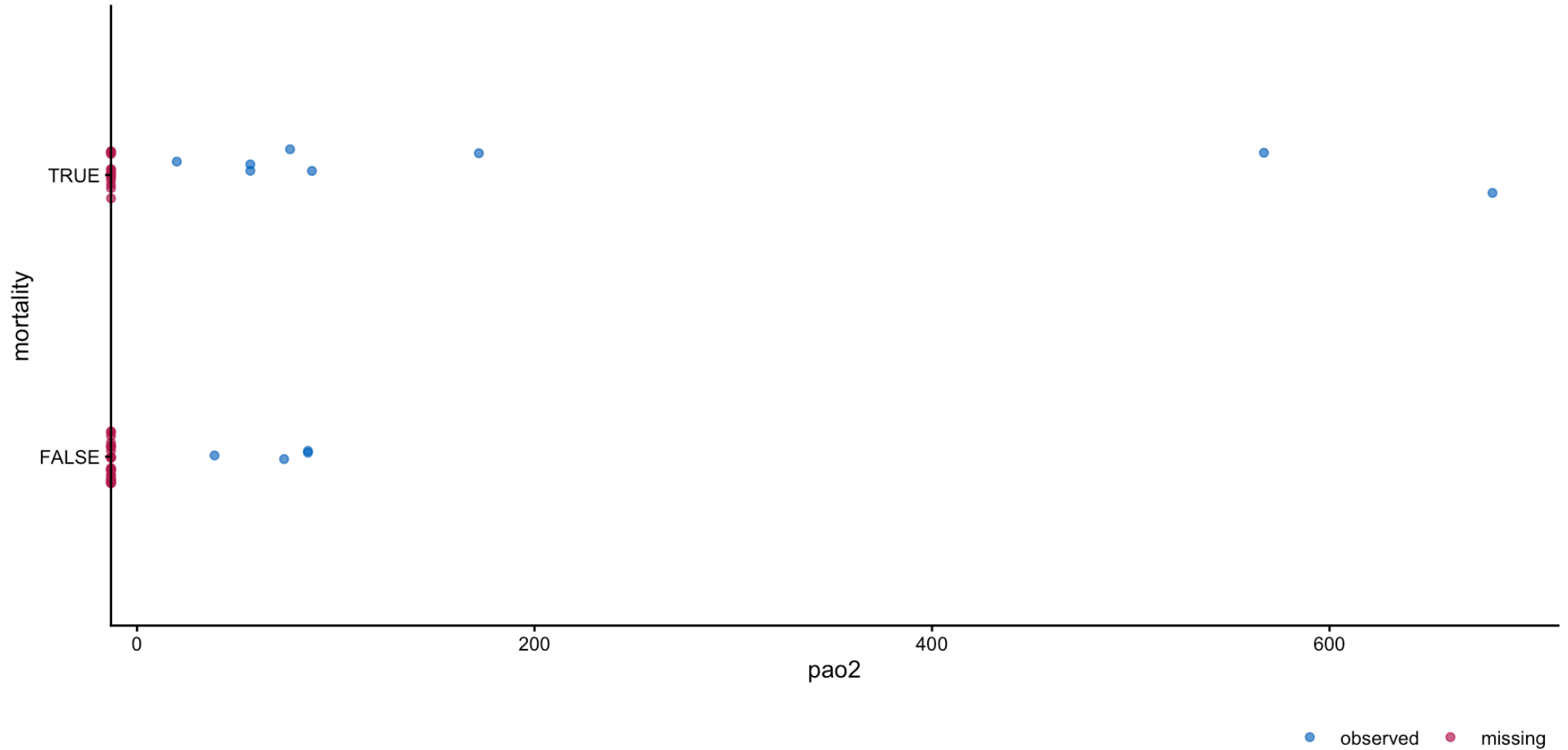
Mortality by blood pressure



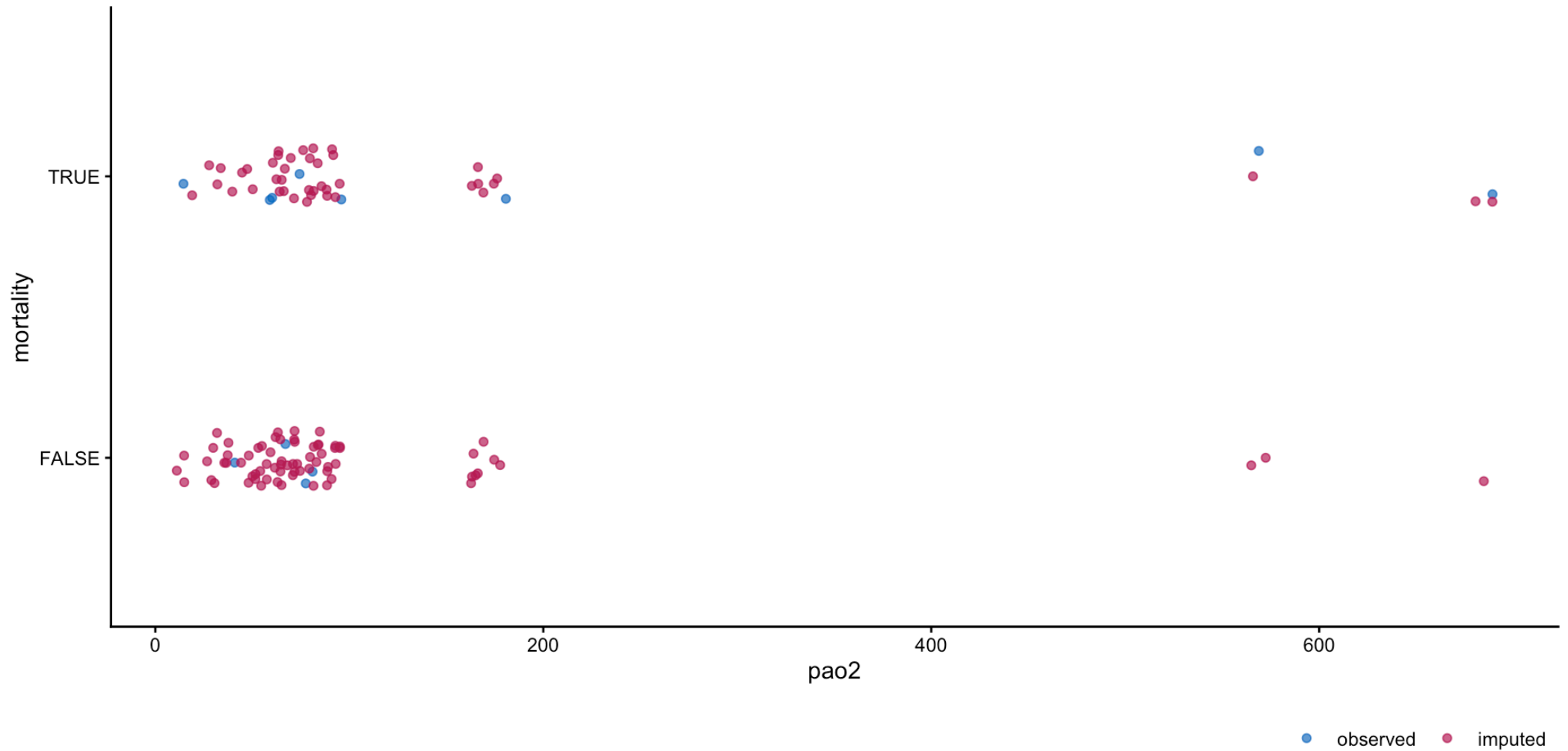
Mortality by blood pressure



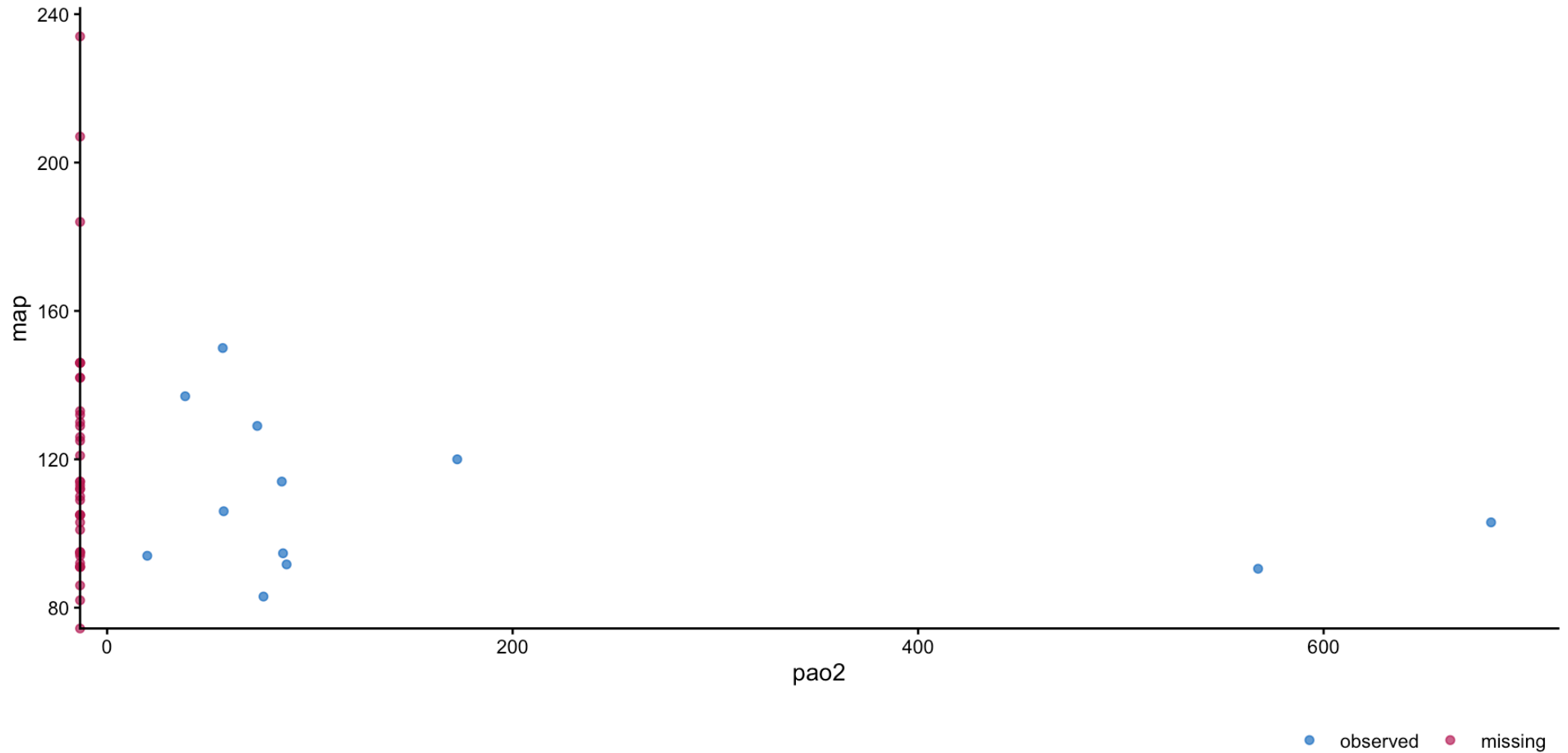
Mortality by oxygenation



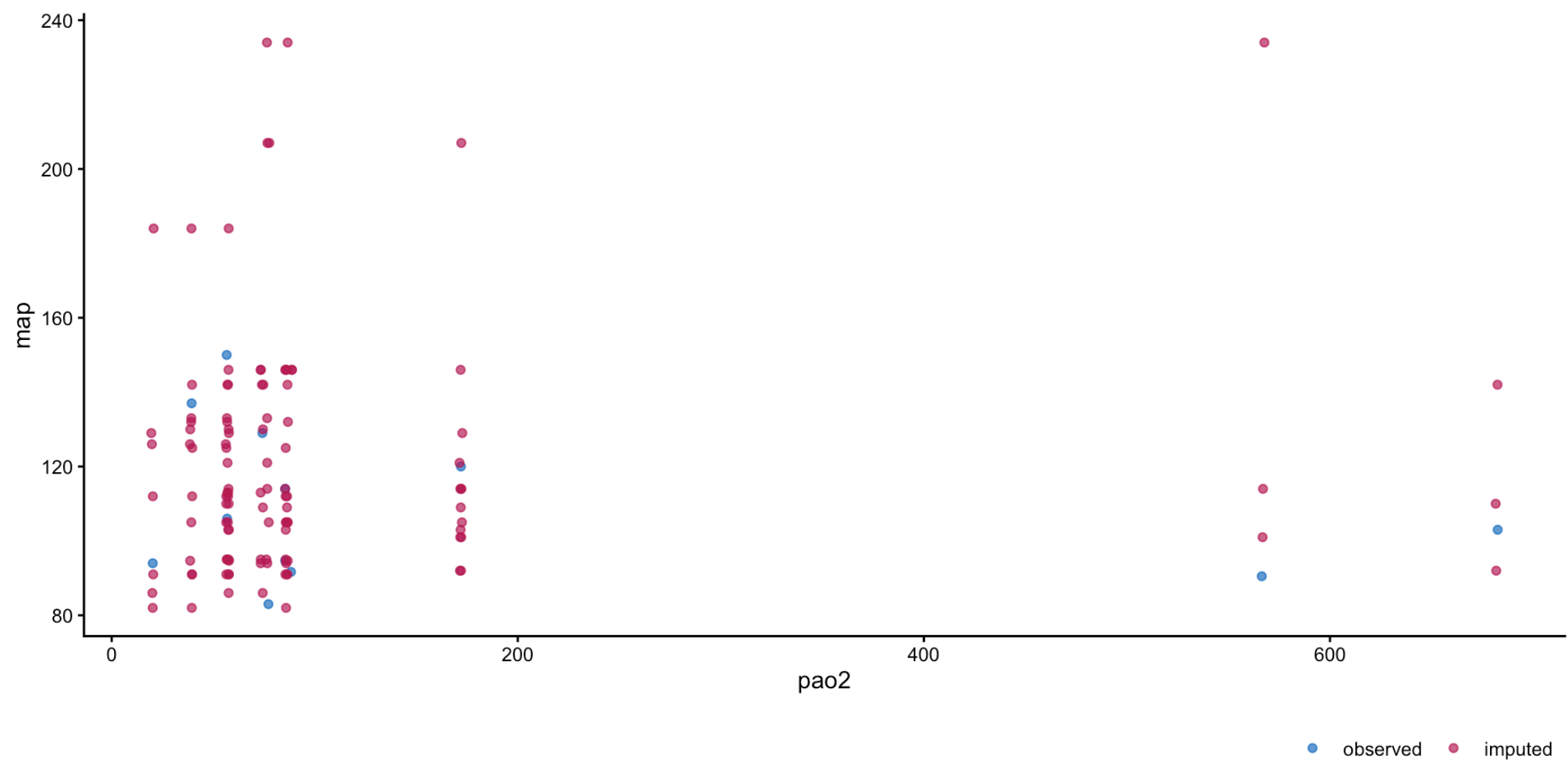
Mortality by oxygenation



Scatter plot

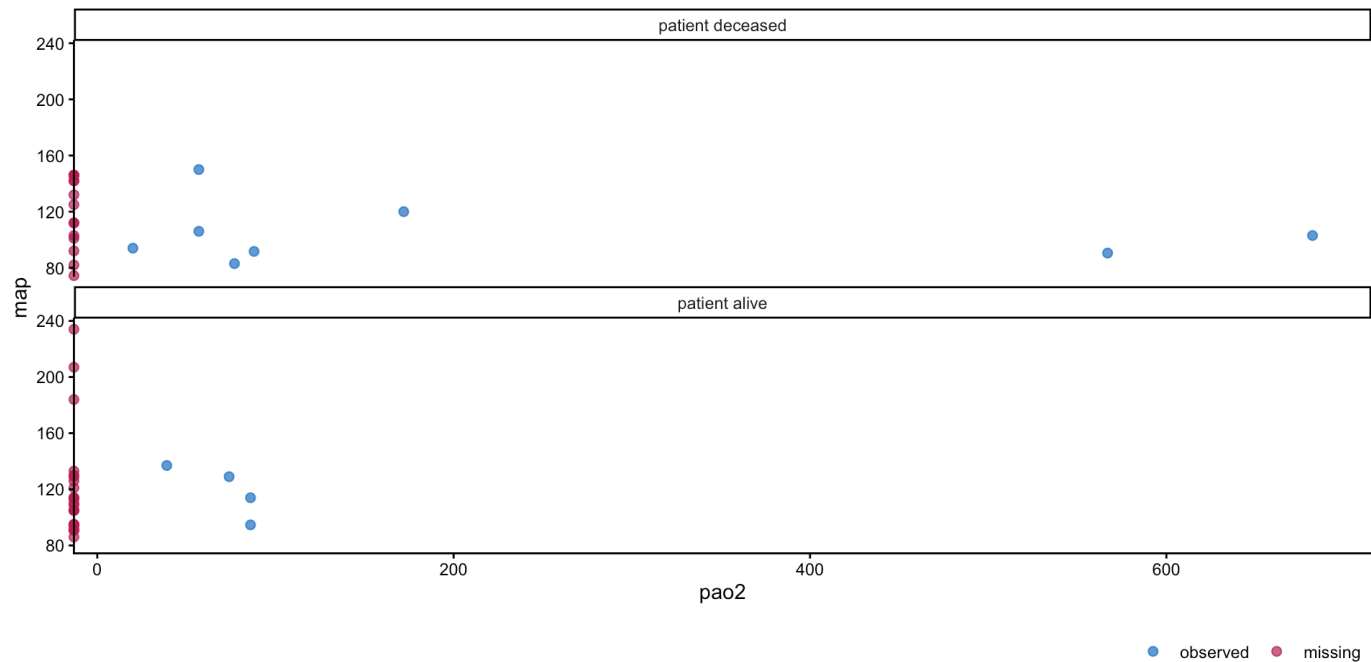


Scatter plot



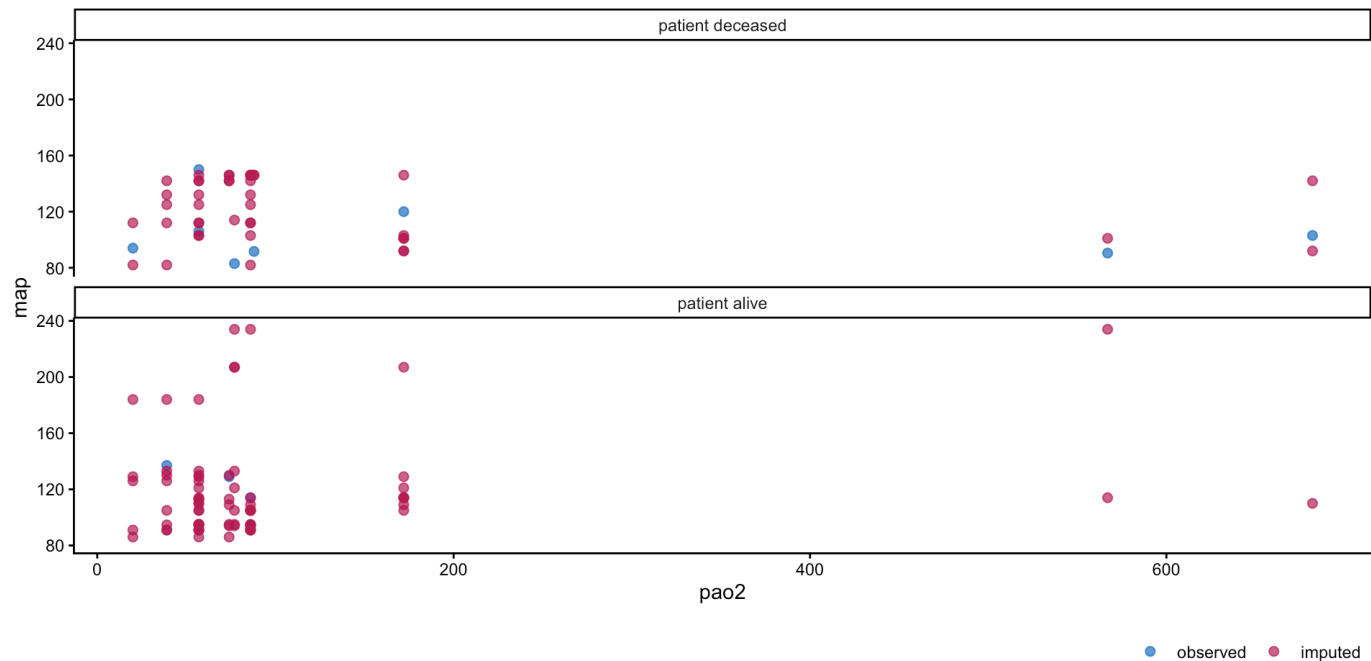
Faceted scatter plot

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



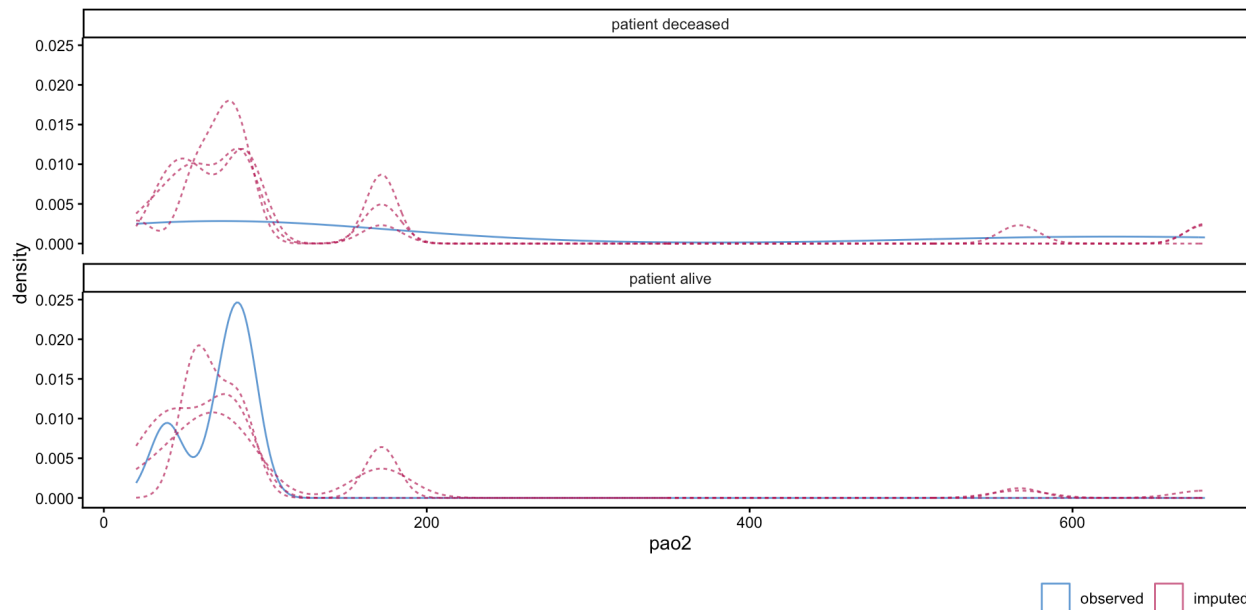
Faceted scatter plot

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



Faceted distribution

```
1 ggmlce(imp, aes(pao2, group = .imp, linetype = !(.imp == 0))) +  
2   geom_density() +  
3   facet_wrap(~factor(  
4     mortality,  
5     levels = c(TRUE, FALSE),  
6     labels = c("patient deceased", "patient alive")  
7   ), ncol = 1) +  
8   scale_linetype(guide = "none")
```



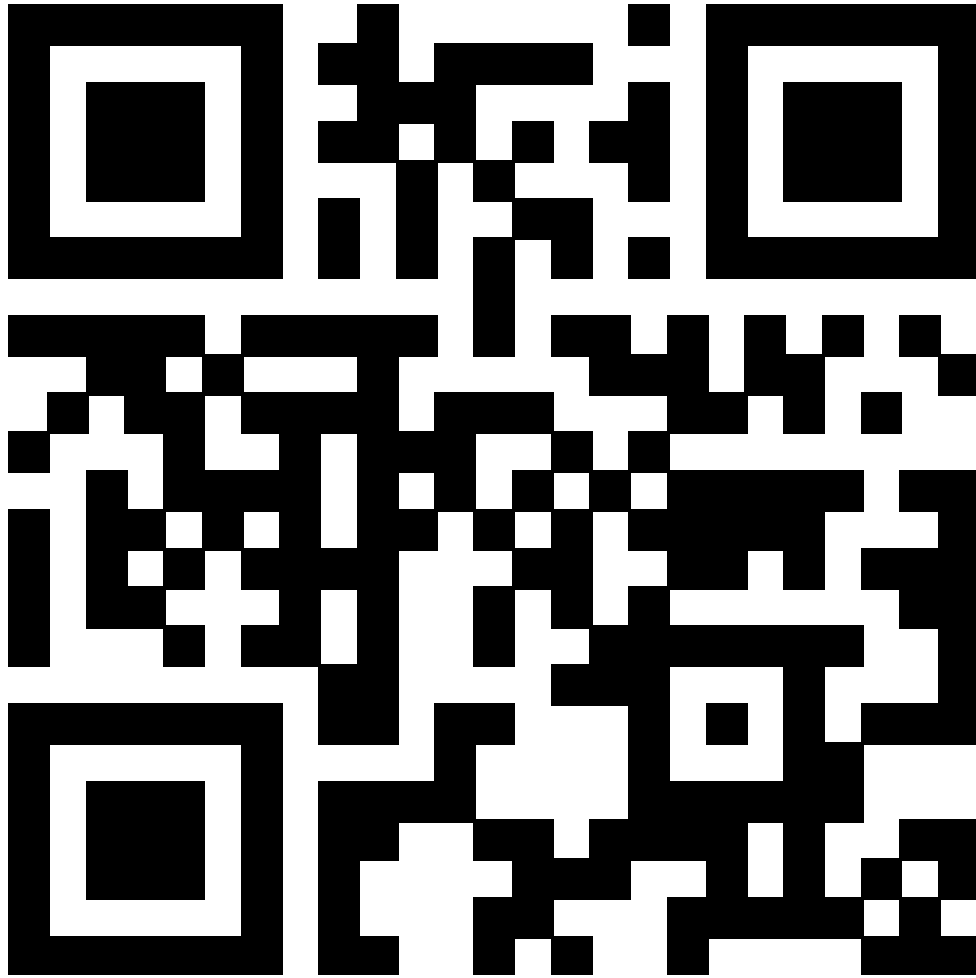
Take aways

Missing data are

- a pervasive problem
- visualizable & analyzable
- informative!



Thank you!



edu.nl/nbd3q

