Por Nuestra Salud Study: Sanity Checks on Curated Data

March 23, 2020

1 About

In this file, we document sanity checks on the Por Nuestra Salud (PNS) study curated data. Curated data are expected to (1) be consistent with the study design, and (2) be internally consistent, i.e., having no conflicting information between rows or columns in the dataset. To this end, we use the `testthat` package to organize *programmatic checks* on the curated data and the `ggplot2` package to create *visual checks* on the curated data.

```
use.samp.size <- 20
pns.quit.dates <- read.csv(file.path(path.pns.input_data, "pns_quit_dates.csv"), header = TRUE)
use.df.ids <- SampleAndRename(df = pns.quit.dates, use.seed = 754369, samp.size = use.samp.size)
```

Visual checks are displayed for the same sample of 20 randomly chosen PNS study participants.

2 Smoking Outcome Curated Datasets

```
df.smoking.01 <- read.csv(file.path(path.pns.output_data, "pns.smoking.01.csv"), header = TRUE)
df.smoking.02 <- read.csv(file.path(path.pns.output_data, "pns.smoking.02.csv"), header = TRUE)
```

2.1 Programmatic Checks

```
df <- df.smoking.01
test_file(file.path(path.pns.code, "pns-test-file-01.R"))
```

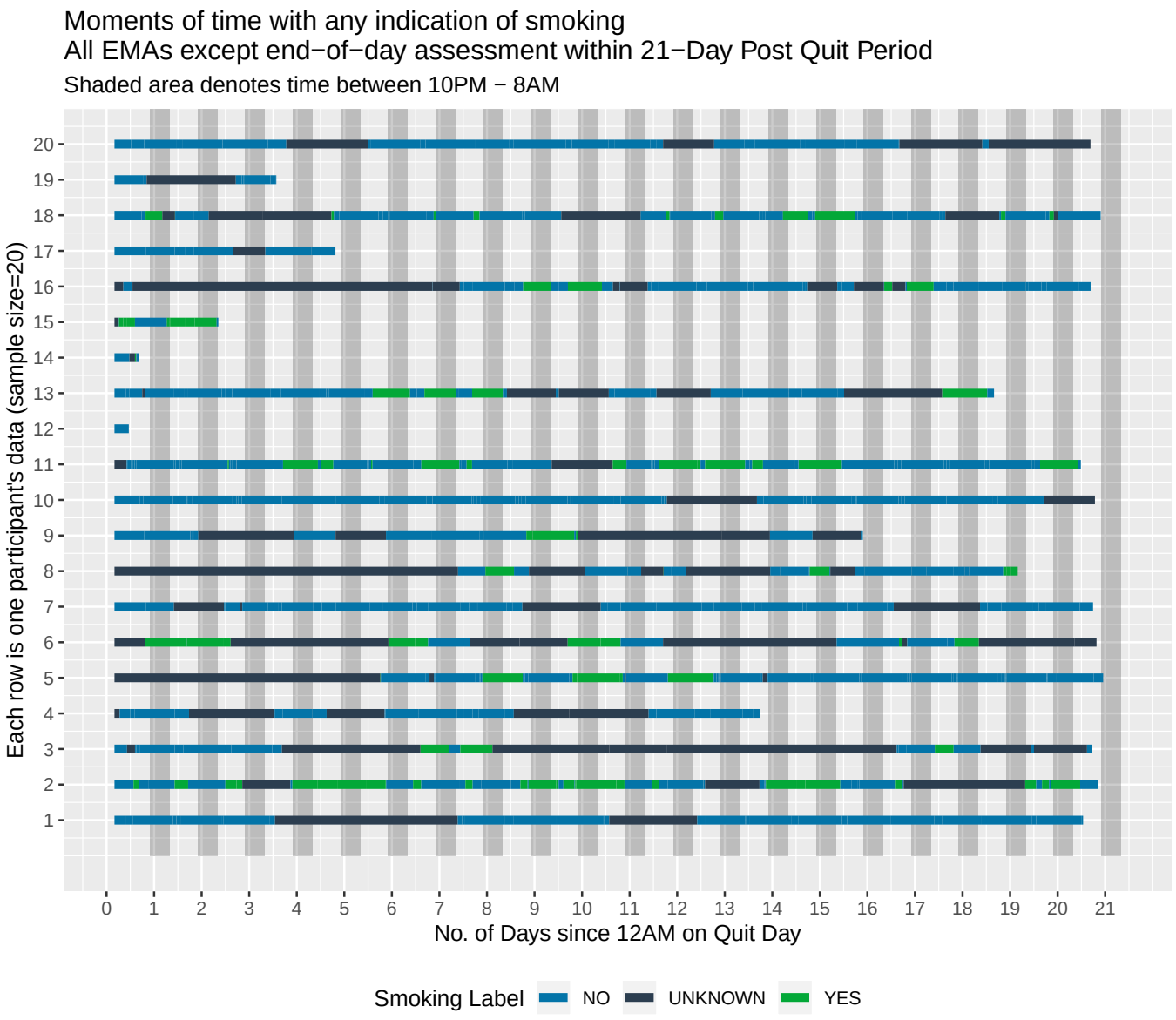
```
## v |  OK F W S | Context
## / |   0       | Construction of smoking intervals: internal consistency of curated data| |   3       | Co
##
## == Results =====================================================================================================
## Duration: 0.3 s
##
## OK:       4
## Failed:   0
## Warnings: 0
## Skipped:  0
```

```
df <- df.smoking.02
test_file(file.path(path.pns.code, "pns-test-file-01.R"))
```

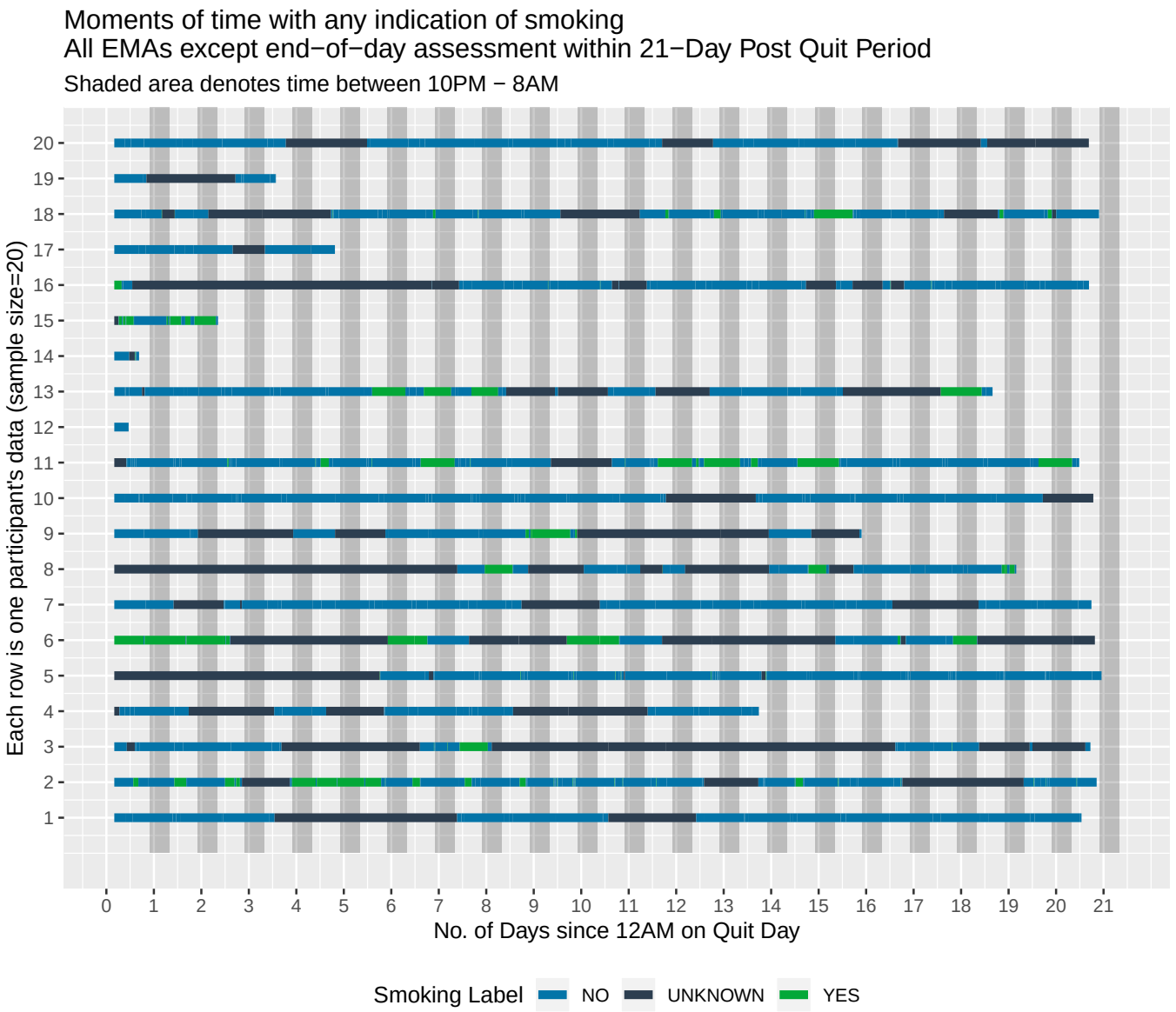
```
## v |  OK F W S | Context
## / |   0       | Construction of smoking intervals: internal consistency of curated data/ |   4       | Co
##
## == Results =====================================================================================================
## Duration: 0.2 s
##
## OK:       4
## Failed:   0
## Warnings: 0
## Skipped:  0
```

2.2 Visual Checks

```
gg.smoking.01 <- PlotSmokingOutcome(df.smoking = df.smoking.01, df.ids = use.df.ids)
gg.smoking.01
```



```
gg.smoking.02 <- PlotSmokingOutcome(df.smoking = df.smoking.02, df.ids = use.df.ids)
gg.smoking.02
```



3 Post-Quit Random Curated Datasets

```
df.post.quit.random.01 <- read.csv(file.path(path.pns.output_data, "pns.postquitrandom.01.csv"), header = TR
```

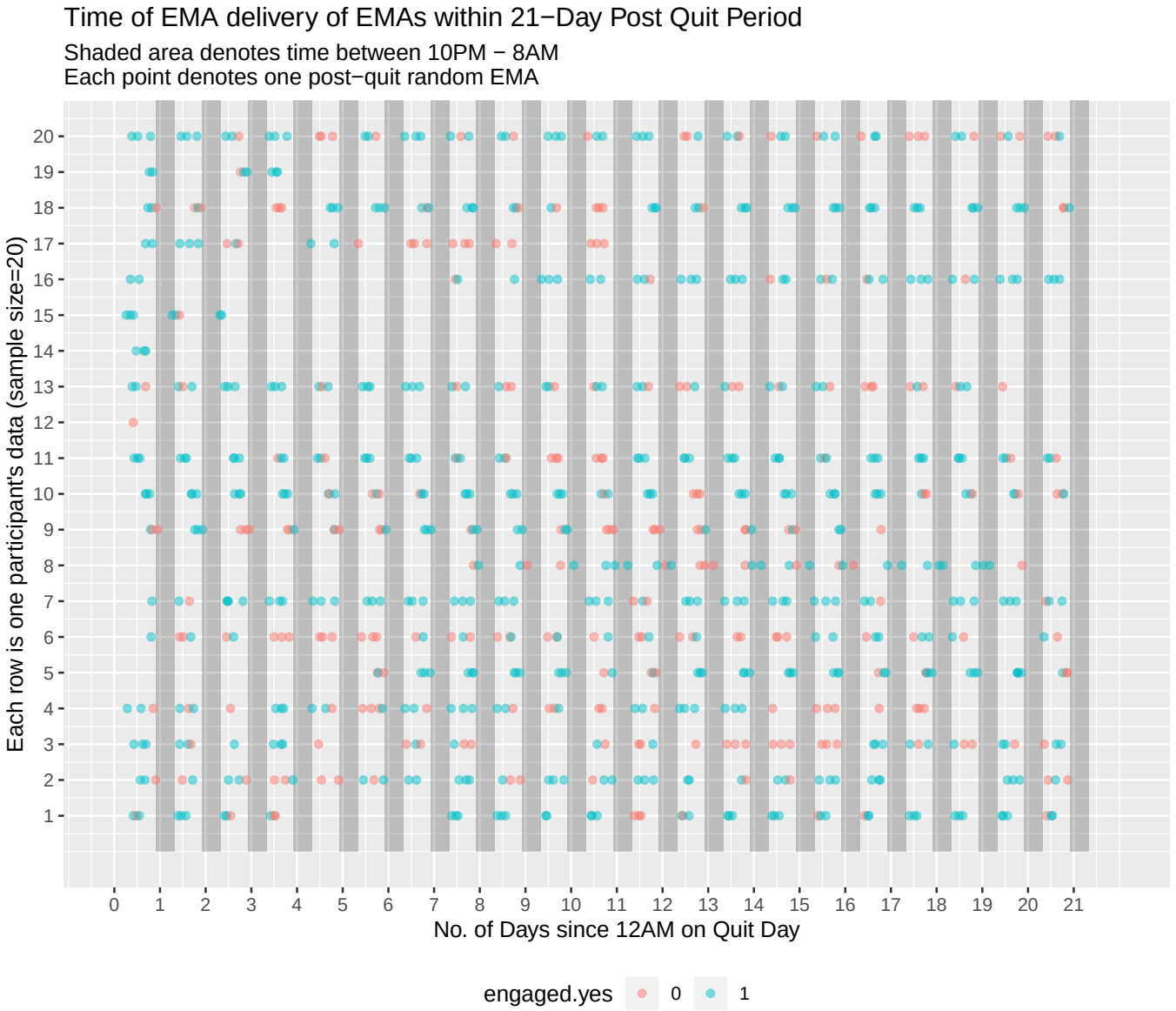
3.1 Programmatic Checks

```
df <- df.post.quit.random.01
test_file(file.path(path.pns.code, "pns-test-file-02.R"))
```

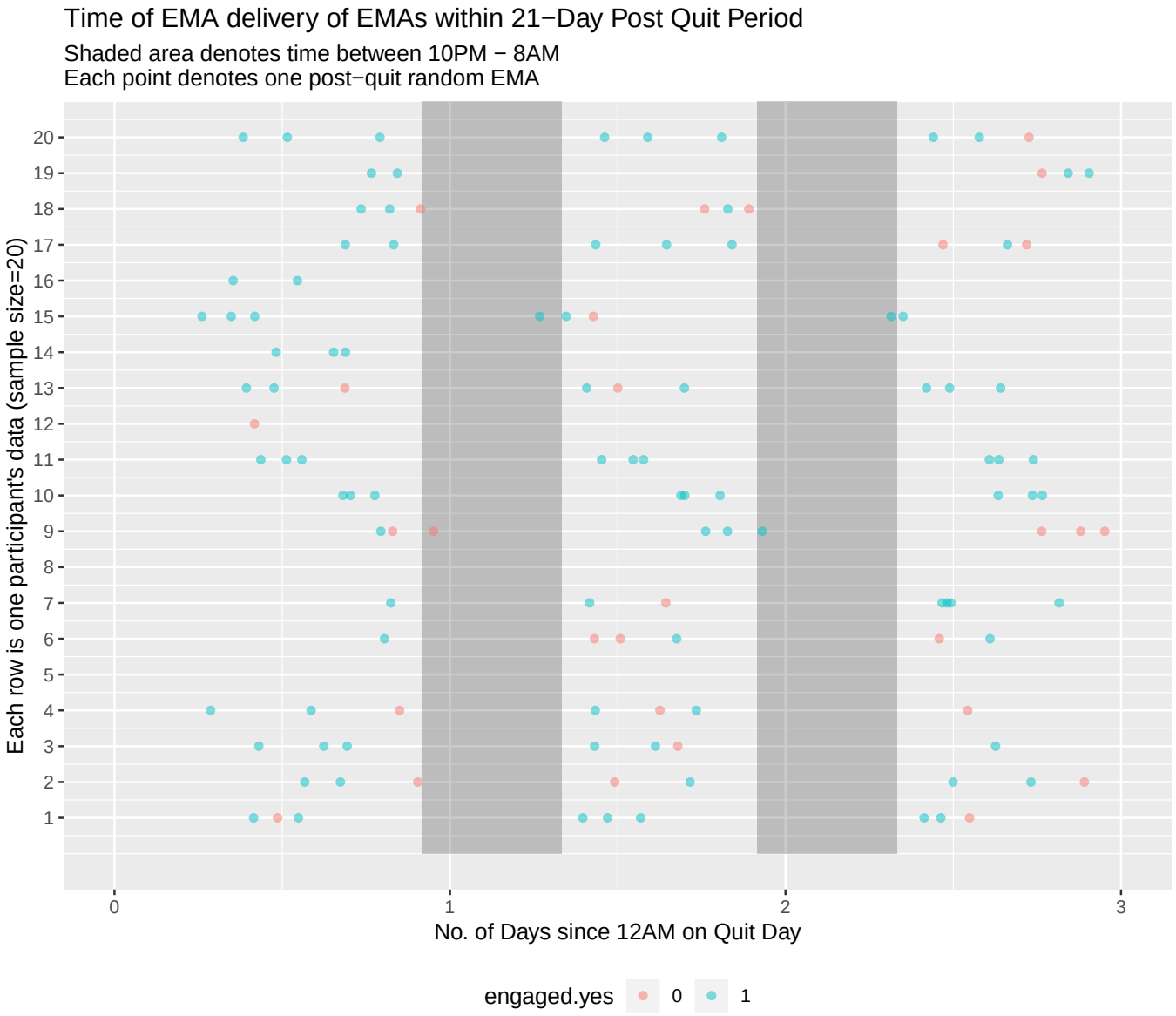
```
## v |  OK F W S | Context
## / |   0       | Construction of Post-Quit Random EMA Datasetsv |   1       | Construction of Post-Quit Ra
##
## == Results =====================================================================================================
## OK:       1
## Failed:   0
## Warnings: 0
## Skipped:  0
##
## You are a coding rockstar!
```

3.2 Visual Checks

```
gg.pq.random <- PlotPostQuitEMATime(df.post.quit = df.post.quit.random.01,
                                    df.ids = use.df.ids,
                                    plot.days = 22,
                                    ema.type="random")

gg.pq.random
```

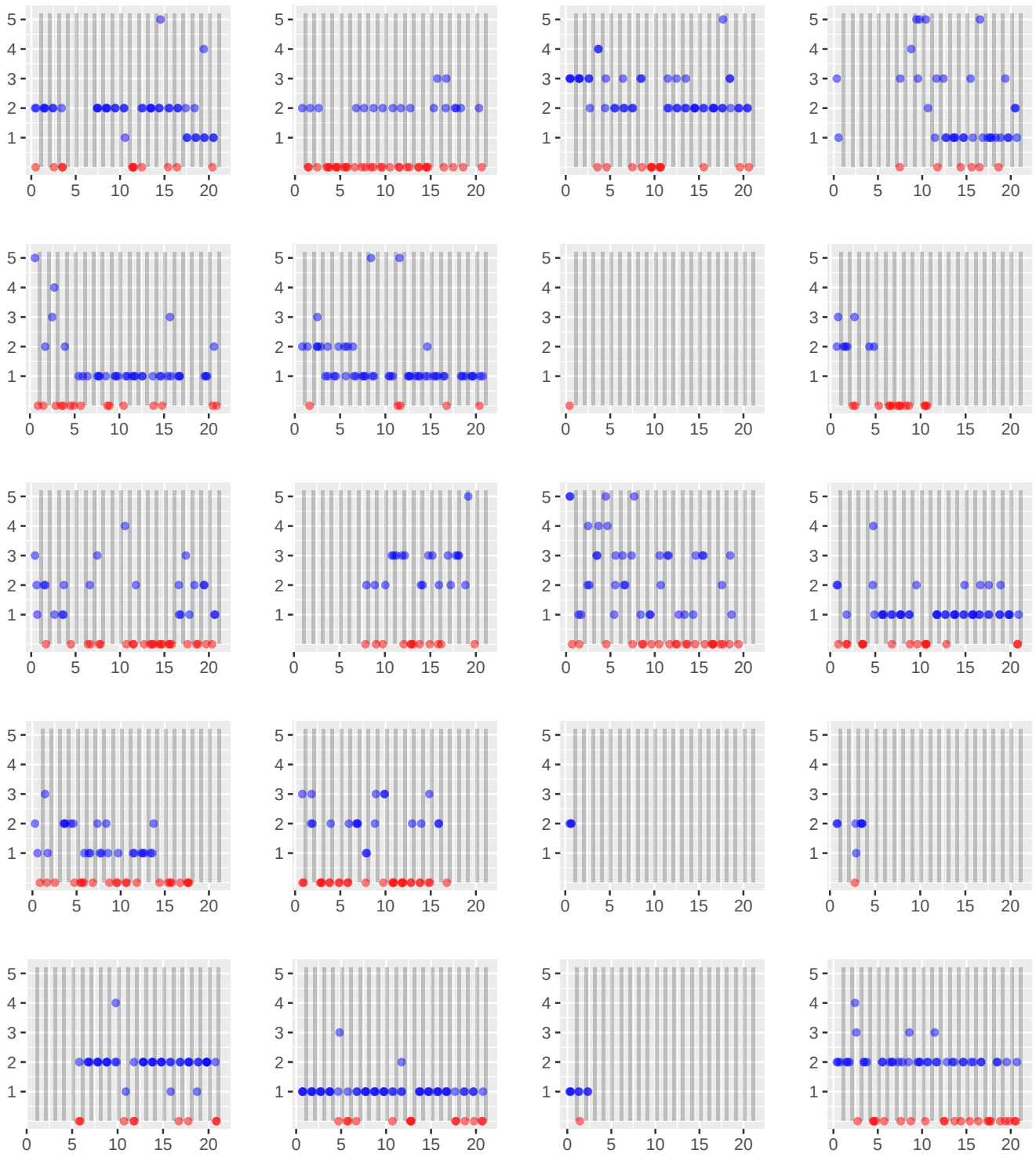


```
gg.pq.random.zoom <- PlotPostQuitEMATime(df.post.quit = df.post.quit.random.01,
                                         df.ids = use.df.ids,
                                         plot.days = 3,
                                         ema.type="random")

gg.pq.random.zoom
```



```
all.vars <- c(paste("Affect",c(6,8,10),sep=""))
collect.plot.grid <- list()

for(i in 1:length(all.vars)){
  use.var.name <- all.vars[i]

  collect.plots <- PlotPostQuitNumericResponses(df.post.quit = df.post.quit.random.01,
                                                var.name = use.var.name,
                                                df.ids = use.df.ids)

  text.top <- paste(use.var.name,
                    "Time of EMA delivery versus response on a 5-point Likert scale",
                    "All Random EMAs within 21-Day Post Quit Period",
                    sep="\n")
  text.bottom <- paste("Shaded area denotes time between 10PM - 8AM",
                       "Each point denotes one random EMA (red dots: engaged.yes=0, blue dots: engaged.yes=1
                       sep="\n")

  plot.grid <- marrangeGrob(grobs = collect.plots,
                            ncol=4,
                            nrow = 5,
                            top = textGrob(text.top,gp=gpar(fontsize=11,font=3)),
                            bottom = textGrob(text.bottom,gp=gpar(fontsize=11,font=3))
                            )

  collect.plot.grid <- append(collect.plot.grid, list(plot.grid))
}
```

```
collect.plot.grid[[1]]
```

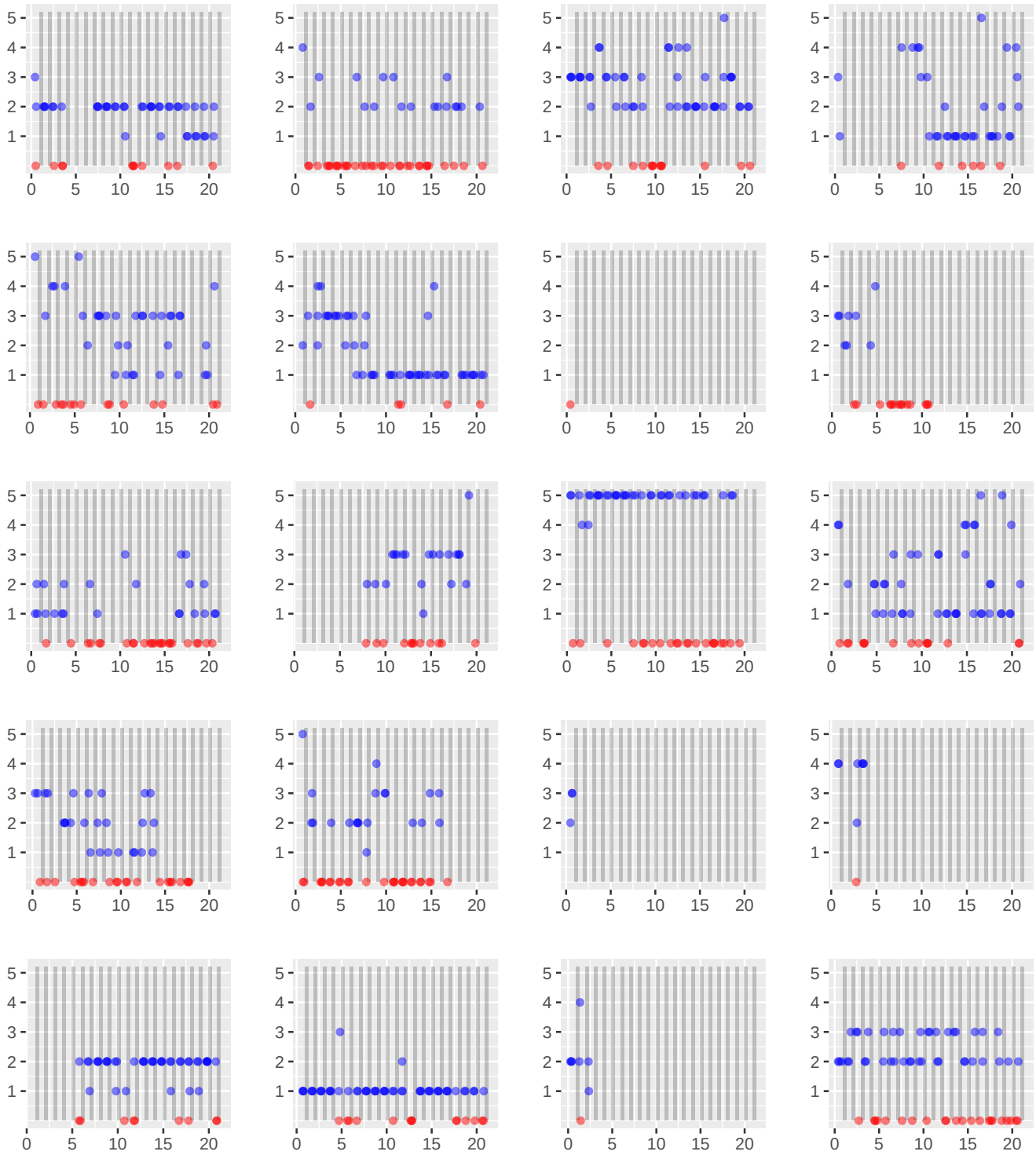


```
collect.plot.grid[[2]]
```

Affect8

Time of EMA delivery versus response on a 5−point Likert scale

All Random EMAs within 21−Day Post Quit Period



Shaded area denotes time between 10PM − 8AM

Each point denotes one random EMA (red dots: engaged.yes=0, blue dots: engaged.yes=1)

```
collect.plot.grid[[3]]
```