PreliminaryData

Melissa Lowe

1/29/2019

Import Data:

```
library(haven)
mort_adjud <- read_dta("~/Desktop/COPDGeneData/Mortality_Jan2019.dta",
    NULL)
```

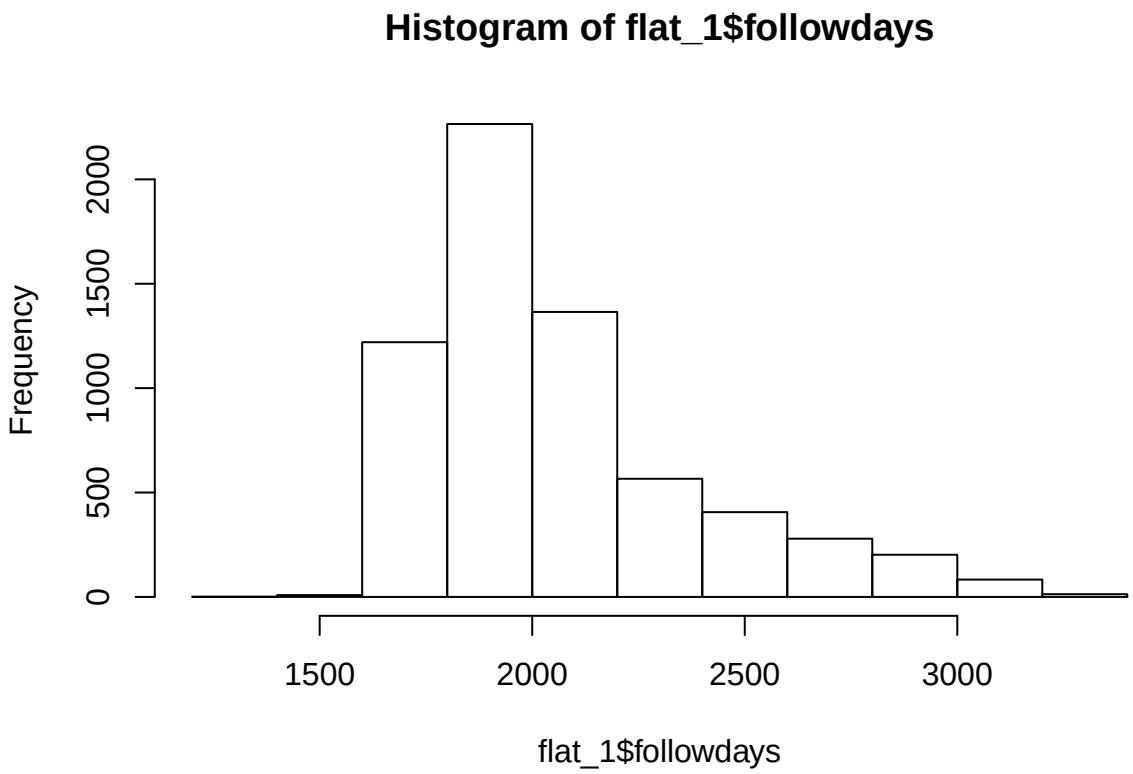
```
#copdgene_p1p2_all <- read_sas("~/Desktop/COPDGeneData/copdgene_p1p2_all_visit_30dec18.sas7bdat",NULL)
```

```
copdgene_p1p2_flat <- read_dta("~/Desktop/COPDGeneData/P1P2_Pheno_Flat_All_sids_Dec18.dta",
    NULL)
```

Create my own dataset of variables of interest:

We know we'll need sid, visit num, ccenter, visit date, gender, race, smoking status, age at baseline, visit type, exclude lungtrans, height, wegiht, distwalked, cigperdaysmoknow, all copdexac, lungproc_lungtransplant, copdafe, emphage, smokstartage, ats_packyears, yearssincequit, fev1pp_utah, fev1_fvc_utah,

```
#need to deal with the atomic labels which are actually storing the subject ids
library(sjlabelled)

#save the old sid values just in case
copdgene_p1p2_flat$sid2 <- copdgene_p1p2_flat$sid

#pull the labels as necessary
copdgene_p1p2_flat$sid <- get_labels(copdgene_p1p2_flat$sid)

#make them both into characters so they are the same.
mort_adjud$sid <- as.character(mort_adjud$sid)


library(tidyverse)


#subset so I don't have an insanely large data set


flat_1 <- copdgene_p1p2_flat %>% select(sid, gender,race,Visit_Date_P1,ccenter_P1, EverSmokedCig_P1, sm
BMI_P2,distwalked_P2,SmokStopAge_P2,ATS_PackYears_P2, YearsSinceQuit_P2, Severe_Exacerbations_P2, Exace

#possibly useful data set
mort_1 <- mort_adjud %>% select(sid, vital_status, mortality_survival_vetted, mortality_survival_vital_

#number of days since January 1, 1960 is the date.

flat_1$followdays <- flat_1$Visit_Date_P2 - flat_1$Visit_Date_P1

hist(flat_1$followdays)
```




```
mean(na.exclude(flat_1$followdays))
```

```
## [1] 2065.648
```

```
summary(na.exclude(flat_1$followdays))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1310    1826    1976    2066    2191    3378
```

```
flat_1$diff_spiro_ratio <- flat_1$FEV1_FVC_utah_P2 -  flat_1$FEV1_FVC_utah_P1

flat_1$diff_spiro_fev1 <- flat_1$FEV1_utah_P2 - flat_1$FEV1_utah_P1

flat_1$diff_emphys <- flat_1$pctEmph_Thirona_P2 - flat_1$pctEmph_Thirona_P1

flat_1$diff_gas_trap <- flat_1$pctGasTrap_Thirona_P2 - flat_1$pctGasTrap_Thirona_P1

flat_1$diff_Pi10 <-  flat_1$Pi10_Thirona_P1 - flat_1$Pi10_Thirona_P2

flat_1$diff_AWT <- flat_1$AWT_seg_Thirona_P1 - flat_1$AWT_seg_Thirona_P2


flag_items <- as.data.frame(cbind(flat_1$diff_spiro_ratio,flat_1$diff_spiro_fev1, flat_1$diff_emphys, fl

names(flag_items) <- c("diff_ratio","diff_fev1", "diff_emph", "diff_gastrap", "diff_pi10", "diff_awt")
```

Summary of Visits:

Summary Tables of Outcomes of Interest

```
outcome_table <- as.data.frame(t(sapply(flag_items, summaries)))
outcome_table
```

```
##               Mean   SD    N Minimum Maximum .05,.5, .95     NA     NA
## diff_ratio   -0.01 0.07 5717   -0.37    0.72      -0.120 -0.010  0.090
## diff_fev1    -0.21 0.30 5718   -1.96    2.80      -0.709 -0.199  0.218
## diff_emph     0.31 3.75 5093  -22.58   27.14      -5.394  0.028  7.023
## diff_gastrap  1.63 8.76 4138  -45.72   41.24     -11.389  0.925 17.165
## diff_pi10    -0.03 0.38 5093   -2.33    2.60      -0.634 -0.030  0.600
## diff_awt      0.00 0.12 5087   -0.63    1.24      -0.194 -0.003  0.189
```

Currently, they're using the 95 percentile to mark where they think a serious change in disease status would be for these markers.

This is obviously a fairly artificial marker.

```
flat_1$flag_spiro_ratio <- ifelse(flat_1$diff_spiro_ratio >= 0.09, 1, 0)

sum(na.exclude(flat_1$flag_spiro_ratio))
```

```
## [1] 252
```

```
flat_1$flag_fev1 <-  ifelse(flat_1$diff_spiro_fev1 >= 0.218, 1, 0)

sum(na.exclude(flat_1$flag_fev1))
```

```
## [1] 285
```

```
flat_1$flag_emphys <-  ifelse(flat_1$diff_emphys >= 7.023, 1, 0)

sum(na.exclude(flat_1$flag_emphys))
```

```
## [1] 255
```

```
flat_1$flag_gastrap <-  ifelse(flat_1$diff_gas_trap >= 17.165, 1, 0)

sum(na.exclude(flat_1$flag_gastrap))
```

```
## [1] 207
```

```
flat_1$flag_Pi10 <-  ifelse(flat_1$diff_Pi10 >= 0.600, 1, 0)

sum(na.exclude(flat_1$flag_Pi10))
```

```
## [1] 254
```

```
flat_1$flag_AWT <-  ifelse(flat_1$diff_AWT >= 0.189, 1, 0)

sum(na.exclude(flat_1$flag_AWT))
```

```
## [1] 259
```

```
flat_1$flagcount <- flat_1$flag_spiro_ratio + flat_1$flag_fev1 + flat_1$flag_emphys + flat_1$flag_Pi10 +

table(flat_1$flagcount)
```

```
##
##    0    1    2    3    4
## 3206  630  185   23    7
```

```
#based on this table, we can see that most people only experience one of these markers if at all but ne
```

In terms of the mortality dataset:

```r
length(mort_adjud$vital_status) #number of subjects in the cohort
```

```
## [1] 10720
```

```r
sum(na.exclude(mort_adjud$vital_status)) #number of deaths in the cohort
```

```
## [1] 1795
```

```r
summaries(mort_adjud$months_followed_net) #83 months of average follow up - lower 5% was 13
```

```
##        Mean          SD           N     Minimum     Maximum .05,.5, .95
##       83.06       31.10    10720.00        0.00      128.60       13.10
##        <NA>        <NA>
##       91.70      119.40
```

```r
summaries(mort_adjud$days_followed) # mean days followed was 2491.7 (little weird, 2065 was average day
```

```
##        Mean          SD           N     Minimum     Maximum .05,.5, .95
##     2491.72      933.00    10720.00        0.00     3858.00      394.00
##        <NA>        <NA>
##     2751.00     3582.00
```

```r
#if we subset to only subjects who died:

mort_dead <- subset(mort_adjud, mort_adjud$vital_status == 1)
summaries(mort_dead$months_followed_net) #52 months of average follow up time, lower 5% was 7
```

```
##        Mean          SD           N     Minimum     Maximum .05,.5, .95
##       52.88       28.75     1795.00        0.00      120.10        7.30
##        <NA>        <NA>
##       51.60       98.96
```

```r
summaries(mort_dead$days_followed) # mean days followed was 1586.3
```

```
##        Mean          SD           N     Minimum     Maximum .05,.5, .95
##     1586.34      862.57     1795.00        0.00     3603.00      218.00
##        <NA>        <NA>
##     1549.00     2969.50
```

Merging the datasets:

```r
fulldata <- merge(flat_1,mort_adjud,by="sid")

fulldata$sid <- as.factor(fulldata$sid)

#subset only to people who had a visit 2 date.

fulldata$check2 <- ifelse(is.na(fulldata$Visit_Date_P2), 1, 0)

datavisit2 <- subset(fulldata, fulldata$check2 == 0)

#subset only to people who had a marked change in one of the the biomarkers of interest and a visit 2

#datasicker <- subset(fulldata, fulldata$flagcount > 0)


#Pull a random sample of subject ids to evaluate the biomarker progression of the subjects
```

```
set.seed(245)
randomsid <- sample(datavisit2$sid, 30, replace=FALSE)
#$randomsidsick <- sample(datasicker$sid, 30, replace=FALSE)

#create binary variable for where the item is true

datavisit2$check <- ifelse(datavisit2$sid %in% randomsid, 1, 0)
#datasicker$check <- ifelse(datasicker$sid %in% randomsidsick, 1, 0)
#subset our dataframe to just have these values:
randomdat <- subset(datavisit2, datavisit2$check == 1)

#randomsick <- subset(datasicker, datasicker$check == 1)
```

Now make all of the necessary plots that show the progression of the different biomarkers.

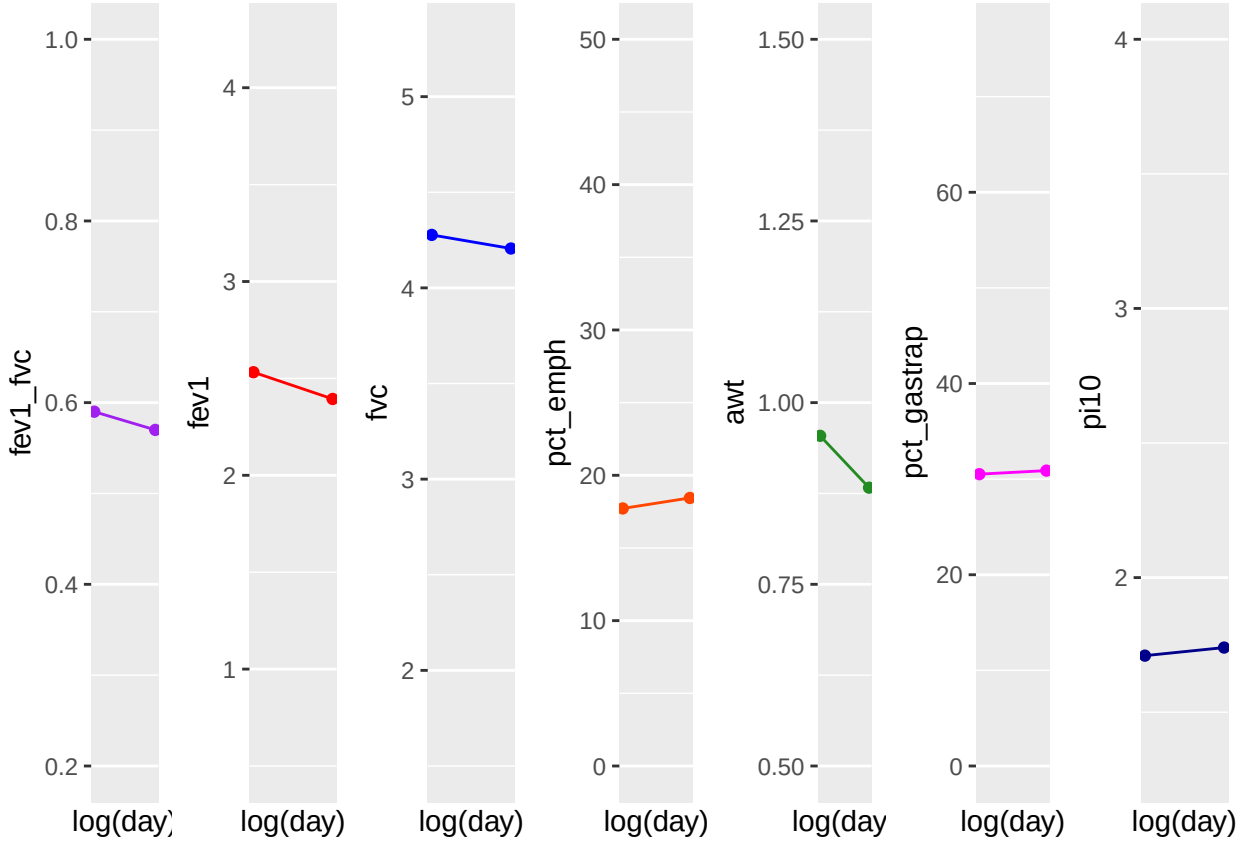
Problem: need to change it to long format instead of wide.

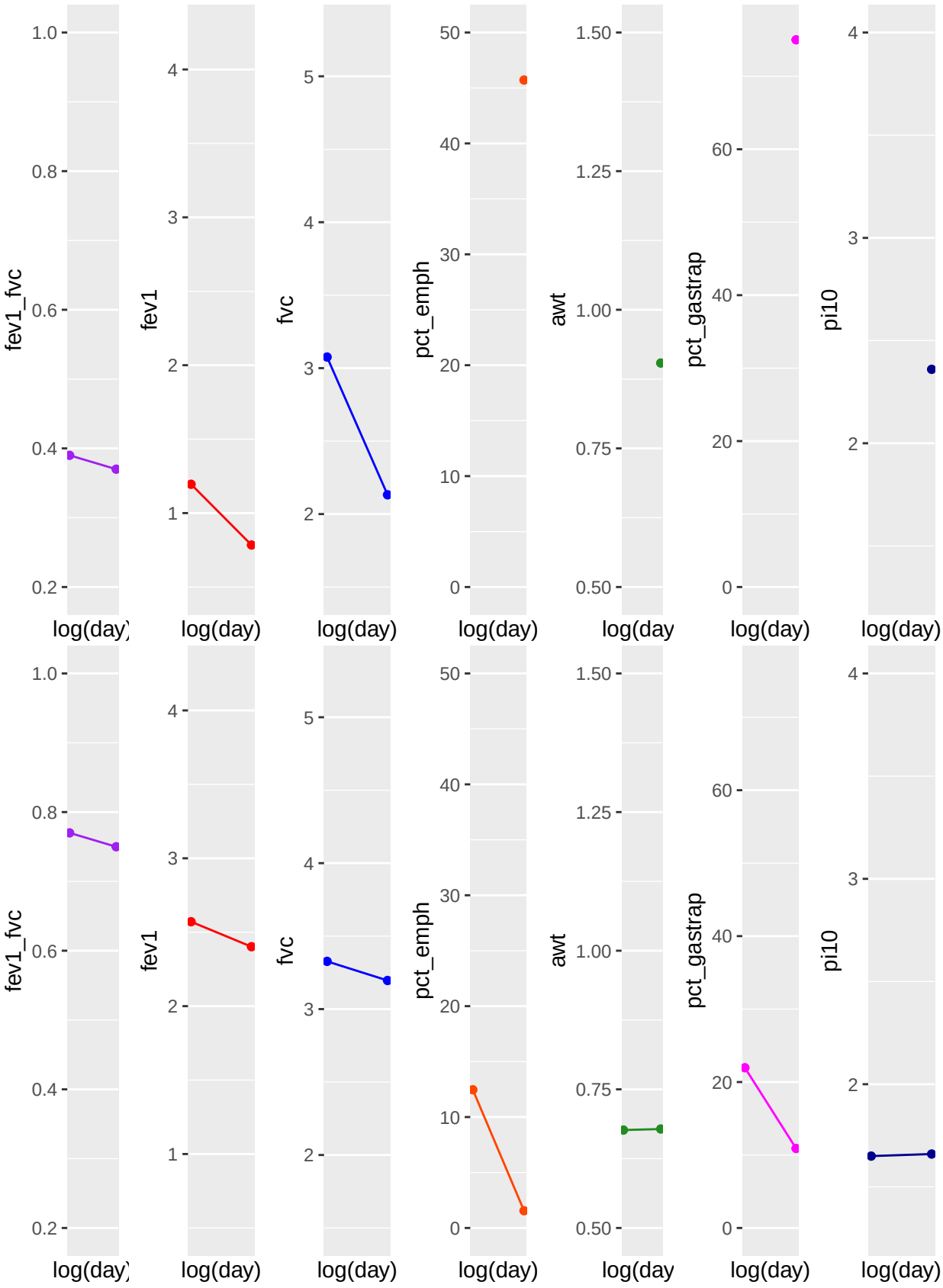
This is for everyone in the data set, even those that don't hit the extra sick markers.

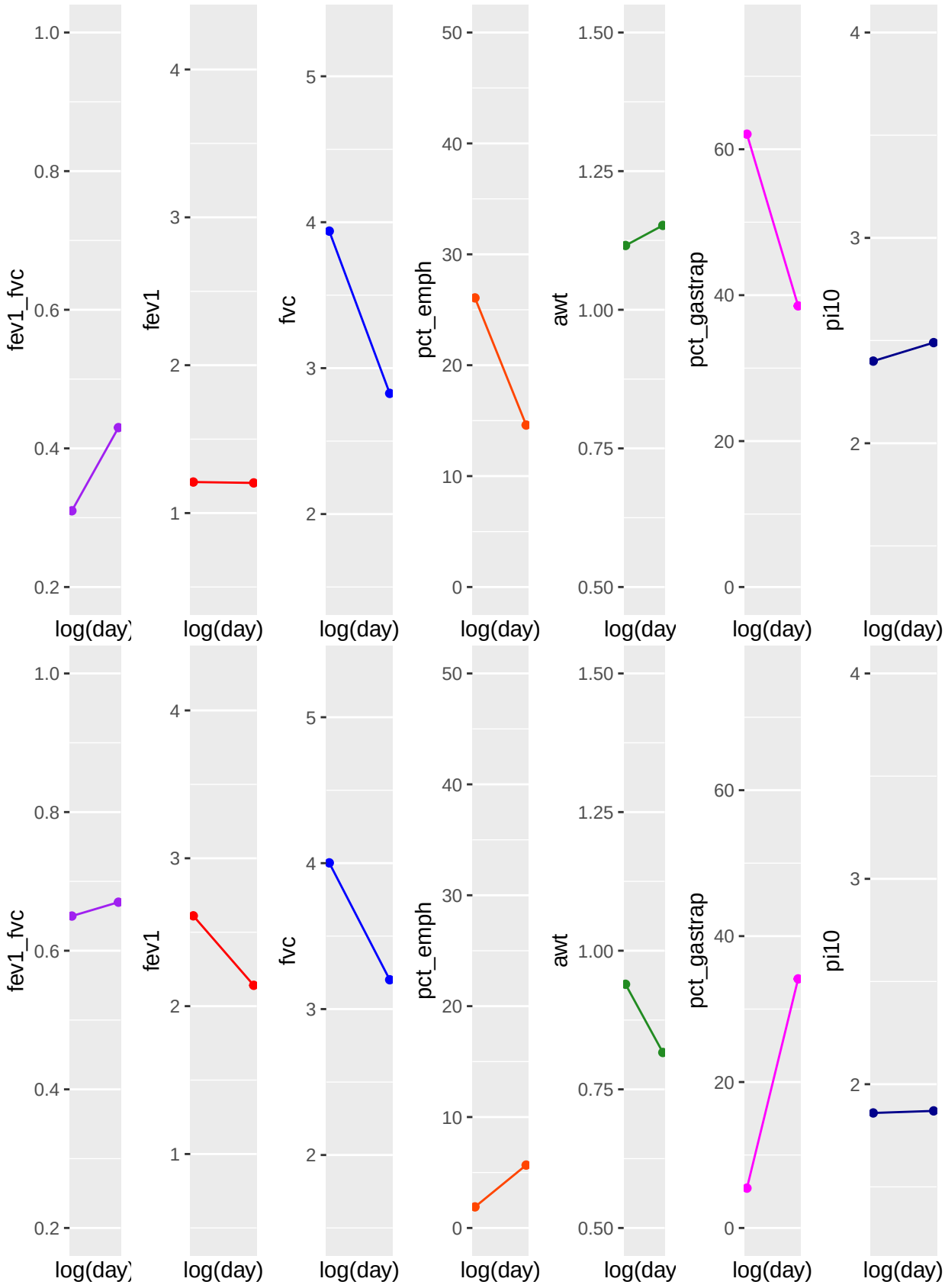
```
widerandom <- randomdat %>% select(sid, Visit_Date_P1, Visit_Date_P2, FEV1_FVC_utah_P1,FEV1_FVC_utah_P2

library(reshape)
library(reshape2)
library(ggplot2)
library(magrittr)
library(dplyr)
library(gridExtra)


#great, now it's in long format and I can start creating the graphs that I need.
longrandom <- reshape(widerandom,idvar='sid', direction='long',
        varying=list(c(2,3), c(4,5), c(6,7), c(8,9), c(10,11), c(12,13), c(14,15), c(16,17)), #note tha
        timevar='visit',
        times=c('p1', 'p2'),
        v.names=c('visitdate','fev1_fvc', 'fev1', 'fvc', 'pct_emph', 'awt', 'pct_gastrap', 'pi10'))


#for practice: we'll just do it on one item first




plotting <-function(x) {
a <- ggplot(x, aes(log(visitdate), fev1_fvc)) +
  geom_point(color = 'purple') +geom_path(color = 'purple') + scale_x_continuous(breaks=seq(9.75,9.85, 9
b <- ggplot(x, aes(log(visitdate), fev1)) +
  geom_point(color = 'red') +geom_path(color = 'red') + scale_x_continuous(breaks=seq(9.75,9.85, 9.96))
c <- ggplot(x, aes(log(visitdate), fvc)) +
  geom_point(color='blue') +geom_path(color='blue') + scale_x_continuous(breaks=seq(9.75,9.85, 9.96)) +
d <- ggplot(x, aes(log(visitdate), pct_emph)) +
  geom_point(color='orangered') +geom_path(color='orangered') + scale_x_continuous(breaks=seq(9.75,9.85
e <- ggplot(x, aes(log(visitdate), awt)) +
  geom_point(color = 'forestgreen') +geom_path(color = 'forestgreen') + scale_x_continuous(breaks=seq(9
f <- ggplot(x, aes(log(visitdate), pct_gastrap)) +
```

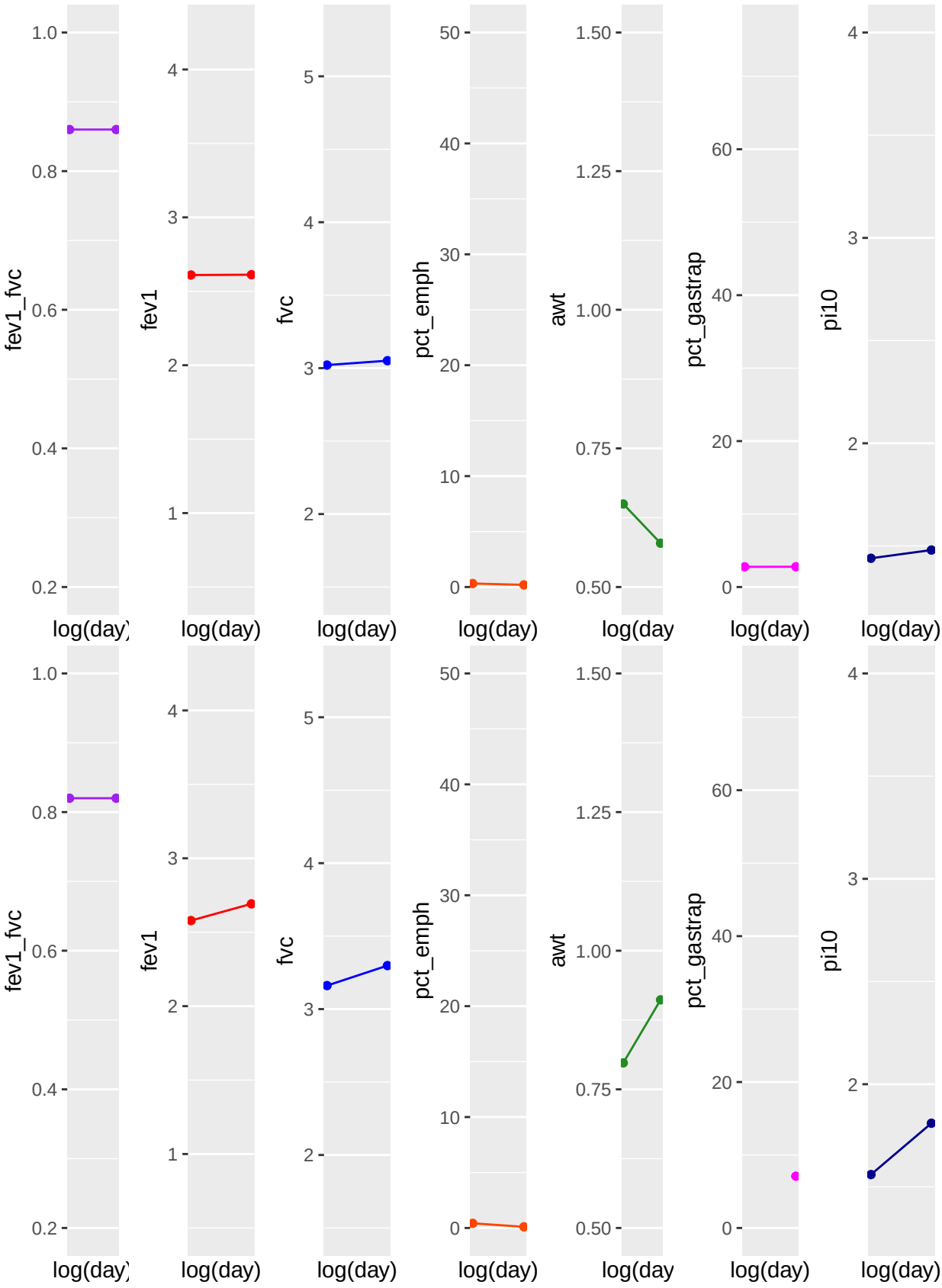
```
  geom_point(color='magenta') +geom_path(color='magenta') + scale_x_continuous(breaks=seq(9.75,9.85, 9.
g <- ggplot(x, aes(log(visitdate), pi10)) +
  geom_point(color = 'darkblue') +geom_path(color = 'darkblue')  + scale_x_continuous(breaks=seq(9.75,9


return(grid.arrange(a,b,c,d,e,f,g, ncol=7))

}
```

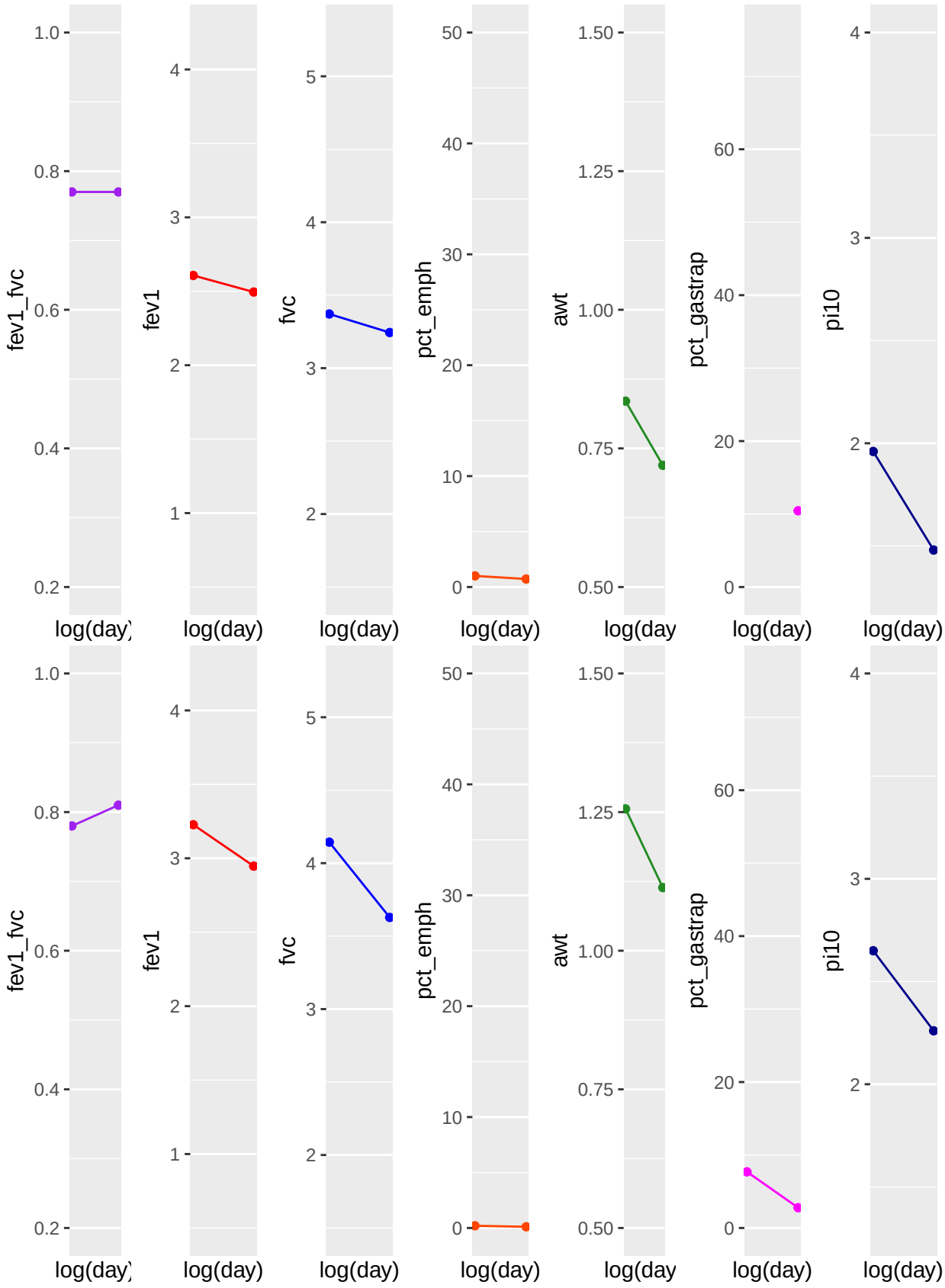
```
subjects <- unique(longrandom$sid)
for (i in 1:30) {
  x <- as.data.frame(subset(longrandom, longrandom$sid == subjects[i]))
  a <- plotting(x)
  a
}
```

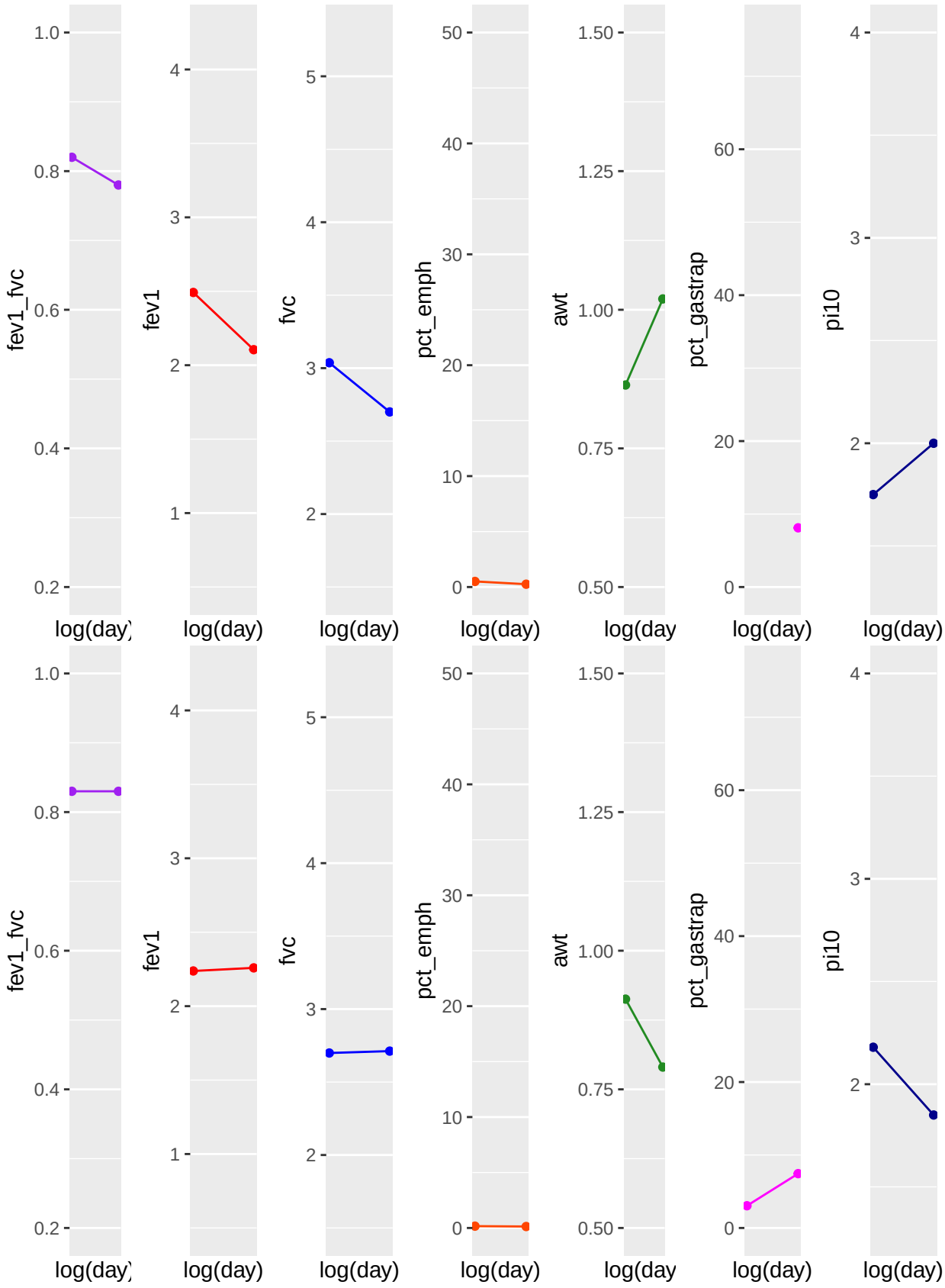


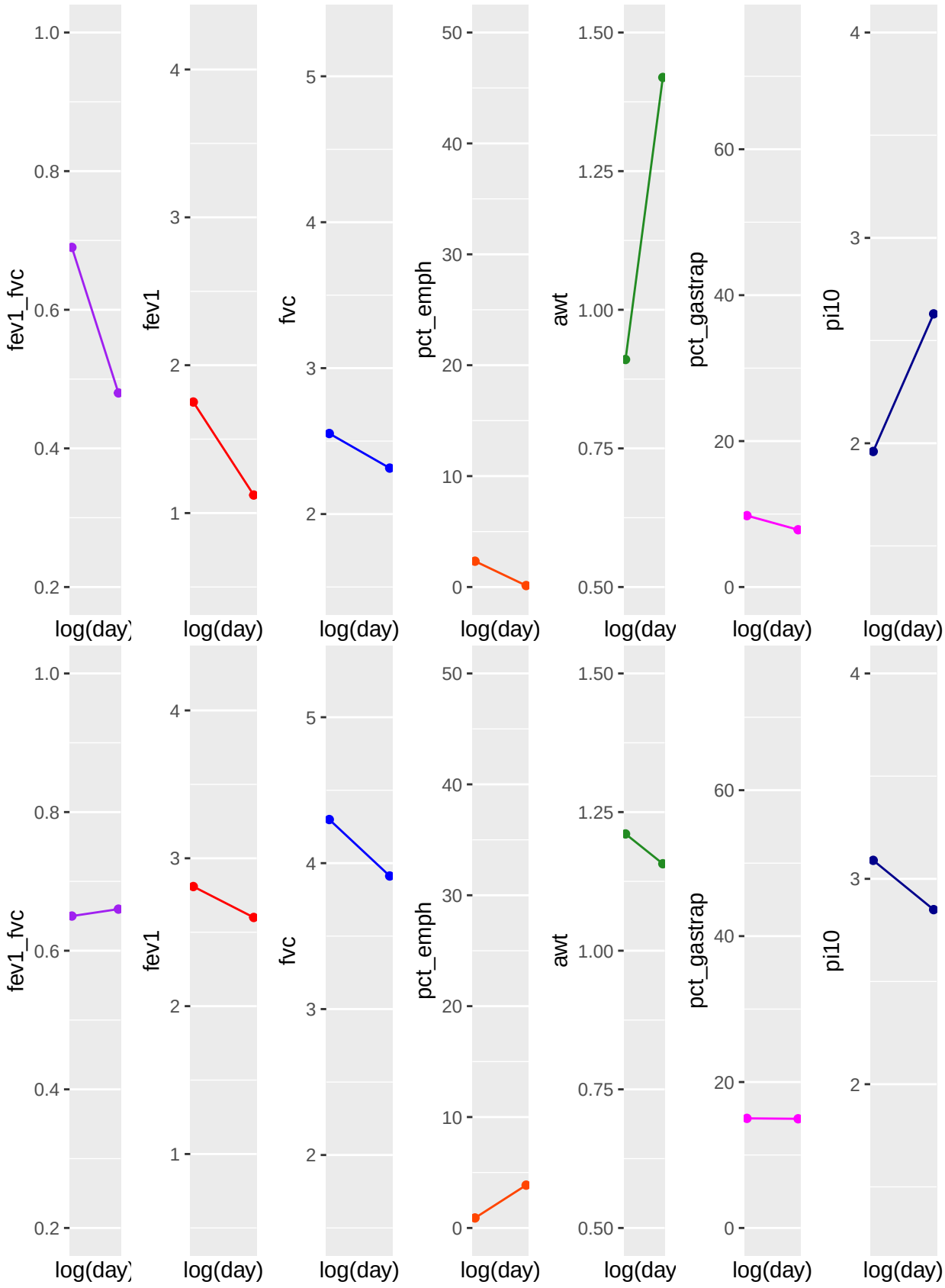

fev1_fvc
fev1
fvc
pct_emph
awt
pct_gastrap
pi10
log(day)

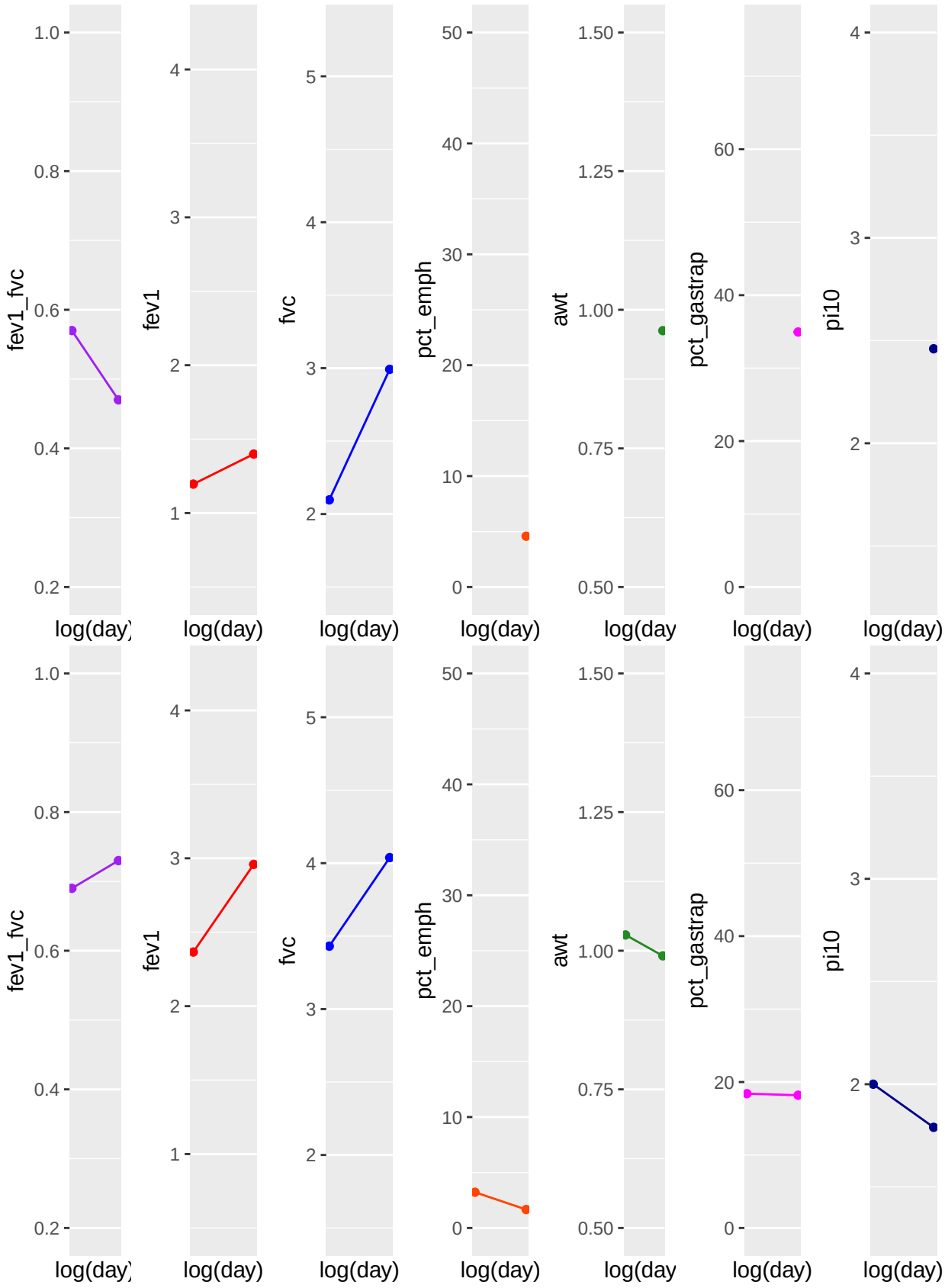

fev1_fvc
fev1
fvc
pct_emph
awt
pct_gastrap
pi10
log(day)

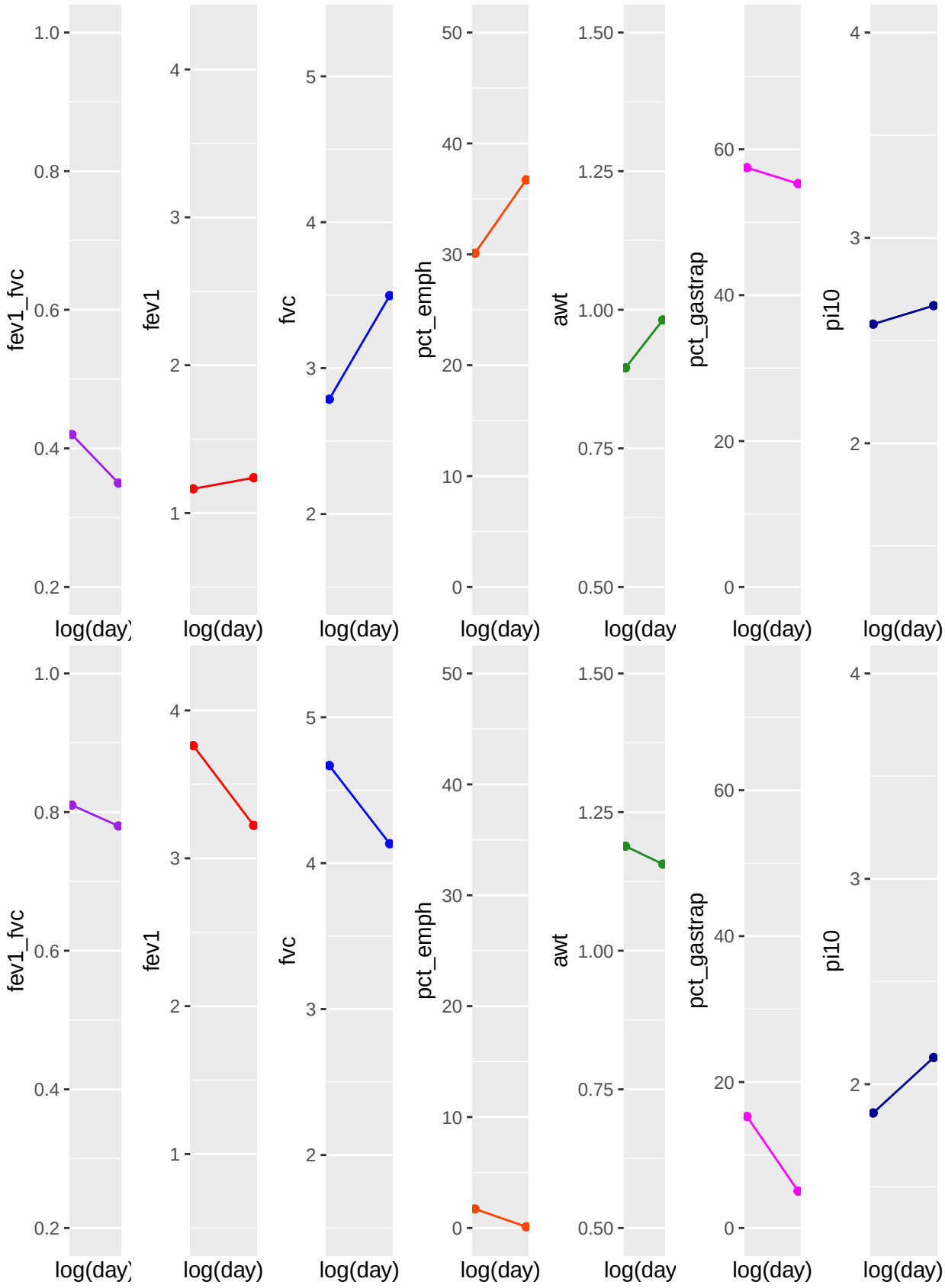

fev1_fvc
fev1
fvc
pct_emph
awt
pct_gastrap
pi10
log(day)

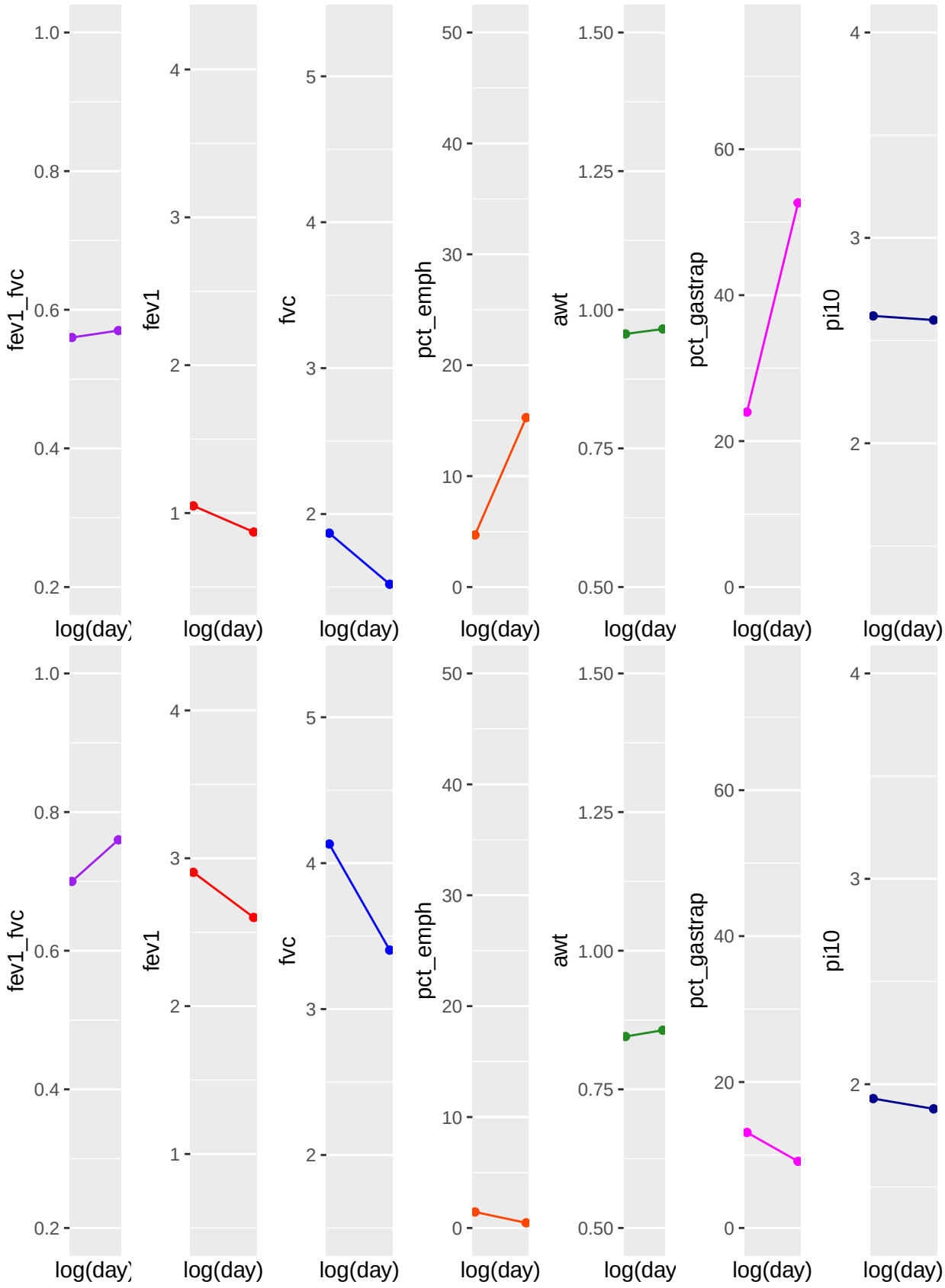

fev1_fvc
fev1
fvc
pct_emph
awt
pct_gastrap
pi10
log(day)

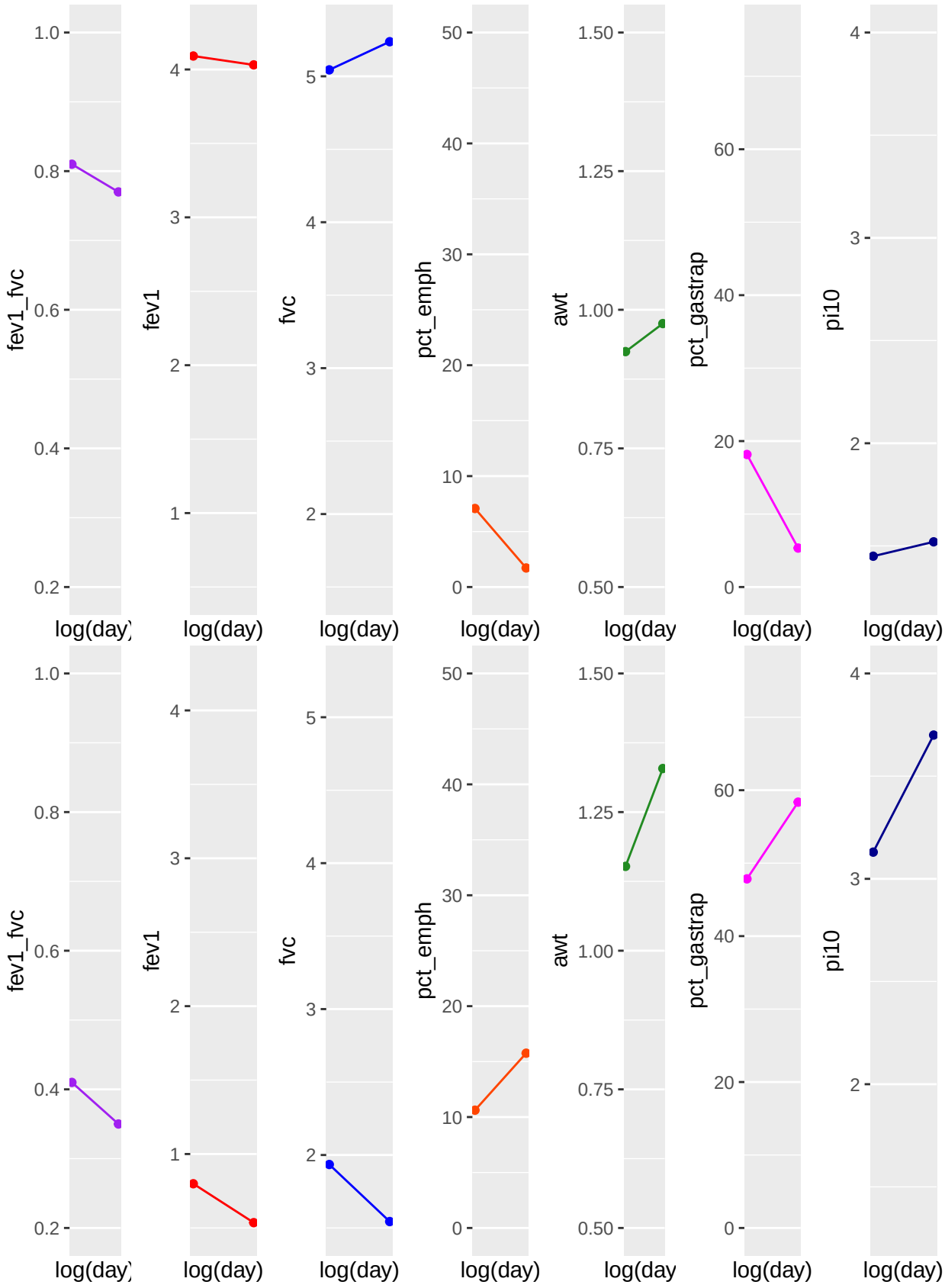

fev1_fvc
fev1
fvc
pct_emph
awt
pct_gastrap
pi10
log(day)

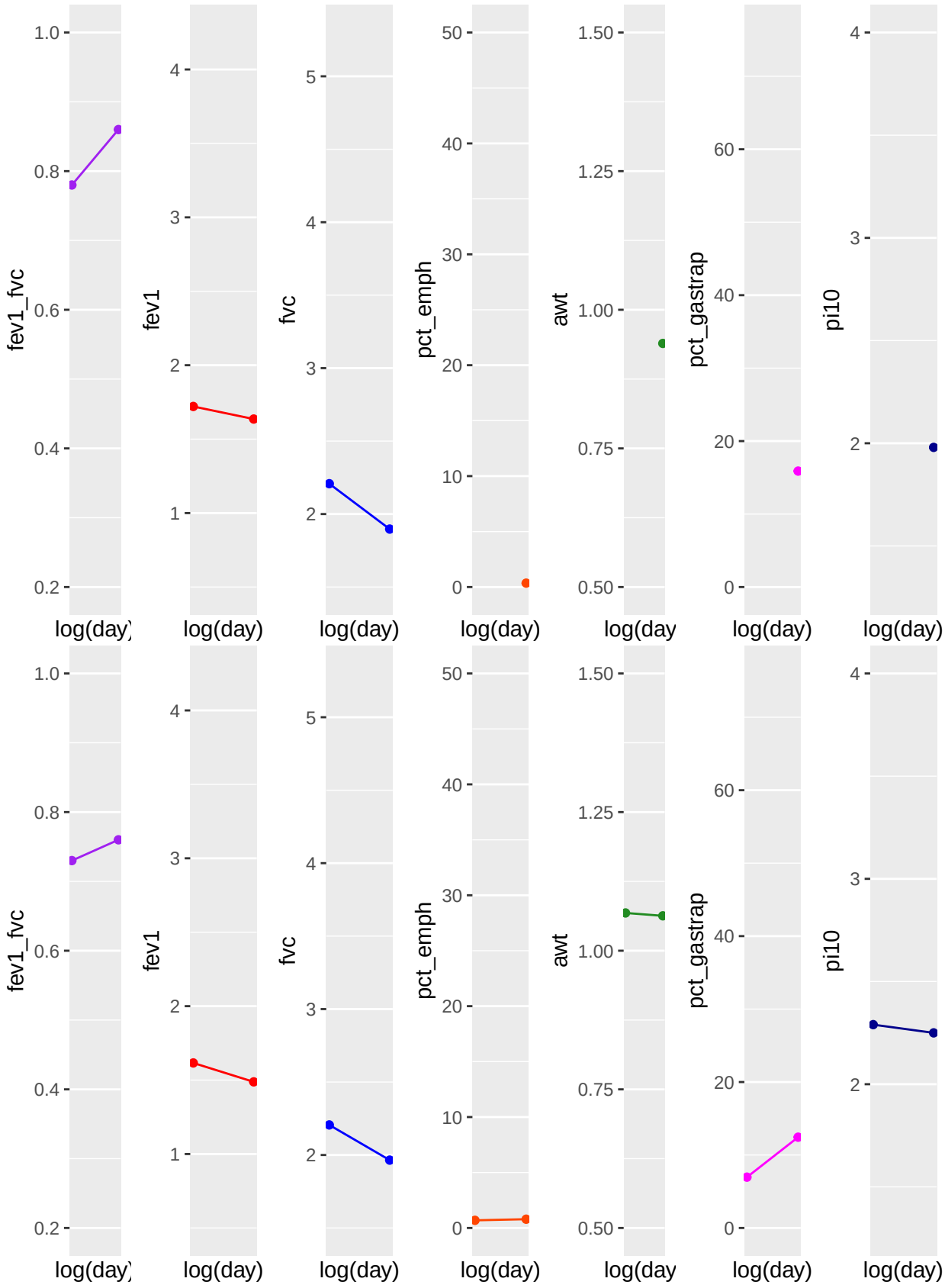

fev1_fvc
fev1
fvc
pct_emph
awt
pct_gastrap
pi10
log(day)

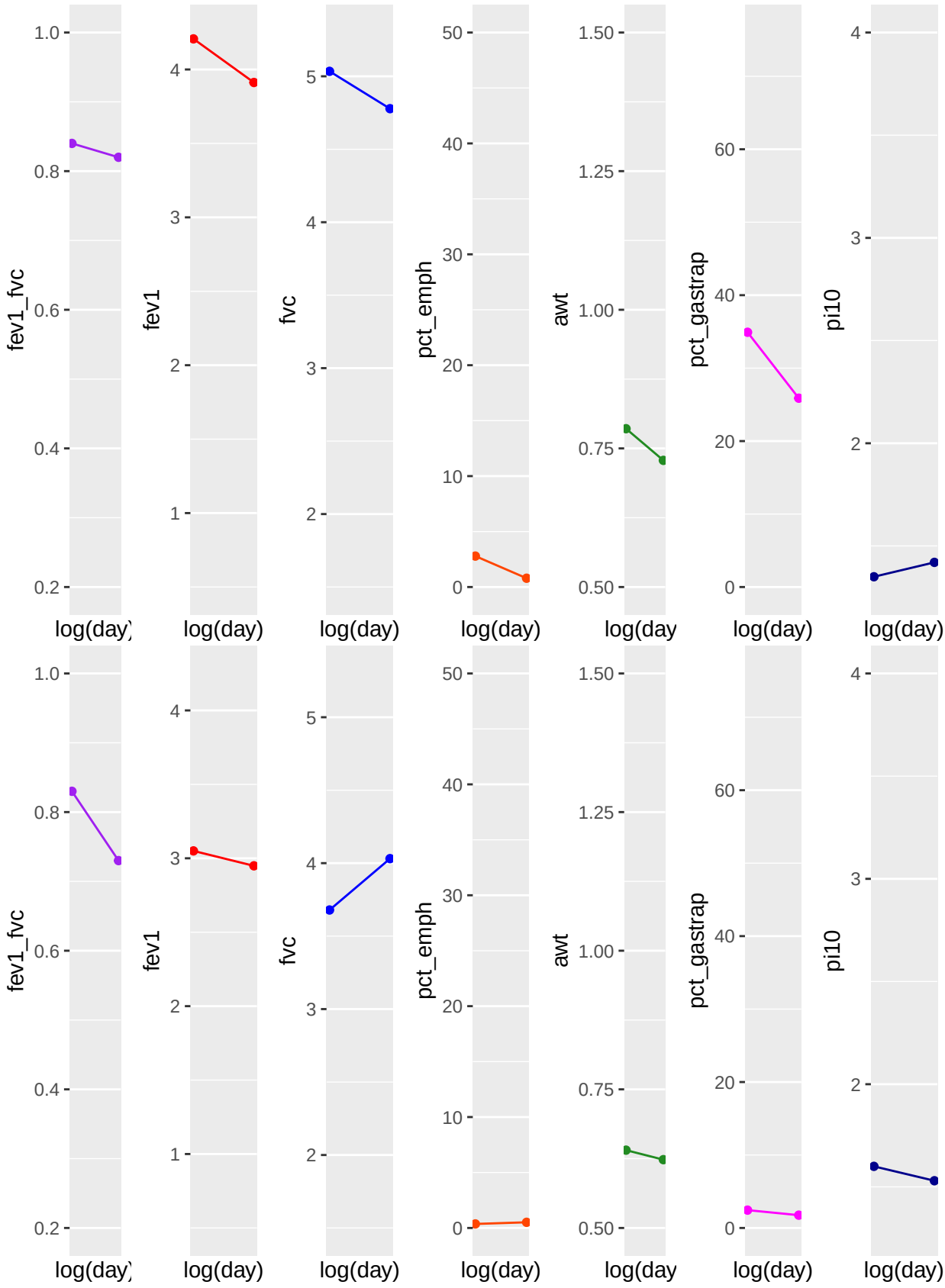

fev1_fvc
fev1
fvc
pct_emph
awt
pct_gastrap
pi10
log(day)

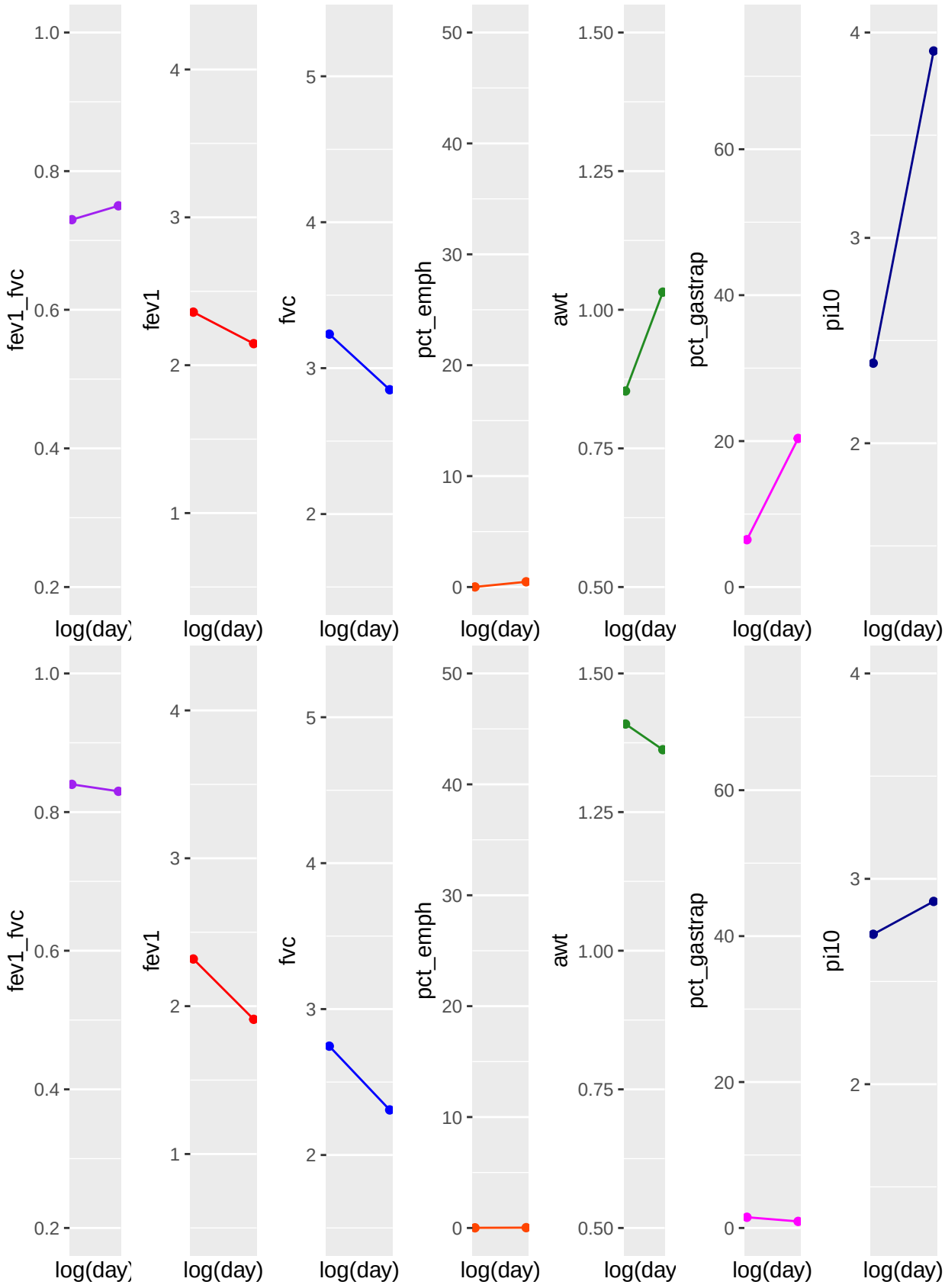

fev1_fvc
fev1
fvc
pct_emph
awt
pct_gastrap
pi10
log(day)

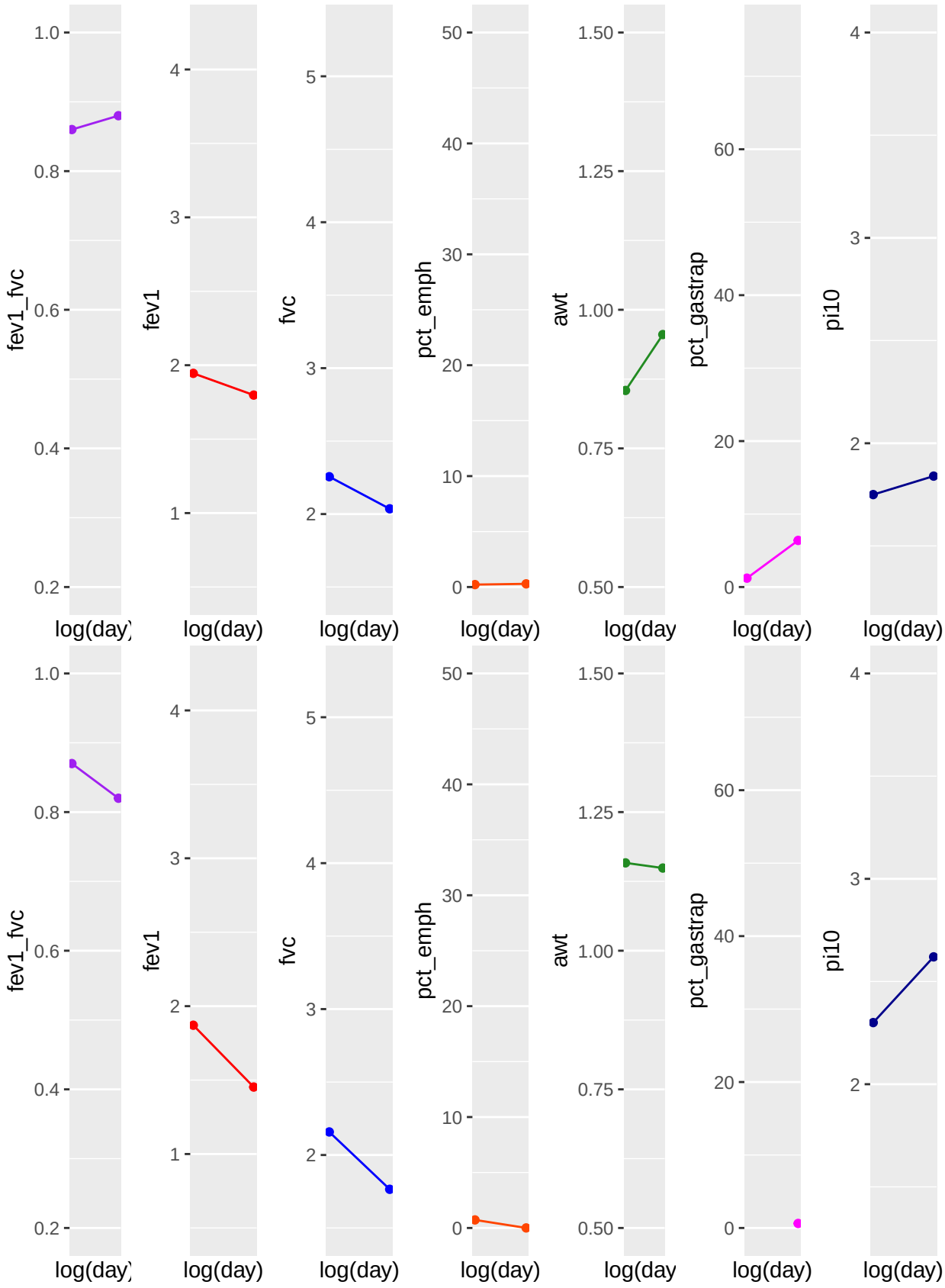

fev1_fvc
fev1
fvc
pct_emph
awt
pct_gastrap
pi10
log(day)

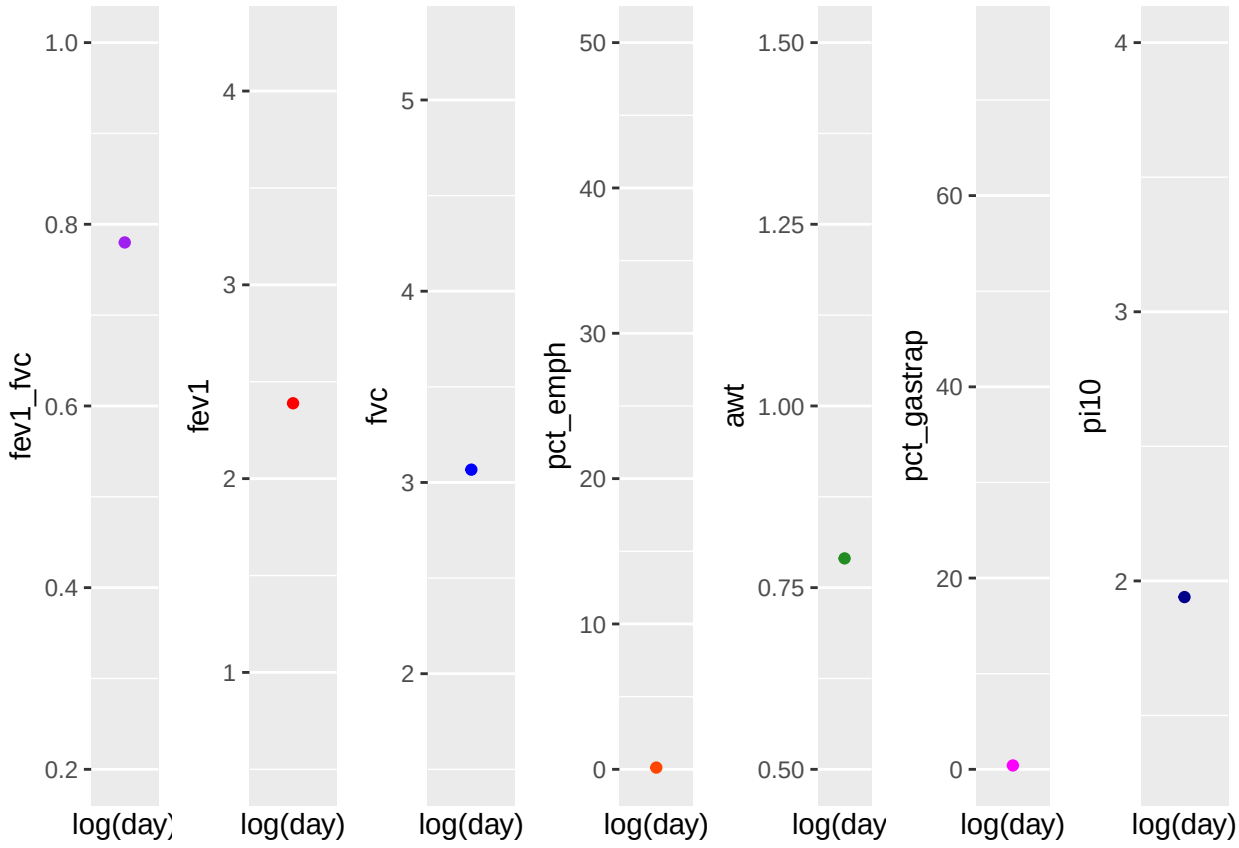

fev1_fvc
fev1
fvc
pct_emph
awt
pct_gastrap
pi10
log(day)


fev1_fvc
fev1
fvc
pct_emph
awt
pct_gastrap
pi10
log(day)


fev1_fvc
fev1
fvc
pct_emph
awt
pct_gastrap
pi10
log(day)


fev1_fvc
fev1
fvc
pct_emph
awt
pct_gastrap
pi10
log(day)


fev1_fvc
fev1
fvc
pct_emph
awt
pct_gastrap
pi10
log(day)



Change variables to fit data structures:

```
fulldata$gender <- as.factor(fulldata$gender)
fulldata$race <- as.factor(fulldata$race)
fulldata$ccenter_P1 <- as.factor(fulldata$ccenter_P1)
fulldata$ccenter_P2 <- as.factor(fulldata$ccenter_P2)
fulldata$finalGold_P1 <- as.factor(fulldata$finalGold_P1)
fulldata$finalGold_P2 <- as.factor(fulldata$finalGold_P2)
fulldata$EverSmokedCig_P1<- as.factor(fulldata$EverSmokedCig_P1)
fulldata$smoking_status_P1<- as.factor(fulldata$smoking_status_P1)
```

Now for summary tables.

We need to define the population of interest, that is; we want all subjects that either complete both visits or failed to complete both visits because they died before visit 1.

```
#subjects that died before visit 2:

fulldata$deadearly <- ifelse(is.na(fulldata$Visit_Date_P2)& fulldata$vital_status==1, 1, 0)

#subjects that made it to visit 2:
fulldata$v2attend<- ifelse(!is.na(fulldata$Visit_Date_P2), 1, 0)

#subjects that made it to visit 2 and are known to be dead

fulldata$deadlater <- ifelse(fulldata$v2attend == 1 & fulldata$vital_status ==1, 1, 0)

subjectdata <- subset(fulldata, fulldata$deadearly ==1 | fulldata$v2attend == 1 | fulldata$deadlater ==
subjectdata$group1 <- rep(0, 8158)
```

```
subjectdata$groupa <- ifelse(subjectdata$deadearly == 1, 1, subjectdata$group1)
subjectdata$groupb <- ifelse(subjectdata$v2attend == 1, 2, subjectdata$groupa)
subjectdata$group <- ifelse(subjectdata$deadlater == 1, 3, subjectdata$groupb)
```

Now, create a general table 1, then a table 1 for each of the other groups.

```
library(tableone)
## Vector of variables to summarize
myVars <- c("Age_P1", "BMI_P1", "ATS_PackYears_P1", "FEV1_FVC_utah_P1", "FEV1_utah_P1", "FVC_utah_P1",
            "pctEmph_Thirona_P1", "AWT_seg_Thirona_P1", "pctGasTrap_Thirona_P1","gender", "race", "Ever
             "finalGold_P1", "finalGold_P2")
## Vector of categorical variables that need transformation
catVars <- c("gender", "race", "EverSmokedCig_P1", "smoking_status_P1",
             "finalGold_P1", "finalGold_P2")
## Create  TableOne objects
tab2 <- CreateTableOne(vars = myVars, data = subjectdata, factorVars = catVars)
tab2 #values for all patients
```

```
##
##                                    Overall
##   n                                 8158
##   Age_P1 (mean (sd))               60.78 (8.98)
##   BMI_P1 (mean (sd))               28.85 (6.28)
##   ATS_PackYears_P1 (mean (sd))     44.49 (25.74)
##   FEV1_FVC_utah_P1 (mean (sd))      0.66 (0.17)
##   FEV1_utah_P1 (mean (sd))          2.18 (0.92)
##   FVC_utah_P1 (mean (sd))           3.26 (1.00)
##   pctEmph_Thirona_P1 (mean (sd))    6.95 (10.22)
##   AWT_seg_Thirona_P1 (mean (sd))    1.06 (0.23)
##   pctGasTrap_Thirona_P1 (mean (sd)) 22.73 (20.11)
##   gender = 2 (%)                    3956 (48.5)
##   race = 2 (%)                      2290 (28.1)
##   EverSmokedCig_P1 = 1 (%)          7718 (98.8)
##   smoking_status_P1 (%)
##      0                                91 ( 1.2)
##      1                              4095 (52.4)
##      2                              3623 (46.4)
##   finalGold_P1 (%)
##      -2                               91 ( 1.2)
##      -1                              913 (11.8)
##      0                              3125 (40.2)
##      1                               605 ( 7.8)
##      2                              1537 (19.8)
##      3                               973 (12.5)
##      4                               525 ( 6.8)
##   finalGold_P2 (%)
##      -2                               394 ( 6.5)
##      -1                               724 (11.9)
##      0                               2438 (40.0)
##      1                                544 ( 8.9)
##      2                               1141 (18.7)
##      3                                602 ( 9.9)
##      4                                253 ( 4.2)
```

```r
summary(tab2)
```

```
##
##      ### Summary of continuous variables ###
##
## strata: Overall
##                           n miss p.miss mean   sd median  p25  p75   min
## Age_P1                 8158  349      4 60.8  9.0   60.7 53.3 67.7 4e+01
## BMI_P1                 8158  349      4 28.8  6.3   28.0 24.4 32.3 1e+01
## ATS_PackYears_P1       8158  353      4 44.5 25.7   40.0 27.0 55.8 0e+00
## FEV1_FVC_utah_P1       8158  389      5  0.7  0.2    0.7  0.6  0.8 1e-01
## FEV1_utah_P1           8158  389      5  2.2  0.9    2.2  1.5  2.8 2e-01
## FVC_utah_P1            8158  389      5  3.3  1.0    3.2  2.5  3.9 6e-01
## pctEmph_Thirona_P1     8158  852     10  6.9 10.2    2.5  0.7  8.3 2e-04
## AWT_seg_Thirona_P1     8158  853     10  1.1  0.2    1.0  0.9  1.2 5e-01
## pctGasTrap_Thirona_P1 8158 1817     22 22.7 20.1   15.5  7.0 33.8 3e-02
##                          max skew   kurt
## Age_P1                    85  0.1 -0.910
## BMI_P1                    64  0.9  1.168
## ATS_PackYears_P1         332  1.5  5.178
## FEV1_FVC_utah_P1           1 -0.9 -0.188
## FEV1_utah_P1               6  0.2 -0.442
## FVC_utah_P1                8  0.4 -0.003
## pctEmph_Thirona_P1        62  2.2  4.469
## AWT_seg_Thirona_P1         2  0.7  0.661
## pctGasTrap_Thirona_P1     84  1.0  0.046
##
## ============================================================================================
##
##      ### Summary of categorical variables ###
##
## strata: Overall
##                  var    n miss p.miss level freq percent cum.percent
##               gender 8158    0    0.0     1 4202    51.5        51.5
##                                          2 3956    48.5       100.0
##
##                 race 8158    0    0.0     1 5868    71.9        71.9
##                                          2 2290    28.1       100.0
##
##    EverSmokedCig_P1 8158  349    4.3     0   91     1.2         1.2
##                                          1 7718    98.8       100.0
##
##   smoking_status_P1 8158  349    4.3     0   91     1.2         1.2
##                                          1 4095    52.4        53.6
##                                          2 3623    46.4       100.0
##
##         finalGold_P1 8158  389    4.8    -2   91     1.2         1.2
##                                         -1  913    11.8        12.9
##                                          0 3125    40.2        53.1
##                                          1  605     7.8        60.9
##                                          2 1537    19.8        80.7
##                                          3  973    12.5        93.2
##                                          4  525     6.8       100.0
##
```

```
##       finalGold_P2 8158 2062   25.3    -2  394     6.5         6.5
##                                        -1  724    11.9        18.3
##                                         0 2438    40.0        58.3
##                                         1  544     8.9        67.3
##                                         2 1141    18.7        86.0
##                                         3  602     9.9        95.8
##                                         4  253     4.2       100.0
##
```

```
subjectdata$group <- as.factor(subjectdata$group)
## Vector of variables to summarize
myVars <- c("Age_P1", "BMI_P1", "ATS_PackYears_P1", "FEV1_FVC_utah_P1", "FEV1_utah_P1", "FVC_utah_P1",
            "pctEmph_Thirona_P1", "AWT_seg_Thirona_P1", "pctGasTrap_Thirona_P1","gender", "race", "Ever
             "finalGold_P1", "finalGold_P2", "group")
## Vector of categorical variables that need transformation
catVars <- c("gender", "race", "EverSmokedCig_P1", "smoking_status_P1",
             "finalGold_P1", "finalGold_P2", "group")
## Create  TableOne objects
tab3 <- CreateTableOne(vars = myVars, strata = c("group"), data = subjectdata, factorVars = catVars)
tab3 #values for all patients
```

```
##                                    Stratified by group
##                                     1              2
##   n                                  1400           6363
##   Age_P1 (mean (sd))                63.71 (9.16)   59.80 (8.71)
##   BMI_P1 (mean (sd))                27.61 (6.65)   29.17 (6.15)
##   ATS_PackYears_P1 (mean (sd))      53.94 (30.64)  41.76 (23.82)
##   FEV1_FVC_utah_P1 (mean (sd))       0.54 (0.20)    0.69 (0.14)
##   FEV1_utah_P1 (mean (sd))           1.61 (0.95)    2.34 (0.85)
##   FVC_utah_P1 (mean (sd))            2.87 (1.04)    3.36 (0.97)
##   pctEmph_Thirona_P1 (mean (sd))    13.54 (14.55)   5.12 (7.90)
##   AWT_seg_Thirona_P1 (mean (sd))     1.15 (0.25)    1.03 (0.22)
##   pctGasTrap_Thirona_P1 (mean (sd)) 36.52 (24.43)  18.86 (17.04)
##   gender = 2 (%)                      545 ( 38.9)   3249 ( 51.1)
##   race = 2 (%)                        379 ( 27.1)   1822 ( 28.6)
##   EverSmokedCig_P1 = 1 (%)           1395 ( 99.6)   5931 ( 98.6)
##   smoking_status_P1 (%)
##      0                                  5 (  0.4)     85 (  1.4)
##      1                                759 ( 54.2)   3106 ( 51.6)
##      2                                636 ( 45.4)   2825 ( 47.0)
##   finalGold_P1 (%)
##      -2                                 5 (  0.4)     85 (  1.4)
##      -1                               139 ( 10.0)    739 ( 12.3)
##      0                                262 ( 18.8)   2794 ( 46.7)
##      1                                 63 (  4.5)    519 (  8.7)
##      2                                266 ( 19.1)   1155 ( 19.3)
##      3                                314 ( 22.6)    549 (  9.2)
##      4                                343 ( 24.6)    143 (  2.4)
##   finalGold_P2 (%)
##      -2                                 0 (  NaN)    391 (  6.8)
##      -1                                 0 (  NaN)    690 ( 12.0)
##      0                                  0 (  NaN)   2394 ( 41.5)
##      1                                  0 (  NaN)    520 (  9.0)
##      2                                  0 (  NaN)   1052 ( 18.2)
##      3                                  0 (  NaN)    518 (  9.0)
```

```
##      4                                    0 (  NaN)    203 (  3.5)
##   group (%)
##      1                                 1400 (100.0)      0 (  0.0)
##      2                                    0 (  0.0)   6363 (100.0)
##      3                                    0 (  0.0)      0 (  0.0)
##                                     Stratified by group
##                                      3              p      test
##   n                                    395
##   Age_P1 (mean (sd))                 65.32 (8.77)   <0.001
##   BMI_P1 (mean (sd))                 28.38 (6.39)   <0.001
##   ATS_PackYears_P1 (mean (sd))       52.71 (25.22)  <0.001
##   FEV1_FVC_utah_P1 (mean (sd))        0.56 (0.17)   <0.001
##   FEV1_utah_P1 (mean (sd))            1.75 (0.87)   <0.001
##   FVC_utah_P1 (mean (sd))             3.07 (1.01)   <0.001
##   pctEmph_Thirona_P1 (mean (sd))     12.68 (12.55)  <0.001
##   AWT_seg_Thirona_P1 (mean (sd))      1.12 (0.23)   <0.001
##   pctGasTrap_Thirona_P1 (mean (sd)) 35.32 (21.81)  <0.001
##   gender = 2 (%)                       162 ( 41.0)  <0.001
##   race = 2 (%)                          89 ( 22.5)   0.021
##   EverSmokedCig_P1 = 1 (%)             392 ( 99.7)   0.001
##   smoking_status_P1 (%)                              <0.001
##      0                                   1 (  0.3)
##      1                                 230 ( 58.5)
##      2                                 162 ( 41.2)
##   finalGold_P1 (%)                                   <0.001
##      -2                                  1 (  0.3)
##      -1                                 35 (  8.9)
##      0                                  69 ( 17.6)
##      1                                  23 (  5.9)
##      2                                 116 ( 29.5)
##      3                                 110 ( 28.0)
##      4                                  39 (  9.9)
##   finalGold_P2 (%)                                    NaN
##      -2                                  3 (  0.9)
##      -1                                 34 ( 10.4)
##      0                                  44 ( 13.4)
##      1                                  24 (  7.3)
##      2                                  89 ( 27.1)
##      3                                  84 ( 25.6)
##      4                                  50 ( 15.2)
##   group (%)                                          <0.001
##      1                                   0 (  0.0)
##      2                                   0 (  0.0)
##      3                                 395 (100.0)
```

```
summary(tab3)
```

```
##
##      ### Summary of continuous variables ###
##
## group: 1
##                          n miss p.miss mean   sd median  p25  p75   min
## Age_P1                1400    0   0.00 63.7  9.2   64.5 56.2 71.1 44.80
## BMI_P1                1400    0   0.00 27.6  6.7   26.6 22.8 31.2 12.29
## ATS_PackYears_P1      1400    1   0.07 53.9 30.6   47.0 34.0 68.0  0.00
```

```
## FEV1_FVC_utah_P1      1400    8   0.57  0.5  0.2    0.5  0.4  0.7  0.15
## FEV1_utah_P1          1400    8   0.57  1.6  1.0    1.4  0.8  2.3  0.22
## FVC_utah_P1           1400    8   0.57  2.9  1.0    2.8  2.1  3.5  0.65
## pctEmph_Thirona_P1    1400  142  10.14 13.5 14.5    7.2  1.4 23.0  0.01
## AWT_seg_Thirona_P1    1400  142  10.14  1.2  0.2    1.1  1.0  1.3  0.64
## pctGasTrap_Thirona_P1 1400  324  23.14 36.5 24.4   34.3 12.8 59.0  0.07
##                         max  skew kurt
## Age_P1                 81.0 -0.16 -1.0
## BMI_P1                 58.6  0.93  1.2
## ATS_PackYears_P1      216.0  1.45  3.0
## FEV1_FVC_utah_P1        0.9  0.05 -1.3
## FEV1_utah_P1            5.5  0.71 -0.4
## FVC_utah_P1             7.7  0.57  0.3
## pctEmph_Thirona_P1     61.7  1.01 -0.1
## AWT_seg_Thirona_P1      2.4  0.65  0.9
## pctGasTrap_Thirona_P1  83.8  0.13 -1.4
## ------------------------------------------------------------
## group: 2
##                         n miss p.miss mean   sd median  p25  p75   min
## Age_P1                6363  347      5 59.8  8.7   59.6 52.6 66.3 4e+01
## BMI_P1                6363  347      5 29.2  6.2   28.2 24.8 32.5 1e+01
## ATS_PackYears_P1      6363  349      5 41.8 23.8   38.0 25.0 52.5 0e+00
## FEV1_FVC_utah_P1      6363  379      6  0.7  0.1    0.7  0.6  0.8 2e-01
## FEV1_utah_P1          6363  379      6  2.3  0.8    2.3  1.7  2.9 3e-01
## FVC_utah_P1           6363  379      6  3.4  1.0    3.3  2.6  4.0 6e-01
## pctEmph_Thirona_P1    6363  678     11  5.1  7.9    1.9  0.6  5.8 9e-04
## AWT_seg_Thirona_P1    6363  679     11  1.0  0.2    1.0  0.9  1.2 5e-01
## pctGasTrap_Thirona_P1 6363 1434     23 18.9 17.0   13.2  6.1 26.5 3e-02
##                        max skew  kurt
## Age_P1                  85  0.2 -0.82
## BMI_P1                  64  0.9  1.22
## ATS_PackYears_P1       332  1.5  6.36
## FEV1_FVC_utah_P1         1 -1.1  0.68
## FEV1_utah_P1             5  0.2 -0.25
## FVC_utah_P1              7  0.5 -0.05
## pctEmph_Thirona_P1      61  2.7  8.23
## AWT_seg_Thirona_P1       2  0.6  0.49
## pctGasTrap_Thirona_P1   81  1.3  1.02
## ------------------------------------------------------------
## group: 3
##                         n miss p.miss mean   sd median  p25  p75   min
## Age_P1                395    2    0.5 65.3  8.8   66.4 58.8 72.2 5e+01
## BMI_P1                395    2    0.5 28.4  6.4   27.5 23.8 31.8 1e+01
## ATS_PackYears_P1      395    3    0.8 52.7 25.2   48.1 35.0 69.0 0e+00
## FEV1_FVC_utah_P1      395    2    0.5  0.6  0.2    0.6  0.4  0.7 2e-01
## FEV1_utah_P1          395    2    0.5  1.7  0.9    1.6  1.0  2.3 3e-01
## FVC_utah_P1           395    2    0.5  3.1  1.0    2.9  2.4  3.7 1e+00
## pctEmph_Thirona_P1    395   32    8.1 12.7 12.6    8.3  2.3 19.6 2e-04
## AWT_seg_Thirona_P1    395   32    8.1  1.1  0.2    1.1  1.0  1.2 6e-01
## pctGasTrap_Thirona_P1 395   59   14.9 35.3 21.8   34.0 15.9 52.7 9e-01
##                         max  skew kurt
## Age_P1                 81.0 -0.31 -0.8
## BMI_P1                 55.3  0.92  1.3
## ATS_PackYears_P1      139.0  0.78  0.4
```

```
## FEV1_FVC_utah_P1        0.9 -0.04 -1.0
## FEV1_utah_P1            4.9  0.78  0.2
## FVC_utah_P1             7.2  0.68  0.5
## pctEmph_Thirona_P1     53.9  1.04  0.2
## AWT_seg_Thirona_P1      2.0  0.67  0.9
## pctGasTrap_Thirona_P1  84.1  0.22 -1.1
##
## p-values
##                              pNormal    pNonNormal
## Age_P1                  1.061156e-71  3.627039e-67
## BMI_P1                  1.864802e-16  4.162716e-21
## ATS_PackYears_P1        2.312096e-66  1.779420e-59
## FEV1_FVC_utah_P1      2.194871e-268 5.136081e-185
## FEV1_utah_P1          6.623453e-186 9.804496e-172
## FVC_utah_P1             1.399822e-63  2.318675e-62
## pctEmph_Thirona_P1    2.211850e-189 7.356458e-112
## AWT_seg_Thirona_P1      3.451840e-69  1.033286e-64
## pctGasTrap_Thirona_P1 1.845400e-191 4.728569e-130
##
## Standardize mean differences
##                           average    1 vs 2     1 vs 3    2 vs 3
## Age_P1                  0.4155411 0.4367032 0.17921933 0.6307008
## BMI_P1                  0.1622994 0.2432284 0.11813619 0.1255337
## ATS_PackYears_P1        0.3115292 0.4441039 0.04412855 0.4463551
## FEV1_FVC_utah_P1        0.6284087 0.9137911 0.12207532 0.8493598
## FEV1_utah_P1            0.5507079 0.8100453 0.15107263 0.6910056
## FVC_utah_P1             0.3255345 0.4875634 0.19179280 0.2972472
## pctEmph_Thirona_P1      0.5012314 0.7195766 0.06387515 0.7202425
## AWT_seg_Thirona_P1      0.3465708 0.5131694 0.12019234 0.4063507
## pctGasTrap_Thirona_P1   0.5771515 0.8386528 0.05203138 0.8407703
##
## ============================================================================================
##
##      ### Summary of categorical variables ###
##
## group: 1
##                 var    n miss p.miss level freq percent cum.percent
##              gender 1400    0    0.0     1  855    61.1        61.1
##                                          2  545    38.9       100.0
##
##                race 1400    0    0.0     1 1021    72.9        72.9
##                                          2  379    27.1       100.0
##
##    EverSmokedCig_P1 1400    0    0.0     0    5     0.4         0.4
##                                          1 1395    99.6       100.0
##
##   smoking_status_P1 1400    0    0.0     0    5     0.4         0.4
##                                          1  759    54.2        54.6
##                                          2  636    45.4       100.0
##
##        finalGold_P1 1400    8    0.6    -2    5     0.4         0.4
##                                         -1  139    10.0        10.3
##                                          0  262    18.8        29.2
##                                          1   63     4.5        33.7
```

```
##                                      2  266    19.1        52.8
##                                      3  314    22.6        75.4
##                                      4  343    24.6       100.0
##
##       finalGold_P2 1400 1400  100.0    -2    0     NaN         NaN
##                                      -1    0     NaN         NaN
##                                       0    0     NaN         NaN
##                                       1    0     NaN         NaN
##                                       2    0     NaN         NaN
##                                       3    0     NaN         NaN
##                                       4    0     NaN         NaN
##
##              group 1400    0    0.0     1 1400   100.0       100.0
##                                       2    0     0.0       100.0
##                                       3    0     0.0       100.0
##
## ------------------------------------------------------------
## group: 2
##                var    n miss p.miss level freq percent cum.percent
##             gender 6363    0    0.0     1 3114    48.9        48.9
##                                       2 3249    51.1       100.0
##
##               race 6363    0    0.0     1 4541    71.4        71.4
##                                       2 1822    28.6       100.0
##
##   EverSmokedCig_P1 6363  347    5.5     0   85     1.4         1.4
##                                       1 5931    98.6       100.0
##
##  smoking_status_P1 6363  347    5.5     0   85     1.4         1.4
##                                       1 3106    51.6        53.0
##                                       2 2825    47.0       100.0
##
##       finalGold_P1 6363  379    6.0    -2   85     1.4         1.4
##                                      -1  739    12.3        13.8
##                                       0 2794    46.7        60.5
##                                       1  519     8.7        69.1
##                                       2 1155    19.3        88.4
##                                       3  549     9.2        97.6
##                                       4  143     2.4       100.0
##
##       finalGold_P2 6363  595    9.4    -2  391     6.8         6.8
##                                      -1  690    12.0        18.7
##                                       0 2394    41.5        60.2
##                                       1  520     9.0        69.3
##                                       2 1052    18.2        87.5
##                                       3  518     9.0        96.5
##                                       4  203     3.5       100.0
##
##              group 6363    0    0.0     1    0     0.0         0.0
##                                       2 6363   100.0       100.0
##                                       3    0     0.0       100.0
##
## ------------------------------------------------------------
## group: 3
```

```
##                var   n miss p.miss level freq percent cum.percent
##             gender 395    0    0.0     1  233    59.0        59.0
##                                        2  162    41.0       100.0
##
##               race 395    0    0.0     1  306    77.5        77.5
##                                        2   89    22.5       100.0
##
##   EverSmokedCig_P1 395    2    0.5     0    1     0.3         0.3
##                                        1  392    99.7       100.0
##
##  smoking_status_P1 395    2    0.5     0    1     0.3         0.3
##                                        1  230    58.5        58.8
##                                        2  162    41.2       100.0
##
##       finalGold_P1 395    2    0.5    -2    1     0.3         0.3
##                                       -1   35     8.9         9.2
##                                        0   69    17.6        26.7
##                                        1   23     5.9        32.6
##                                        2  116    29.5        62.1
##                                        3  110    28.0        90.1
##                                        4   39     9.9       100.0
##
##       finalGold_P2 395   67   17.0    -2    3     0.9         0.9
##                                       -1   34    10.4        11.3
##                                        0   44    13.4        24.7
##                                        1   24     7.3        32.0
##                                        2   89    27.1        59.1
##                                        3   84    25.6        84.8
##                                        4   50    15.2       100.0
##
##              group 395    0    0.0     1    0     0.0         0.0
##                                        2    0     0.0         0.0
##                                        3  395   100.0       100.0
##
##
## p-values
##                        pApprox       pExact
## gender             1.980647e-17 1.518351e-17
## race               2.132701e-02 1.998835e-02
## EverSmokedCig_P1   9.246878e-04 2.866124e-04
## smoking_status_P1  2.972164e-04           NA
## finalGold_P1      1.429091e-295           NA
## finalGold_P2                NaN           NA
## group              0.000000e+00           NA
##
## Standardize mean differences
##                       average     1 vs 2     1 vs 3    2 vs 3
## gender            0.16362970 0.24570341 0.04255609 0.2026296
## race              0.09344789 0.03486925 0.10526695 0.1402075
## EverSmokedCig_P1  0.08638990 0.11290368 0.01859921 0.1276668
## smoking_status_P1 0.12939939 0.12006442 0.08798054 0.1801532
## finalGold_P1      0.75842156 0.96870512 0.44342841 0.8631312
## finalGold_P2             NaN        NaN        NaN 0.9401010
## group                    NaN        NaN        NaN       NaN
```

```r
sum(is.na(flat_1$Visit_Date_P2)) #3962 subjects did not have a visit 2 date
```

```
## [1] 3962
```

```r
sum(is.na(flat_1$Visit_Date_P1)) #349 subjects do not have a visit 1 date
```

```
## [1] 349
```

```r
sum(is.na(flat_1$FEV1_utah_P2)) #4624 patients don't have fev1 for visit 2
```

```
## [1] 4624
```

```r
sum(is.na(flat_1$FEV1_utah_P1)) #416 patients don't have fev1 for visit 1
```

```
## [1] 416
```

```r
sum(is.na(flat_1$FEV1_FVC_utah_P2)) #4625 patients don't have fev1/fvc for visit 2
```

```
## [1] 4625
```

```r
sum(is.na(flat_1$FEV1_FVC_utah_P1)) #416 patients don't have fev1/fvc for visit 1
```

```
## [1] 416
```

```r
sum(is.na(flat_1$pctEmph_Thirona_P2)) #4995 patients don't have emphysema percentage for visit 2
```

```
## [1] 4995
```

```r
sum(is.na(flat_1$pctEmph_Thirona_P1)) #1072 patients didn't have visit 1 emphysema
```

```
## [1] 1072
```

```r
sum(is.na(flat_1$AWT_seg_Thirona_P2)) #5000 patients don't have AWT measures for visit 2
```

```
## [1] 5000
```

```r
sum(is.na(flat_1$AWT_seg_Thirona_P1)) #1073 patients dont' have AWT measures for visit 1
```

```
## [1] 1073
```

```r
sum(is.na(flat_1$Pi10_Thirona_P2)) #4995 patients don't have Pi10 for visit 2
```

```
## [1] 4995
```

```r
sum(is.na(flat_1$Pi10_Thirona_P1)) #1072 patients don't have Pi10 values for visit 1
```

```
## [1] 1072
```

```r
sum(is.na(flat_1$pctGasTrap_Thirona_P2)) #5505 patients don't have gas trapping for visit 2
```

```
## [1] 5505
```

```r
sum(is.na(flat_1$pctGasTrap_Thirona_P1)) #2435 patients don't have gas trapping for visit 1
```

```
## [1] 2435
```