

# Sequence Analysis

*Morten Dybdahl Krebs*

17/01/2019

## Contents

<b>1 Sequence Analysis in Schizophrenia</b>	<b>1</b>
1.1 Libraries and Data . . . . .	1
1.2 Diagnosis State Sequences: . . . . .	2
1.3 Sequence Dissimilarities: . . . . .	4
1.4 Multidimensional scaling . . . . .	7
1.5 Correlation with gender and year of birth . . . . .	9
1.6 Dimension Characteristics: . . . . .	12
1.7 Clustering . . . . .	18
1.8 Association with putative risk variables . . . . .	20
1.9 Replication . . . . .	28
1.10 Robustness . . . . .	32
1.11 Post-hoc Regressions . . . . .	36
1.12 Browser . . . . .	74

This is a document providing the code for the analyses in the paper “...” by Krebs et al.

## 1 Sequence Analysis in Schizophrenia

### 1.1 Libraries and Data

Load required libraries including

```
library(devtools)
install_github("MortenKrebs/diagtraject")
library(diagtraject)
```

Load data and create sequence state object:

The Study was performed using the iPSYCH cohort (See [Pedersen et al](#):

```
# 3 data.frames with one row pr individual and collums containing id, birthday and date of first diagno

load(diagdates) # individuals diagnosed with Schizophrenia before Dec 31, 2012
load(diagdates_random) # random population sample
load(diagdates2016) # individuals diagnosed with Schizophrenia 2013-2016
```

```

colnames(diagdates)

## [1] "pid"      "birthdate" "F1"       "F3"       "F4"
## [6] "F50"      "F60"       "F70"       "F84"      "F9"
## [11] "SCZ"      "gender"

nrow(diagdates)

## [1] 5432

```

Diagnosis	ICD.10
Substance abuse	F10-F19
Mood disorders*	F30-39
Anxiety disorders and Obsessive compulsive disorder	F40.0-F40.2 F41.0-F41.1
Eating disorders	F50
Personality disorder	F60
Mental Retardation	F70-79
Pervasive developmental disorders	F84
Behavioural and emotional disorders with onset usually occurring in childhood and adolescence	F90-98

\*For recurrent depression onset was defined as the second admission that occurred at least 8 weeks after last discharge with these ICD-8 codes

## 1.2 Diagnosis State Sequences:

```

dt <- data.table(diagdates)

fdate <- function(x) as.Date(x,"%d/%m/%Y" )

dt[,birthdate:= fdate(birthdate) ]
dt[,c("F1","F3", "F4","F50","F60", "F70", "F84", "F9","SCZ") := 
  lapply(list(F1,F3,F4,F50,F60,F70,F84,F9,SCZ),function(x){
    (as.numeric(fdate(x))- as.numeric(birthdate))/365})]
dt[,censored:= (as.numeric(fdate("31/12/2016"))-as.numeric(birthdate))/365] #Calculates age in 31/12/2016

dt.m <-melt(dt, id.vars = 'pid', direction = "long", measure.vars = list(c(3:10,13)), value.name = "time")
dt.m<- dt.m[order(pid)]
dt.m[,time:=time1]
dt.m <- dt.m[!is.na(time)]

events <- levels(dt.m$event)
drop <- matrix(FALSE, nrow = length(events), ncol = length(events), dimnames = list(events,events))
drop['censored',] <- T
drop[, 'censored'] <- T
diag(drop) <- F

e2sm <-seqe2stm(events = events, dropMatrix = drop)

```

Creating sequence object with three different settings:

### 1.2.1 Varying sequences lengths

```
seq_mis <- seqdef(TSE_to_STS(dt.m,
                               id = "pid", timestamp = "time", event= "event",
                               tmin=1, tmax=ceiling(max(dt.m$time)), stm = e2sm)
                           , firstState ="None", missing="censored")
```

### 1.2.2 Plotting most frequent states:

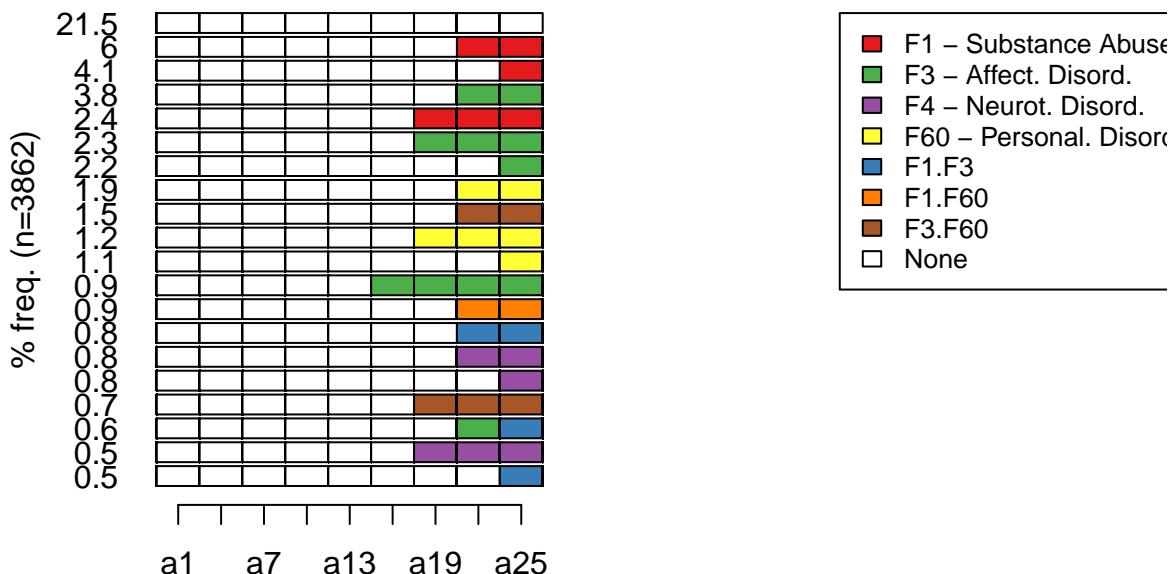
Top-20 most frequent sequences in first 25 years for individuals followed more than 25 years

```
setkey(dt,pid)
seq25 <- seq_mis[seqlength(seq_mis)>=25 & dt[,SCZ] < 25,]
seq25 <- seq25[,seq(1,25,3)]

cols <- brewer.pal(9, "Set1")
tab <- seqtab(seq25,tlim = 1:20)
lab <- attributes(seq25)$labels[which(lapply(attributes(seq25)$labels, function(x) sum(grepl(x, rownames(o<-c(1,4,6,7,2,3,5,8)
lab[o][1:4] <- paste(lab[o][1:4], c("Substance Abuse", "Affect. Disord.", "Neurot. Disord.", "Personal. Disord."))

attributes(seq25)$cpal <- rep("grey", length(attributes(seq25)$labels) )
attributes(seq25)$cpal[which(lapply(attributes(seq25)$labels, function(x) sum(grepl(x, rownames(attributes(seq25)$labels)) > 1))] <- cols[c(1,2,5,3,7,4,6)][o], "white"),cex=.8)

par(mfrow=c(1,2))
seqfplot(seq25,tlim=20:1, pbarw=F, with.legend=F, yaxis="pct")
plot(0,type='n',axes=FALSE,ann=FALSE)
legend( x=0.7,y=1, lab[o] ,fill = c(cols[c(1,2,5,3,7,4,6)][o], "white"),cex=.8)
```



### 1.2.3 “censored” as separate state

For later use a version keeping "censored" as a separate state was created:

```
seq <- seqdef(TSE_to_STS(dt.m, id = "pid", timestamp = "time", event= "event",
                         tmin=1, tmax=ceiling(max(dt.m$time)),
                         stm = e2sm),firstState ="None")

range(seqlength(seq))

## [1] 36 36

length(alphabet(seq))

## [1] 194
```

### 1.2.4 Additional last state

Identifying the state at time of censoring:

```
drop['censored',] <- F
drop[, 'censored'] <- T
e2sm <- seqe2stm(events = events, dropMatrix = drop)

seq_last <- seqdef(
  TSE_to_STS(dt.m, id = "pid", timestamp = "time",
              event= "event", tmin=1,tmax=ceiling(max(dt.m$time))+1,
              stm = e2sm),
  firstState ="None", missing="censored")
seq_mis_last <- seq_mis
seq_mis_last$last <- seq_last[,ncol(seq_last)]
attributes(seq_mis_last)$alphabet <- c(alphabet(seq_mis_last), unique(seq_mis_last$last)[which(!unique
attributes(seq_mis_last)$alphabet <- attributes(seq_mis_last)$alphabet[order(attributes(seq_mis_last)$a
attributes(seq_mis_last)$labels <- attributes(seq_mis_last)$alphabet
```

## 1.3 Sequence Dissimilarities:

### 1.3.1 Transition Rates:

To obtain population-wide estimates of transition probabilities, the random population cohort of 30000 individuals was included (See Pedersen et al for details) and estimates were computed using inverse sampling probability weighting (Bowman II):

```
load(diagdates_random)
load(diagdates2016)

dt_pop <- data.table(rbind(diagdates, diagdates2016, diagdates_random))

dt_pop[,year := format(birthdate, '%Y')]
dt_pop[,year := as.factor(as.numeric(year))]
sample_dist <- dt_pop[pid %in% diagdates_random$pid,.N,c("year", "gender")]
```

Distribution of age and gender in the Danish population publicly available from Statistics Denmark:

```
DK_pop <- data.table("year"=1981:2005,
                      "M"= c(27117*7/12, 27063, 26001, 26572, 27465, 28434, 29079, 30324, 31475, 32620, 33005, 34811,
                            "F"= c(25972*7/12, 25595, 24821, 25228, 26284, 26878, 27142, 28520, 29876, 30813, 31353, 32914)
DK_pop <-melt(DK_pop, id.vars = "year", value.name = "n")
DK_pop <-DK_pop[year<2002]
```

Inverse sampling probability weighting:

```
sample_dist <- sample_dist[order(gender,year)]
sample_dist[,pop_weights := DK_pop[,n]/sample_dist[,N]]
dt_pop <- merge(dt_pop, sample_dist, by=c("year", "gender"))
dt_pop[!is.na(SCZ), pop_weights := 1]
setkey(dt_pop,pid)
```

Formatted as above

```
dt_pop[,"birthdate":= fdate(birthdate) ]
dt_pop[,c("F1", "F3", "F4", "F50", "F60", "F70", "F84", "F9", "SCZ") := 
  lapply(list(F1,F3,F4,F50,F60,F70,F84,F9,SCZ),function(x){
    (as.numeric(fdate(x))- as.numeric(birthdate))/365})]
dt_pop[,censored:= (as.numeric(fdate("31/12/2016"))-as.numeric(birthdate))/365]
dt.m <-melt(dt_pop, id.vars = 'pid', direction = "long", measure.vars = list(c(5:12,16)), value.name =
dt.m<- dt.m[order(pid)]
dt.m[,time:=time1]
dt.m <- dt.m[!is.na(time)]

events <- levels(dt.m$event)
drop <- matrix(FALSE, nrow = length(events), ncol = length(events), dimnames = list(events,events))
drop['censored',] <- T
drop[, 'censored']= T
diag(drop) <- F

e2sm <-seq2stm(events = events, dropMatrix = drop)

seq_pop <- seqdef(TSE_to_STS(dt.m,id = "pid", timestamp = "time", event= "event",
                             tmin=1, tmax=ceiling(max(dt.m$time)),
                             stm = e2sm),firstState ="None")

seq_mis_pop <- seqdef(TSE_to_STS(dt.m,
                                    id = "pid", timestamp = "time", event= "event",
                                    tmin=1, tmax=ceiling(max(dt.m$time)),stm = e2sm)
                           , firstState ="None", missing="censored")

drop['censored',] <- F
drop[, 'censored']= T
e2sm <-seq2stm(events = events, dropMatrix = drop)

seq_last <- seqdef(
  TSE_to_STS(dt.m, id = "pid", timestamp = "time",
             event= "event", tmin=1,tmax=ceiling(max(dt.m$time))+1,
```

```

        stm = e2sm),
firstState ="None", missing="censored")
seq_mis_last_pop <- seq_mis_pop
seq_mis_last_pop$last <- seq_last[,ncol(seq_last)]
attributes(seq_mis_last)$alphabet <- c(alphabet(seq_mis_last), unique(seq_mis_last$last)[which(!unique
attributes(seq_mis_last)$alphabet <- attributes(seq_mis_last)$alphabet[order(attributes(seq_mis_last)$al
attributes(seq_mis_last)$labels <- attributes(seq_mis_last)$alphabet

```

Time-varying Transition Rates computed with TraMineR::seqrate::

```

seq_mis_last_uni_pop <- unique(seq_mis_last_pop)
attributes(seq_mis_last_uni_pop)$weights <- match(seqconc(seq_mis_last_pop),seqconc(seq_mis_last_uni_pop))

tr_t <- seqrate(seq_mis_last_uni_pop, time.varying = T,weighted = T)

```

### 1.3.2 Substitution Costs:

Substitution Cost Matrix, using Jaccard distance:

$$d_J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

```

alp <- seqdef(as.data.frame(alphabet(seq_pop)),stsep=".[]")
rownames(alp) <- alphabet(seq_pop)
sub.cost_jacc <- matrix(0, nrow(alp),nrow(alp))
alp <- alp[order(seqlength(alp)),]
for(i in 1:nrow(alp))
  for(j in 1:nrow(alp))
    if(i>=j){
      sub.cost_jacc[i,j] <- 1-sum(unlist(alp[i,1:seqlength(alp[i,])])) %in% unlist(alp[j,1:seqlength(alp[j,])])
      length(unique(c(unlist(alp[i,1:seqlength(alp[i,])]), unlist(alp[j,1:seqlength(alp[j,])]))))} else
sub.cost_jacc <- as.matrix(as.dist(sub.cost_jacc ))
rownames(sub.cost_jacc) <- colnames(sub.cost_jacc) <- rownames(alp)
sub.cost_jacc <- sub.cost_jacc[order(rownames(sub.cost_jacc)),order(rownames(sub.cost_jacc))]
sub.cost <- seqsubm(seqdata = seq_pop, method = "CONSTANT")
rownames(sub.cost_jacc) <- rownames(sub.cost)
colnames(sub.cost_jacc) <- colnames(sub.cost)
sub.cost_jacc["censored->"] <- sub.cost_jacc[, "censored->"] <- 0

dim(sub.cost_jacc)

## [1] 202 202

kable(sub.cost_jacc[1:5,1:5])

```

	censored->	F1->	F1.F3->	F1.F3.F4->	F1.F3.F4.F50->
censored->	0	0.0000000	0.0000000	0.0000000	0.00
F1->	0	0.0000000	0.5000000	0.6666667	0.75
F1.F3->	0	0.5000000	0.0000000	0.3333333	0.50
F1.F3.F4->	0	0.6666667	0.3333333	0.0000000	0.25
F1.F3.F4.F50->	0	0.7500000	0.5000000	0.2500000	0.00

### 1.3.3 Handling right-censoring:

To calculate dissimilarities between sequences with right censoring, we used inferred states weighted by the probabilities of that state, given the last observed state in the sequence. When calculating the dissimilarity between two sequences  $i$  and  $j$  of unequal length, the dissimilarity,  $D(i, j)$

$$D(i, j) = d_{obs} + d_{inf}$$

where  $d_{obs}(i, j)$  is the dissimilarity between the sequences before right censoring occurs and  $d_{inf}$  is the sum of the substitution cost matrix weighted by the inferred probabilities.

Dissimilarities for observed states using optimal matching (OM):

```
sms <- which(rownames(sub.cost_jacc) %in% rownames(seqsubm(seq, "CONSTANT")))
d_OM <- seqdist(seq, method = "OM", indel=.5,
                 sm = sub.cost_jacc[sms,sms])
```

### 1.3.4 Imputation:

$$d_{inf}(i, j) = \sum_{t=1}^{t_{max}} Pr(i)_t Pr(j)_t^T \circ SC$$

And for imputed states using diagtraject::mis.cost() :

```
ms <- mis.cost(seq_mis, last, tr_t, sm = sub.cost_jacc,
                cens.type = "right", imp.length = "max",
                diag=F, sum_to_1=T, resol.comp = resol.comp,
                resol.ratio = resol.reduc, mc.cores=27)
```

### 1.3.5 Overall dissimilarities

Obtained by  $d_{obs} + d_{inf}$

```
dist_OM <- d_OM + as.matrix(ms$dist)
```

## 1.4 Multidimensional scaling

Metric multidimensional scaling using vegan::wcmdscale() :

```
mcor <- match(seqconc(seq_mis_last), seqconc(unique(seq_mis_last))) # finding unique sequences
uni <- (!duplicated(mcor))
dist <- list("dist"=dist_OM[uni,uni], "weight"= table(mcor), "mcor"= mcор)
```

```
library(vegan)
library(ggplot2)
wmd <- mclapply(1:13, function(x) wcmdscale(dist$dist, w = table(dist$mcор), eig = T, k=x), mc.cores=12)
```

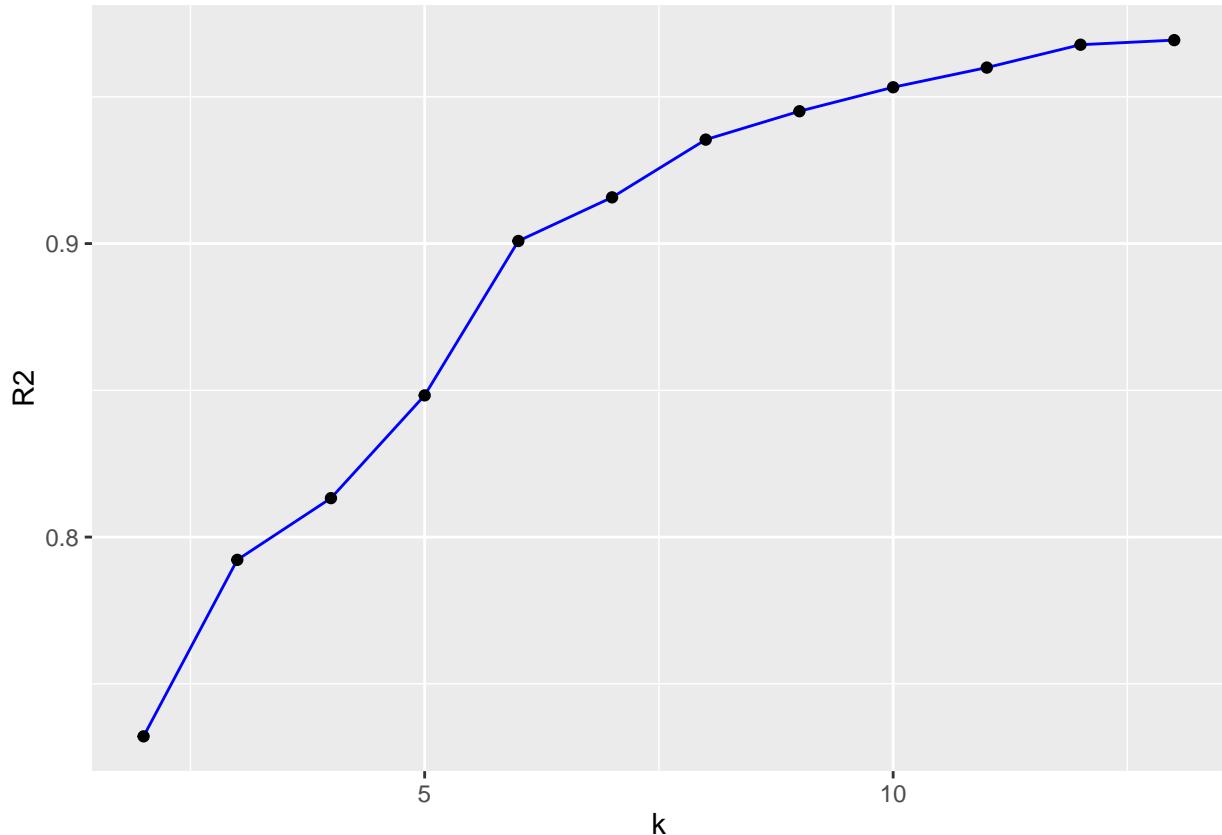
Goodness of fit for k=2:13:

```

R2 <- lapply(wmd,function(x) {
NewDists <- dist(x$points[mcor,], diag=TRUE, upper=TRUE)
r <- cor(c(as.dist(dist$dist[mcor, mcor], diag=TRUE, upper=TRUE)), c(NewDists))
r^2})

R2 <- data.frame(k=2:13,R2=do.call('c', R2) )
ggplot(data=R2, aes(x=k,y=R2)) + geom_line(color="blue") + geom_point()

```

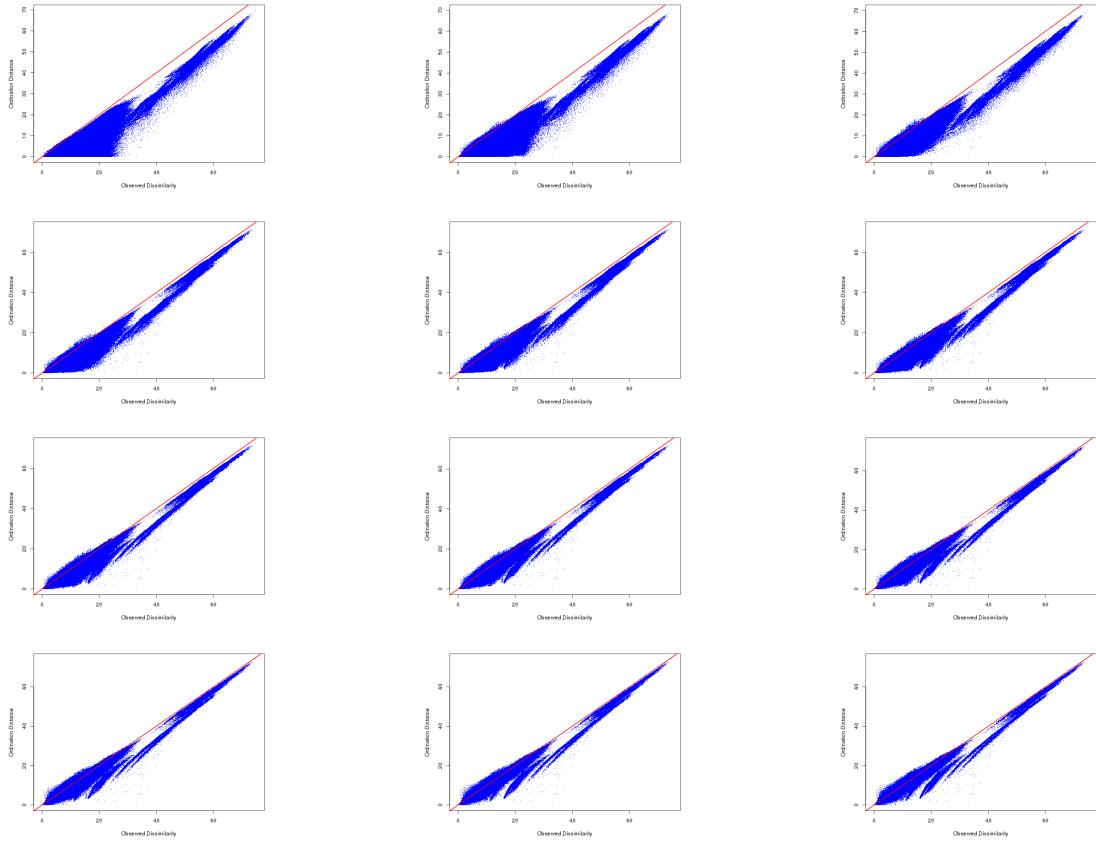


Stressplots for k=2:13:

```

wmd_all <-wcmdscale(dist_unique_0M$dist[1:100, 1:100],w = table(mcor)[1:100], eig = T)
par(mfrow=c(3,4))
for(i in 2:13)
stressplot(wmd_all, k=i, p.col="blue", l.col="red", lwd=2)

```



Computing bootstrap stability of MDS over 100 permutations using `diagtraject::bootmds()`:

```
bootmds(dist_unique_0M$dist,k = 7, nrep = 100, w = table(mcov))
```

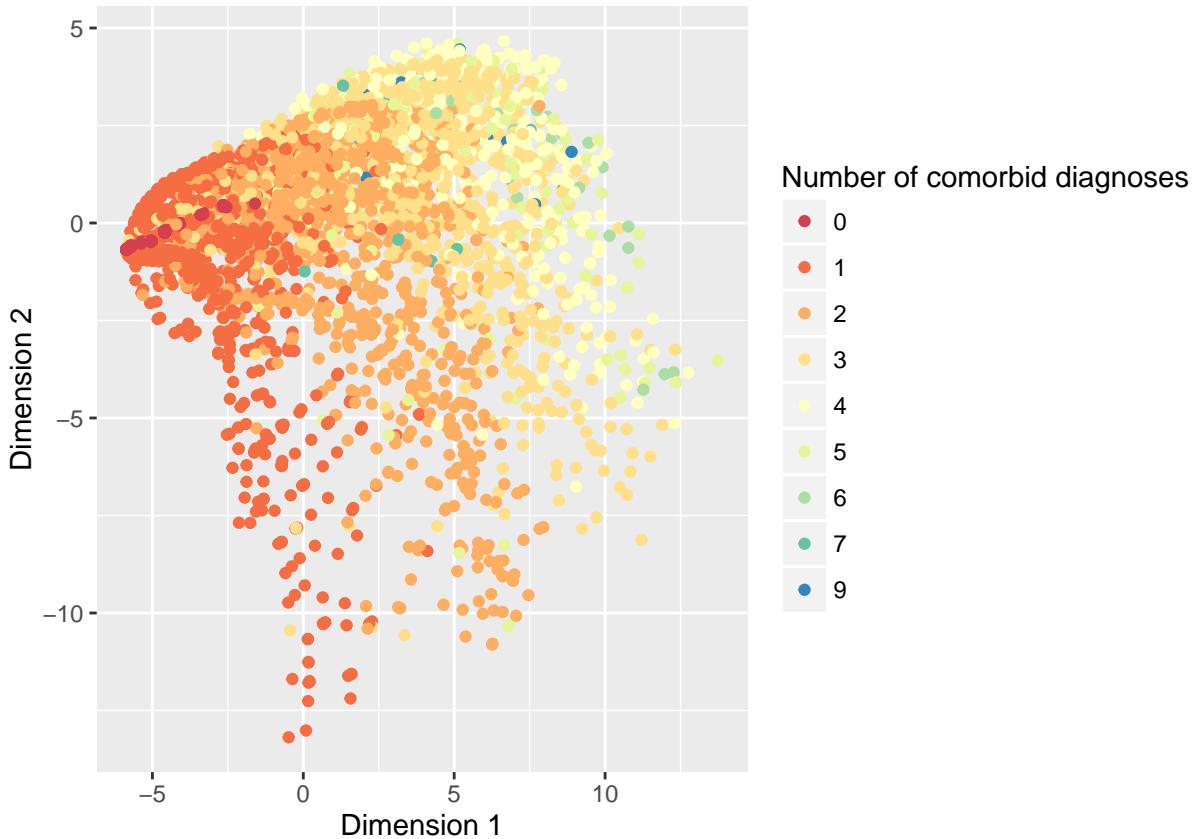
```
## [1] 0.9988969
```

## 1.5 Correlation with gender and year of birth

```
dt_mds <- cbind(dt, wmd[[6]]$points[mcov,])
dt_mds[,n_dia := dt.m[,N, pid]$N-1]
```

Dimension 1+2 and number of diagnoses:

```
ggplot(dt_mds, aes(x=Dim1, y=Dim2, color=factor(n_dia)))+ geom_jitter(h=.05,w=.05)+ scale_colour_brewer(palette = "Spectral") + labs(x="Dimension 1", y="Dimension 2",color= "Number of dia")
```



Dimension 1 and year of birth

```

first_dia <- melt(dt_mds, id.vars = 'pid', direction = "long", measure.vars =  list(c(3:12)),value.name
setkey(first_dia,pid)
setkey(dt_mds,pid)
dt_mds[,age_min := first_dia[,min(time1,na.rm=T),pid]$V1]

dt_mds[,dia_year:=as.numeric(format(SCZ*365+birthdate,"%Y"))]

newdata=data.frame(birthdate=dt_mds[,birthdate],
                    age_min=rep(15,nrow(dt_mds)),
                    gender=rep("M",nrow(dt_mds)))

lm <- lm(Dim1~birthdate, data=dt_mds)
conf_interval <- predict(lm, newdata=newdata, interval="confidence",level = 0.95)
dt_mds[,fit:=conf_interval[,1]]
dt_mds[,lwr:=conf_interval[,2]]
dt_mds[,upr:=conf_interval[,3]]

lm <- lm(Dim1~age_min+gender+birthdate, data=dt_mds)
conf_interval <- predict(lm, newdata=newdata, interval="confidence", level = 0.95)
dt_mds[,fit1:=conf_interval[,1]]
dt_mds[,lwr1:=conf_interval[,2]]
dt_mds[,upr1:=conf_interval[,3]]

p<- ggplot(dt_mds, aes(y=Dim1, x=birthdate))+
  geom_line(aes(x =birthdate, y=fit ),color="red")+
  geom_line(aes(x =birthdate, y=upr ),color="red", linetype="dashed")+

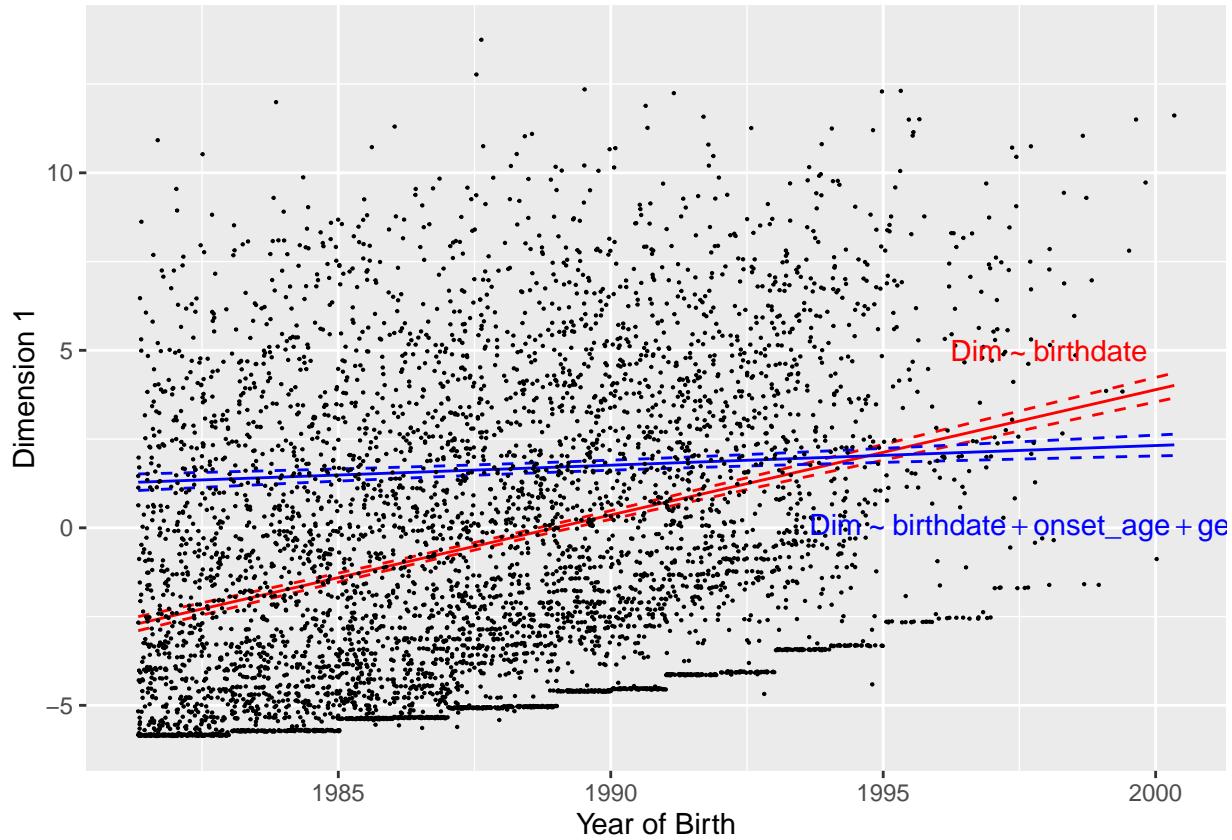
```

```

geom_line(aes(x =birthdate, y=lwr ),color="red", linetype="dashed")+
geom_line(aes(x =birthdate, y=fit1 ),color="blue")+
geom_line(aes(x =birthdate, y=upr1 ),color="blue", linetype="dashed")+
geom_line(aes(x =birthdate, y=lwr1 ),color="blue", linetype="dashed")+
geom_jitter(h=.05,w=.5,size=.1)+labs(y="Dimension 1", x="Year of Birth")+
annotate("text",x=as.Date("1998",format = "%Y"), y=0, parse=T,color= "blue",label =
as.character(expression(Dim%~%birthdate+onset_age+gender)))+
annotate("text",x=as.Date("1998",format = "%Y"), y=5, parse=T,color= "red",label =
as.character(expression(Dim%~%birthdate)))

```

p

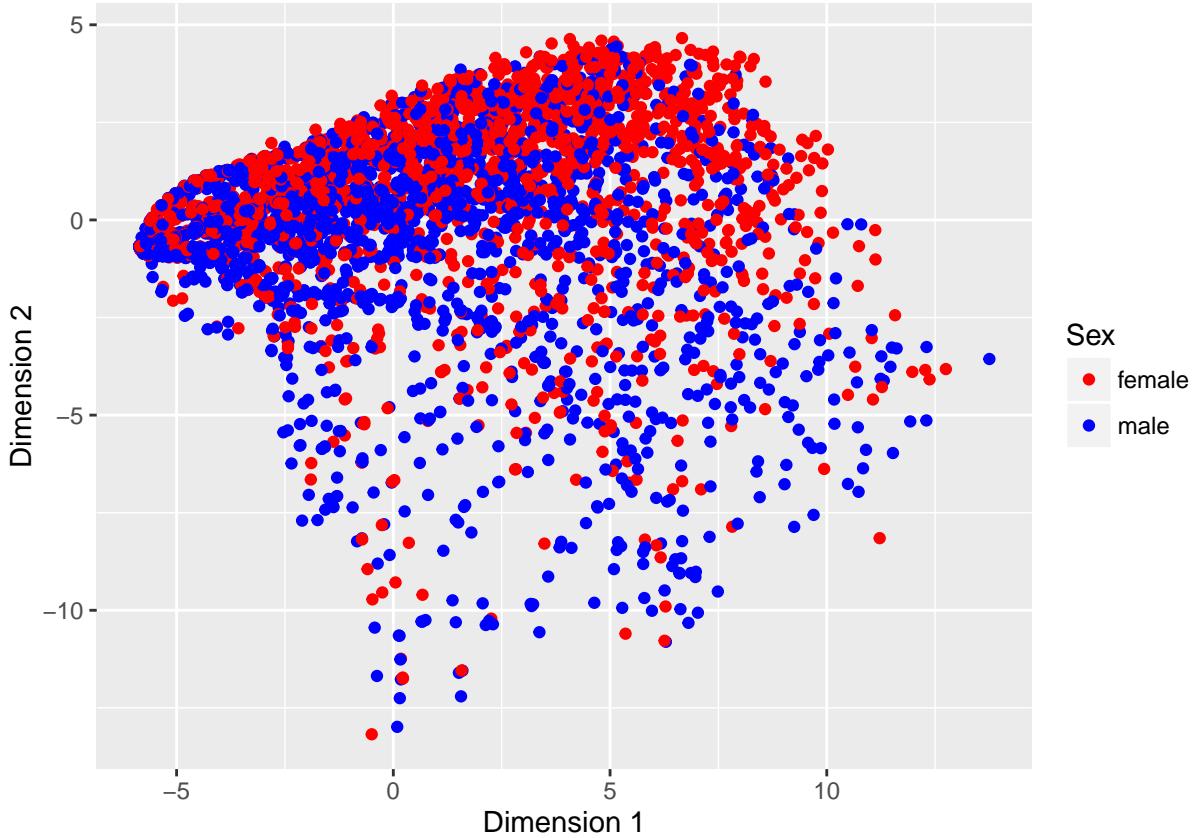


Dimension 1 and 2 ~ gender

```

ggplot(dt_mds, aes(x=Dim1, y=Dim2, color=factor(gender,labels = c("female","male"))))+ geom_jitter(h=.
scale_colour_manual(values = c("red","blue"))+labs(x="Dimension 1", y="Dimension 2",color= "Sex")

```



## 1.6 Dimension Characteristics:

Characteristics for Dimension 1-7:

```
dt_mds[,paste0("Dim",1:7,"r")] := lapply(list(Dim1,Dim2,Dim3,Dim4,Dim5,Dim6,Dim7), rank,ties.method="random")
dt_mds[,paste0("Dim",1:7,"f")] := lapply(list(Dim1r,Dim2r,Dim3r,Dim4r,Dim5r,Dim6r,Dim7r), function(x) cut(x, 6))
dim <- paste0("Dim",1:7,"f")

p_10 <- lapply(dim, function(x) {
  x2 <- eval(noquote(paste0(x)))
  t1 <- dt_mds
  t1[, ':=` (count=.N), by=x2]
  t1[, "xv":=t1[,x2, with=F]]
  t1[, cut:=cut(F10,breaks = seq(0,36,6))]
  t1 <- t1[,.sum(!is.na(F10)),mean(count)), by=.(xv,cut)]
  t1[, V3:=V1/V2*100]
  setkey(t1,xv)
  setkey(t1,cut)
  ggplot(data=t1, aes(x=xv, y=V3, fill=cut)) +
    geom_bar(stat="identity")+
    scale_fill_brewer(palette = "YlGn")+
    scale_y_continuous(limits =c(0,100))+
```

```

    axis.ticks.x=element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    axis.line = element_line(colour = "black"))
})

p_30 <- lapply(dim, function(x) {
  x2 <- eval(noquote(paste0(x)))
  t1 <- dt_mds
  t1[, ' :=' (count=.N), by=x2]
  t1[, "xv":=t1[,x2, with=F]]
  t1[, cut:=cut(F30,breaks = seq(0,36,6))]
  t1 <- t1[,(sum(!is.na(F30)),mean(count)), by=.(xv,cut)]
  t1[, V3:=V1/V2*100]
  setkey(t1,cut)
  ggplot(data=t1, aes(x=xv, y=V3, fill=cut)) +
    geom_bar(stat="identity")+
    scale_fill_brewer(palette = "YlGn")+
    scale_y_continuous(limits =c(0,100))+ 
    labs(title="Aff", x= sub("_f","",x),y="%")+
    theme(legend.position="none",
          axis.text.x=element_blank(),
          axis.ticks.x=element_blank(),
          panel.grid.major = element_blank(),
          panel.grid.minor = element_blank(),
          panel.background = element_blank(),
          axis.line = element_line(colour = "black"))

})

p_40 <- lapply(dim, function(x) {
  x2 <- eval(noquote(paste0(x)))
  t1 <- dt_mds
  t1[, ' :=' (count=.N), by=x2]
  t1[, "xv":=t1[,x2, with=F]]
  t1[, cut:=cut(F40,breaks = seq(0,36,6))]
  t1 <- t1[,(sum(!is.na(F40)),mean(count)), by=.(xv,cut)]
  t1[, V3:=V1/V2*100]
  setkey(t1,cut)
  ggplot(data=t1, aes(x=xv, y=V3, fill=cut)) +
    geom_bar(stat="identity")+
    scale_fill_brewer(palette = "YlGn")+
    scale_y_continuous(limits =c(0,100))+ 
    labs(title="Neuro", x= sub("_f","",x),y="%")+
    theme(legend.position="none",
          axis.text.x=element_blank(),
          axis.ticks.x=element_blank(),
          panel.grid.major = element_blank(),
          panel.grid.minor = element_blank(),
          panel.background = element_blank(),
          axis.line = element_line(colour = "black"))
})

```

```

})

p_50 <- lapply(dim, function(x) {
  x2 <- eval(noquote(paste0(x)))
  t1 <- dt_mds
  t1[, ':=` (count=.N), by=x2]
  t1[, "xv":=t1[,x2, with=F]]
  t1[, cut:=cut(F50, breaks = seq(0,36,6))]
  t1 <- t1[,.sum(!is.na(F50)),mean(count)), by=.(xv,cut)]
  t1[, V3:=V1/V2*100]
  setkey(t1,cut)
  ggplot(data=t1, aes(x=xv, y=V3, fill=cut)) +
    geom_bar(stat="identity")+
    scale_fill_brewer(palette = "YlGn")+
    scale_y_continuous(limits =c(0,100))+ 
    labs(title="ED", x= sub("_f","",x),y="%")+
    theme(legend.position="none",
          axis.text.x=element_blank(),
          axis.ticks.x=element_blank(),
          panel.grid.major = element_blank(),
          panel.grid.minor = element_blank(),
          panel.background = element_blank(),
          axis.line = element_line(colour = "black")))
})

p_60 <- lapply(dim, function(x) {
  x2 <- eval(noquote(paste0(x)))
  t1 <- dt_mds
  t1[, ':=` (count=.N), by=x2]
  t1[, "xv":=t1[,x2, with=F]]
  t1[, cut:=cut(F60, breaks = seq(0,36,6))]
  t1 <- t1[,.sum(!is.na(F60)),mean(count)), by=.(xv,cut)]
  t1[, V3:=V1/V2*100]
  setkey(t1,cut)
  ggplot(data=t1, aes(x=xv, y=V3, fill=cut)) +
    geom_bar(stat="identity")+
    scale_fill_brewer(palette = "YlGn")+
    scale_y_continuous(limits =c(0,100))+ 
    labs(title="PD", x= sub("_f","",x),y="%")+
    theme(legend.position="none",
          axis.text.x=element_blank(),
          axis.ticks.x=element_blank(),
          panel.grid.major = element_blank(),
          panel.grid.minor = element_blank(),
          panel.background = element_blank(),
          axis.line = element_line(colour = "black")))
})

p_70 <- lapply(dim, function(x) {
  x2 <- eval(noquote(paste0(x)))
  t1 <- dt_mds
  t1[, ':=` (count=.N), by=x2]

```

```

t1[, "xv":=t1[,x2, with=F]]
t1[, cut:=cut(F70,breaks = seq(0,36,6))]
t1 <- t1[,(sum(!is.na(F70)),mean(count)), by=.(xv,cut)]
t1[, V3:=V1/V2*100]
setkey(t1,cut)
ggplot(data=t1, aes(x=xv, y=V3, fill=cut)) +
  geom_bar(stat="identity")+
  scale_fill_brewer(palette = "YlGn")+
  scale_y_continuous(limits =c(0,100))+ 
  labs(title="MR", x= sub("_f","",x),y="%")+
  theme(legend.position="none",
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black"))

})

p_84 <- lapply(dim, function(x) {
  x2 <- eval(noquote(paste0(x)))
  t1 <- dt_mds
  t1[, '':=' (count=N), by=x2]
  t1[, "xv":=t1[,x2, with=F]]
  t1[, cut:=cut(F84,breaks = seq(0,36,6))]
  t1 <- t1[,(sum(!is.na(F84)),mean(count)), by=.(xv,cut)]
  t1[, V3:=V1/V2*100]
  setkey(t1,cut)
  ggplot(data=t1, aes(x=xv, y=V3, fill=cut)) +
    geom_bar(stat="identity")+
    scale_fill_brewer(palette = "YlGn")+
    scale_y_continuous(limits =c(0,100))+ 
    labs(title="ASD", x= sub("_f","",x),y="%")+
    theme(legend.position="none",
          axis.text.x=element_blank(),
          axis.ticks.x=element_blank(),
          panel.grid.major = element_blank(),
          panel.grid.minor = element_blank(),
          panel.background = element_blank(),
          axis.line = element_line(colour = "black"))

})

p_90 <- lapply(dim, function(x) {
  x2 <- eval(noquote(paste0(x)))
  t1 <- dt_mds
  t1[, '':=' (count=N), by=x2]
  t1[, "xv":=t1[,x2, with=F]]
  t1[, cut:=cut(F90,breaks = seq(0,36,6))]
  t1 <- t1[,(sum(!is.na(F90)),mean(count)), by=.(xv,cut)]
  t1[, V3:=V1/V2*100]
  setkey(t1,cut)
}

```

```

ggplot(data=t1, aes(x=xv, y=V3, fill=cut)) +
  geom_bar(stat="identity")+
  scale_fill_brewer(palette = "YlGn")+
  scale_y_continuous(limits =c(0,100))+ 
  labs(title="ChD", x= sub("_f","",x),y ="%", fill= "Age at onset in years")+
  theme(#legend.position="none",
    legend.direction="vertical",
    #legend.title=element_text(),
    axis.text.x=element_blank(),
    axis.ticks.x=element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.background = element_blank(),
    axis.line = element_line(colour = "black")) +
  guides(fill=guide_legend(ncol=2))
})

p_n_dia <- lapply(dim, function(x) {
  x2 <- eval(noquote(paste0(x)))
  t1 <- dt_mds
  t1[,count := .N, by=x2]
  t1[, "xv":=t1[,x2, with=F]]
  t1[, cut:=factor(n_dia,levels=(c(1:7,0)))]
  t1 <- t1[,.length(n_dia),mean(count)), by=.(xv,cut)]
  t1[, V3:=V1/V2*100]
#  t1 <- t1[, n_dia, x2]
#  t1[, V3:=xv]
  setkey(t1,cut)
  ggplot(data=t1, aes(x=xv, y=V3,fill=cut)) +
    geom_bar(stat="identity")+
    scale_fill_manual(values=c(brewer.pal(7,"BuPu"),"white"))+
    scale_y_continuous(limits =c(0,101))+ 
    labs(title="", x= sub("_f","",x),y ="%", fill="Number of diagnoses")+
    theme(legend.direction="vertical",
      #legend.title=element_blank(),
      axis.text.x=element_blank(),
      axis.ticks.x=element_blank(),
      panel.grid.major = element_blank(),
      panel.grid.minor = element_blank(),
      panel.background = element_blank(),
      axis.line = element_line(colour = "black"))+
    guides(fill=guide_legend(ncol=2)) })

g_legend<-function(a.gplot){
  tmp <- ggplot_gtable(ggplot_build(a.gplot))
  leg <- which(sapply(tmp$grobs, function(x) x$name) == "guide-box")
  legend <- tmp$grobs[[leg]]
  legend
}

leg <- g_legend(p_n_dia[[1]])
leg2 <- g_legend(p_90[[1]])
p_n <- lapply(p_n_dia, function(x) x +theme(legend.position="none"))

```

```

p_90 <- lapply(p_90, function(x) x +theme(legend.position="none"))

library(grid)
library(gridExtra)

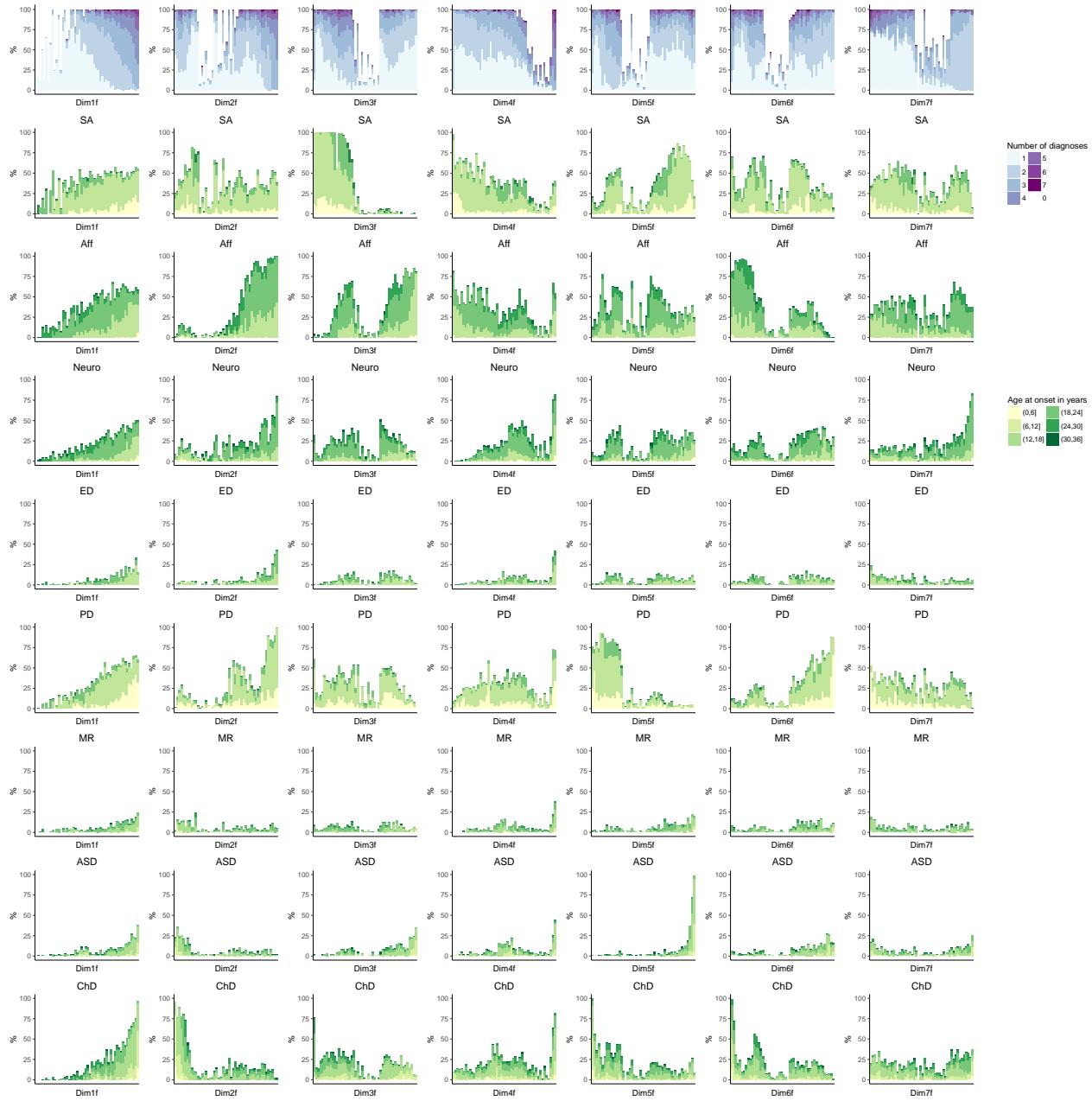
emp <-textGrob("")

# all plots for SM:
plist_all<- c(p_n,list(emp),p_10,list(leg),p_30,list(emp),p_40,list(leg2),p_50,list(emp),p_60,list(emp))
#Figure1
plist1<- list(p_n[[1]], leg,leg2, p_30[[1]],p_60[[1]],p_90[[1]])
#Figure2
plist2<- list(p_n[[2]], leg,leg2,p_30[[2]],p_84[[2]],p_90[[2]])

#Figure3
plist3<- list(p_n[[3]], leg,leg2,p_10[[3]],p_30[[3]],p_84[[3]])

#All
do.call("grid.arrange", c(plist_all, ncol=8))

```



## 1.7 Clustering

Hierarchical clustering based on Sequence Analysis:

```
library(WeightedCluster)

c <- dist_unique_OM$mcov[which(dt[, !is.na(case2012)))]
clust <- hclust(as.dist(dist_unique_OM$dist[c,c]), method="ward.D2")
dt3 <- dt[!is.na(case2012)]
dt3[,cl2:= factor(cutree(clust,2))]
dt3[,cl3:= factor(cutree(clust,3))]
dt3[,cl4:= factor(cutree(clust,4))]
```

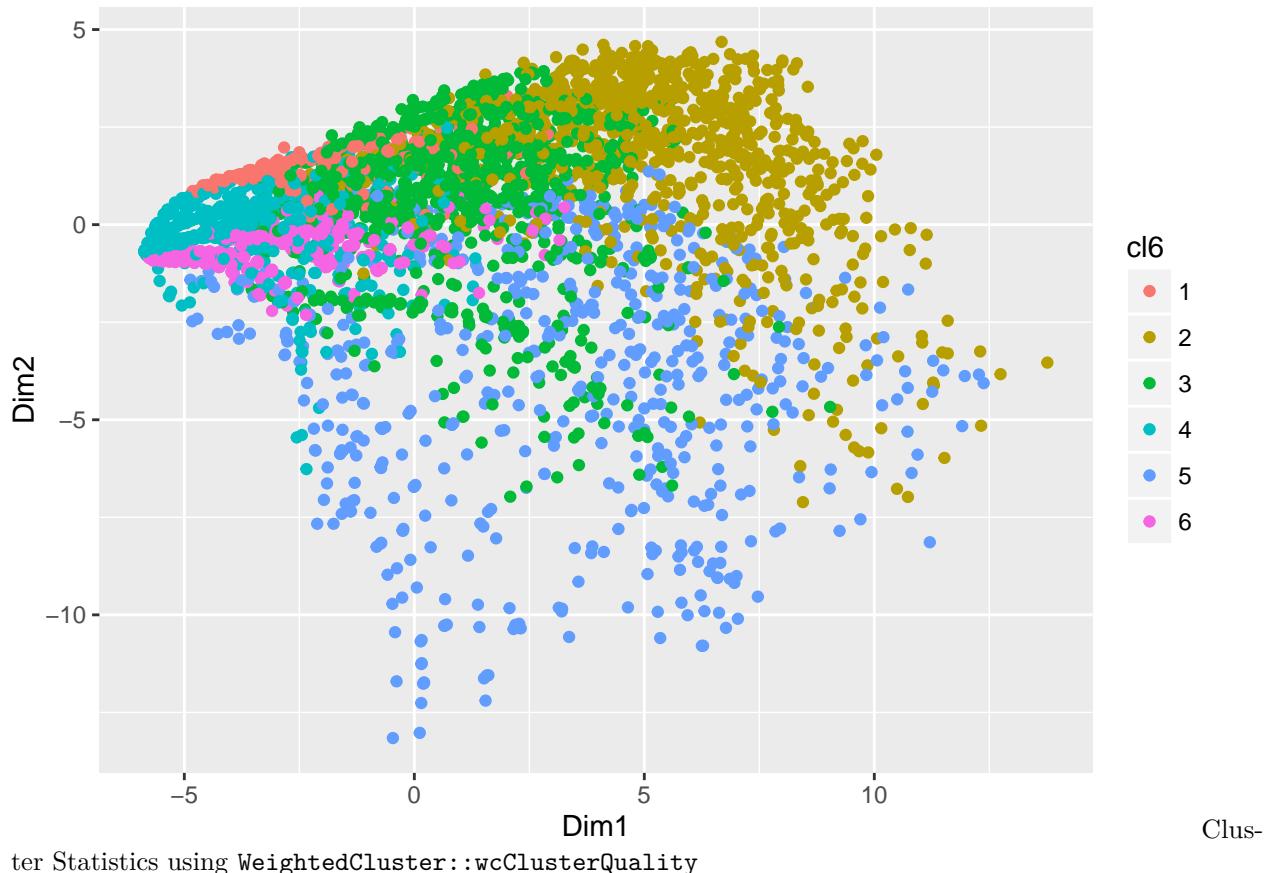
```

dt3[,cl5:= factor(cutree(clust,5))]
dt3[,cl6:= factor(cutree(clust,6))]

dt3 <- merge(dt_mds, dt3[,(pid,cl2,cl3,cl4,cl5,cl6)],by="pid")

ggplot(dt3, aes(x=Dim1, y=Dim2, color=cl6))+ geom_jitter(h=.05,w=.05)

```



```
wcClusterQuality(as.dist(dist_unique_0M$dist[,c,c]),dt3[,cl6])
```

```

## $stats
##          PBC          HG          HGSD          ASW          ASWw          CH
## 0.4306869 0.5881585 0.5881552 0.1705803 0.1714198 415.2389361
##          R2          CHsq          R2sq          HC
## 0.2767450 795.0543851 0.4228441 0.1818011
##
## $ASW
##          ASW          ASWw
## 1 0.44955449 0.45088647
## 2 0.06632928 0.06730067
## 3 0.02841912 0.02917935
## 4 0.26306096 0.26351166
## 5 -0.05018844 -0.04872292
## 6 0.35150270 0.35238364

```

Assesment of clusterstability in bootstrap test using 100 permutations of data and computing the mean

jaccard index using a `fpc::clusterboot`:

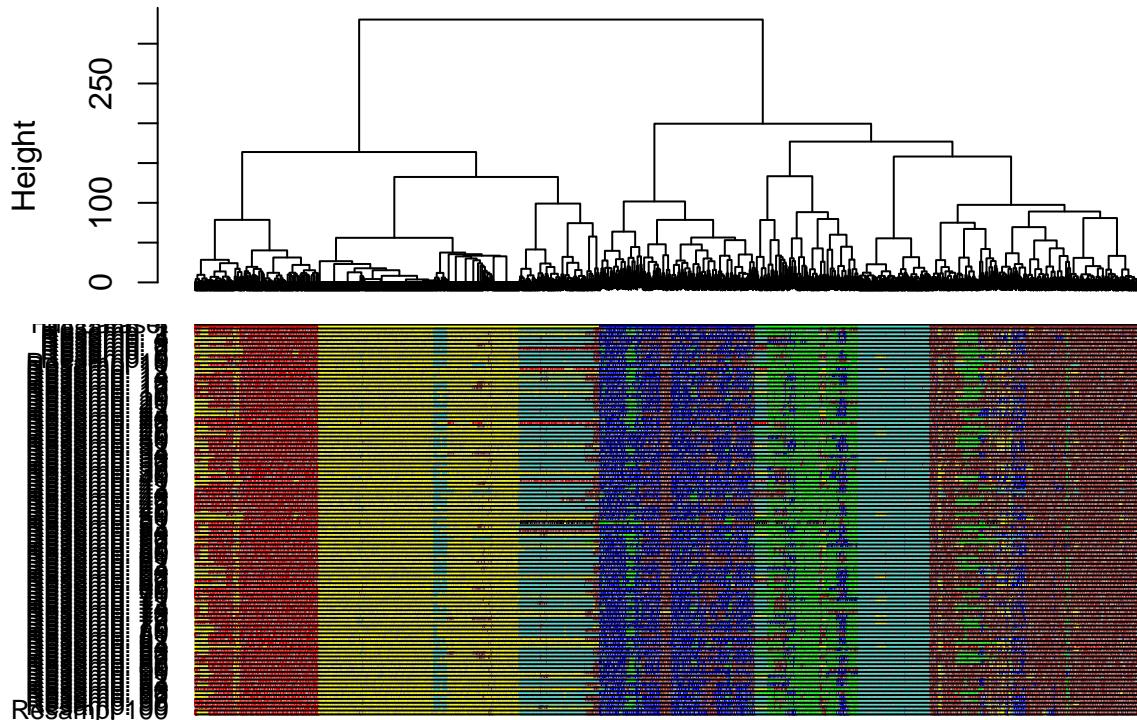
```
library(fpc)
boot<-clusterbootw(as.dist(dist_unique_0M$dist[c,c]),clustermethod=disthclustCBI,k= 6,B=100,method="ward")
boot$bootmean

## [1] 0.5307359 0.5876539 0.6132175 0.6501216 0.5281436 0.8256241
```

Visualizing stability using WGCNA:

```
library(WGCNA)
labels <- matrix(NA, ncol=length(boot$partition), nrow=101)
labels[1,] <- boot$partition
for(i in 2:101) labels[i,] <- matchLabels(boot$bootpartition[, (i-1)], boot$partition)

colors <- labels2colors(labels)
plotDendroAndColors(boot$result$result,t(colors), main="",
                     c("Full dataset", paste("Resamp.", c(1:(dim(labels)[2]-1)))),
                     dendroLabels = F, hang=.03,autoColorHeight = T)
```



## 1.8 Association with putative risk variables

Load additional genetic and registry variables

```
load("additional.Rda")
```

### 1.8.1 Polygenic Scores:

Table 3: Summary Statistics used for Polygenic Score Calculations

Phenotype	n.cases	n.controls	PMID
Anorexia Nervosa	3495	10982	28494655
Anxiety	7016	14745	26754954
Bipolar Affective Disorder	9412	137760	173062
Birth Weight	153781	NA	27680694
Body mass index	339224	NA	25673413
Cannabis use, Lifetime	32330	NA	27023175
Depressive Symptoms	161460	NA	27089181
Education, Years	293723	NA	27225129
Extraversion	160713	NA	24828478
Neuroticism	170911	NA	27089181
Schizophrenia	36989	113075	25056061
Subjective Well-being	298420	NA	27089181

n cases indicate number of cases in the discovery samples

Polygenic Scores were computed using ...

Association with schizophrenia is based on a binomial logistic regression of 2861 cases and 18843 controls - adjusting for age, gender, 10 principal components of genetic similarity and genotype wave. Pseudo  $R^2$  is calculated using Nagelkerke formula (`rcompanion::nagelkerkes`) - comparing the full model to the model without the PGS. p-values printed in **bold** indicate significance after correction for multiple testing (p <0.05/12).

```
source("script/Nagelkerke.R")

dim(additional)
dt_pg <- merge(additional,dt,by = "pid")

dt_pg[!is.na(C1), .N, case]

pval <- lapply(12:23, function(x) {
  tmp = dt_pg
  tmp$y = tmp[,x,with=F]
  summary(glm(!is.na(case)~y+gender+C1+C2+C3+C4+C5+C6+C7+C8+C9+C10+wave, data=tmp, family="binomial"))

null <- glm(!is.na(case)~birthdate+gender+C1+C2+C3+C4+C5+C6+C7+C8+C9+C10+wave, data=dt_pg, family="binomial")

nag <- lapply(12:23, function(x) {
  tmp = dt_pg
  tmp$y = tmp[,x,with=F]
  nagelkerke(fit = glm(!is.na(case)~y+birthdate+gender+C1+C2+C3+C4+C5+C6+C7+C8+C9+C10+wave, data=tmp, family="binomial"))

options(digits=3, scipen=T)
kable(data.frame(Phenotype= name, p= unlist(pval), R2= unlist(nag)), digits=c(1,32,4))
```

Phenotype	p	R2
Anorexia Nervosa	2.11e-02	0.0005
Anxiety	5.39e-01	0.0000
Bipolar Affective Disorder	3.34e-02	0.0004

Phenotype	p	R2
Birth Weight	7.61e-04	0.0011
Body mass index	4.29e-01	0.0001
Cannabis use, Lifetime	3.69e-03	0.0008
Depressive Symptoms	2.40e-12	0.0046
Education, Years	7.66e-05	0.0014
Extraversion	6.89e-03	0.0007
Neuroticism	1.80e-12	0.0046
Schizophrenia	2.24e-29	0.0116
Subjective Well-being	4.38e-11	0.0040

### 1.8.2 Rare Variants:

Computed by Andrea Ganna ...

### 1.8.3 Merge MDS with risk variables:

```
data <- merge(dt_mds[,1:33,with=F],additional,by="pid")
dim(data)

## [1] 5432   91

ind_birth <- c(34:41,86:87)
ind_mat_infek <- 42:43
ind_prs <- (44:55)[pval<.05/12]
ind_pc_wave <- 56:66
ind_sev <- 67:73
ind_infek <- 74:77
ind_rare <- 80:81
ind_parental <- 88:91
```

### 1.8.4 Registry Variables:

Recode variables {#Recode}

```
table(data$apgar5)[table(data$apgar5)>4]
##
##    0    10     5     6     7     8     9     A
##   15 4978    10    18    34    85   221    51
data$apgar5=factor(data$apgar5,levels=c("0","1","2","3","4","5","6","7","8","9","10","A"))
data$apgar5=as.character(data$apgar5)
data$apgar5[data$apgar5=="A"]=NA
data$apgar5=as.numeric(data$apgar5)
table(data$apgar5)[table(data$apgar5)>4]
##
##    0    5     6     7     8     9    10
##   15    10    18    34    85   221  4978
```

```

names(table(data$langde))
## [1] "1"  "10" "11" "15" "16" "18" "20" "21" "22" "23" "24" "25" "26" "27"
## [15] "28" "29" "30" "31" "32" "33" "34" "35" "36" "37" "38" "39" "40" "41"
## [29] "42" "43" "44" "45" "46" "47" "48" "49" "5"   "50" "51" "52" "53" "54"
## [43] "55" "56" "57" "58" "59" "60" "61" "62" "63" "64" "65" "66" "67" "70"
## [57] "74" "80" "9"  "90" "A"
langde=data$langde
langde[langde=="A"]=NA
langde[langde=="1"]=NA
langde[langde=="10"]=NA
langde=as.numeric(as.character(langde))
data$langde=langde;rm(langde)

table(data$B_RYGER)
##
##    0    1
## 565 468
names(table(data$C_RYGER))
## [1] "0"  "10" "20" "21" "22" "99"
data$C_RYGER[data$C_RYGER==99]=NA
table(data$C_RYGER)[table(data$C_RYGER)>4]
##
##    0 21 22
## 35 7 10
data$RYGER <- data$B_RYGER
data$RYGER[!is.na(data$C_RYGER)] <-1

data$wave[is.na(data$wave)]="Other"
data <- data[,wave:=factor(wave)]

table(data$apgar5)
##
##    0    1    2    3    4    5    6    7    8    9    10
## 15    2    1    2    4    10   18   34   85   221  4978
table(data$langde)[table(data$langde)>4]
##
##    36   37   38   39   41   42   43   44   45   46   47   48   49   50   51   52   53   54
##    5    5    7    5   11   14   14   23   35   66  122  252  397  807  799  905  752  540
##    55   56   57   58   59
## 298 154  62  27    5
table(data$RYGER)
##
##    0    1
## 565 525
ind_birth <- c(ind_birth[-c(7:8)],92); names(data)[ind_birth]
## [1] "maternal_age" "paternal_age" "langde"      "apgar5"
## [5] "gest_age"     "bw_score"     "B_SECTIOU"   "B_I11"
## [9] "RYGER"

nms = grep("d2100",names(data))
nms <- names(data)[nms];nms
## [1] "d2100_ptype0_contacts" "d2100_ptype0_days"       "d2100_ptype1_contacts"
## [4] "d2100_ptype1_visits"    "d2100_ptype2_visits"    "d2100_ptype2_novisits"

```

```

## [7] "d2100_ptype3_contacts"
data <- data[,contacts:= Reduce(`+`, .SD), .SDcols=nms[c(1,3,7)]]
data <- data[,visits:= Reduce(`+`, .SD), .SDcols=nms[4:5]]
ind_sev <- 93:94; names(data)[ind_sev]
## [1] "contacts" "visits"






```

### 1.8.5 Standardize PRS scores

```

prs = names(data)[ind_prs]
data[!is.na(C1),c(prs):= lapply(.SD,scale),.SDcols=prs]

```

### 1.8.6 Association Analysis

Wrapper functions for `stats::manova` and `stats::anova`:

```
mancova.func <- function(tmp,ind_x,ind_y,mf){  
tmp<- as.data.frame(tmp)  
p.val <- Fstat <- rep(NA,length(ind_x))  
for(p in (1:length(ind_x))){  
tmp$x=tmp[,ind_x[p]]  
tmp$y = as.matrix(tmp[,ind_y])  
manova.y=manova(mf, data = tmp[!is.na(tmp$x),]);  
p.val[p] = summary(manova.y)$stats[1,6]  
Fstat[p] = summary(manova.y)$stats[1,3]  
}  
names(p.val)=names(tmp)[ind_x]  
names(Fstat)=names(tmp)[ind_x]  
list(p.val=p.val,Fstat=Fstat) }  
  
ancova.func <- function(tmp,ind_x,ind_y,mf){  
tmp<- as.data.frame(tmp)  
p.val <- betas <- array(NA,dim=c(length(ind_y),length(ind_x)))  
for(p in (1:length(ind_x))){  
for(q in (1:length(ind_y))){  
tmp$x=tmp[,ind_x[p]]  
tmp$y = as.matrix(tmp[,ind_y[q]])  
lm = lm(mf, data = tmp[!is.na(tmp$x),])  
anova.y=anova(lm)  
p.val[q,p] = anova.y[1,5]  
betas[q,p] = coef(lm)[2]}  
colnames(p.val)=names(tmp)[ind_x]  
colnames(betas)=names(tmp)[ind_x]  
list(p.val=p.val,beta=betas) }
```

Number of tests in genetic risk analysis:

```
n_test_gen <- length(c(ind_prs, ind_rare));n_test_gen
```

```
## [1] 9
```

Number of tests in registry variable analysis:

```
n_test_regist <- length(c(ind_sev,ind_birth,ind_infek,ind_mat_infek, ind_parental));n_test_regist
```

```
## [1] 21
```

Polygenic Scores:

```
ind_y=23:29;  
ind_x=ind_prs;names(data)[ind_x]  
## [1] "BIP_2016_SCORES_AT_ALL_THRESHOLDS_05"  
## [2] "Cannib SCORES_AT_ALL_THRESHOLDS_05"
```

```

## [3] "DS_2016_SCORES_AT_ALL_THRESHOLDS_05"
## [4] "EduYears_2016_SCORES_AT_ALL_THRESHOLDS_05"
## [5] "NEURO_2016_SCORES_AT_ALL_THRESHOLDS_05"
## [6] "SCZ_2015_SCORES_AT_ALL_THRESHOLDS_05"
## [7] "SWB_2016_SCORES_AT_ALL_THRESHOLDS_05"
mf = formula(y ~ x + birthdate*gender + wave + C1 + C2 + C3 + C4 + C5 + C6 + C7 + C8 + C9 + C10)

manova.results_prs<-mancova.func(data[!is.na(data$C1),], ind_x,ind_y,mf)

ind_x_sig = ind_x[manova.results_prs$p.val <= .05/n_test_gen]
anova.results_prs<-ancova.func(data[!is.na(data$C1),], ind_x_sig,ind_y,mf)

```

Association with rare mutations:

```

ind_x=ind_rare;names(data)[ind_x]

## [1] "disruptive_and_damaging" "synonymous"

mf = formula(y ~ x + birthdate*gender+C1+C2+C3+C4+C5+C6+C7+C8+C9+C10)

manova.results_rare<-mancova.func(data, ind_x,ind_y,mf)

ind_x_sig = ind_x[manova.results_rare$p.val <= .05/n_test_gen]
anova.results_rare<-ancova.func(data, ind_x_sig,ind_y,mf)

```

Contacts and visits

```

ind_x=ind_sev;names(data)[ind_x]
## [1] "contacts" "visits"
mf = formula(y ~ x + birthdate*gender)

manova.results_sev<-mancova.func(data, ind_x,ind_y,mf)

ind_x_sig = ind_x[manova.results_sev$p.val <= .05/n_test_regist]
anova.results_sev<-ancova.func(data, ind_x_sig,ind_y,mf)

```

Association with birth & pregnancy variables:

```

ind_x=c(ind_birth);names(data)[ind_x]
## [1] "maternal_age" "paternal_age" "langde"      "apgar5"
## [5] "gest_age"     "bw_score"     "B_SECTIOU"   "B_I11"
## [9] "RYGER"
mf = formula(y ~ x + birthdate*gender)

manova.results_birth<-mancova.func(data, ind_x,ind_y,mf)

ind_x_sig = ind_x[manova.results_birth$p.val <= .05/n_test_regist]
anova.results_birth<-ancova.func(data, ind_x_sig,ind_y,mf)

```

Association with infection diagnoses during pregnancy

```

ind_x=ind_mat_infek;names(data)[ind_x]
## [1] "ind_mat_preg_Bacterial_infection" "ind_mat_preg_Viral_infection"
summary(data[,ind_x])
##      Min. 1st Qu. Median   Mean 3rd Qu.   Max.
##    42.0    42.2    42.5    42.5    42.8    43.0
mf = formula(y ~ x + birthdate*gender)

manova.results_mat_infek<-mancova.func(data, ind_x,ind_y,mf)

ind_x_sig = ind_x[manova.results_mat_infek$p.val <= .05/n_test_regist]
#anova.results_mat_infek<-ancova.func(data, ind_x_sig,ind_y,mf)

```

Parental diagnosis of Schizophrenia and all Fxx diagnosis

```

ind_x=ind_parental;names(data)[ind_x]
## [1] "ind_Fxxxx_m" "ind_F2100_m" "ind_Fxxxx_f" "ind_F2100_f"
mf = formula(y ~ x + birthdate*gender)

manova.results_fam<-mancova.func(data, ind_x,ind_y,mf)

ind_x_sig = ind_x[manova.results_fam$p.val <= .05/n_test_regist]
anova.results_fam<-ancova.func(data, ind_x_sig,ind_y,mf)

```

Association with infection diagnoses

```

ind_x=ind_infek;names(data)[ind_x]
## [1] "ind_Otitis_infection"      "ind_Bacterial_infection"
## [3] "ind_CNS_infection"        "ind_Viral_infection"
mf = formula(y ~ x + birthdate*gender)

manova.results_infek<-mancova.func(data, ind_x,ind_y,mf)

ind_x_sig = ind_x[manova.results_infek$p.val <= .05/n_test_regist]
anova.results_infek<-ancova.func(data, ind_x_sig,ind_y,mf)

```

Heatmap of associations using twerk of corrplot::corrplot included in diagtraject::corrplot:

```

anovas <- list(
  anova.results_prs,
  anova.results_rare,
  anova.results_birth,
  anova.results_infek,
  #anova.results_mat_infek,
  anova.results_sev,
  anova.results_fam)
betas <- t(Reduce('cbind',lapply(anovas,function(x) x$beta)))
p.vals <- t(Reduce('cbind',lapply(anovas,function(x) x$p.val)))

library(corrplot)

## corrplot 0.84 loaded

```

```

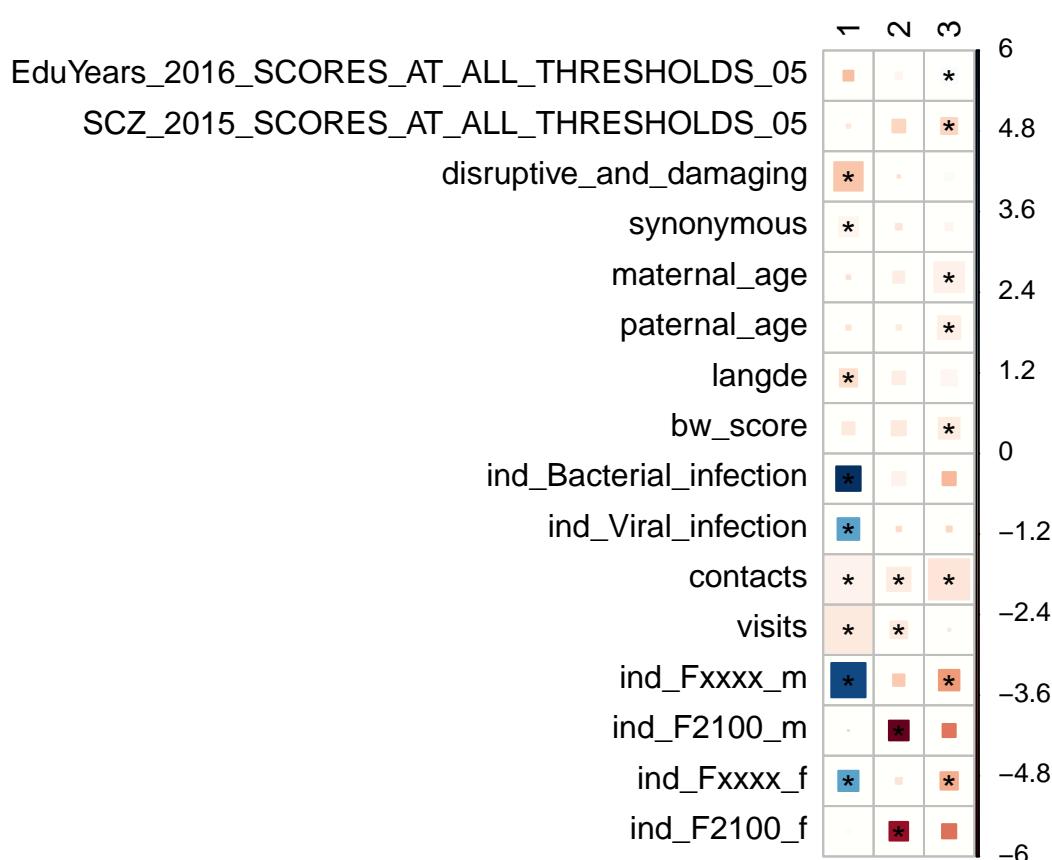
## 
## Attaching package: 'corrplot'

## The following object is masked from 'package:diagtraject':
## 
##     corrplot

load_all("~/trajectory/single/all/diagtraject0.5/")

## Loading diagtraject

corr.plot <- corrplot(betas[,1:3], is.corr=FALSE, p.mat=-log10(p.vals[,1:3])/max(-log10(p.vals[,1:3])),
```



## 1.9 Replication

### 1.9.1 Replication dataset:

Sequence Analysis is performed in non-overlapping dataset consisting of patients diagnosed January 1 2013 - December 31 2016.

```

load(diagdates2016)
diagdates_rep <- data.table(rbind(diagdates,diagdates2016))
```

Sequence object is created as above repeated for dataset of replication samples and primary cohort:

```

dt_rep <- data.table(diagdates_rep)

dt_rep[, "birthdate" := fdate(birthdate) ]
dt_rep[, c("F1", "F3", "F4", "F50", "F60", "F70", "F84", "F9", "SCZ") := 
  lapply(list(F1,F3,F4,F50,F60,F70,F84,F9,SCZ), function(x){
    (as.numeric(fdate(x))- as.numeric(birthdate))/365})]
dt_rep[, censored:= (as.numeric(fdate("31/12/2016"))-as.numeric(birthdate))/365] #Calculates age in 31/12/2016

dt.m <- melt(dt_rep, id.vars = 'pid', direction = "long", measure.vars = list(c(3:10,12)), value.name = "time")
dt.m<- dt.m[order(pid)]
dt.m[,time:=time1]
dt.m <- dt.m[!is.na(time)]

events <- levels(dt.m$event)
drop <- matrix(FALSE, nrow = length(events), ncol = length(events), dimnames = list(events,events))
drop['censored',] <- T
drop[, 'censored'] <- T
diag(drop) <- F

e2sm <- seqe2stm(events = events, dropMatrix = drop)

seq_rep <- seqdef(TSE_to_STS(dt.m,id = "pid", timestamp = "time", event= "event",
                             tmin=1, tmax=ceiling(max(dt.m$time)),
                             stm = e2sm),firstState ="None")

seq_mis_rep <- seqdef(TSE_to_STS(dt.m,
                                   id = "pid", timestamp = "time", event= "event",
                                   tmin=1, tmax=ceiling(max(dt.m$time)),stm = e2sm)
                        , firstState ="None", missing="censored")

drop['censored',] <- F
drop[, 'censored'] <- T
e2sm <- seqe2stm(events = events, dropMatrix = drop)

seq_last <- seqdef(
  TSE_to_STS(dt.m, id = "pid", timestamp = "time",
              event= "event", tmin=1,tmax=ceiling(max(dt.m$time))+1,
              stm = e2sm),
  firstState ="None", missing="censored")

seq_mis_last <- seq_mis
seq_mis_last$last <- seq_last[,ncol(seq_last)]
attributes(seq_mis_last)$alphabet <- c(alphabet(seq_mis_last), unique(seq_mis_last$last)[which(!unique(seq_mis_last$last))])
attributes(seq_mis_last)$alphabet <- attributes(seq_mis_last)$alphabet[order(attributes(seq_mis_last)$labels)]
attributes(seq_mis_last)$labels <- attributes(seq_mis_last)$alphabet

sms <- which(rownames(sub.cost_jacc) %in% rownames(seqsubm(seq,"CONSTANT")))

d_OM <- seqdist(seq_rep, method = "OM", indel=.5, sm = sub.cost_jacc[sms,sms])
ms <- mis.cost(seq_mis_rep, seq_mis_last$last, tr_t, sm = sub.cost_jacc,
               cens.type = "right", imp.length = "max",
               diag=F, sum_to_1=T, resol.comp = resol.comp,
               resol.ratio = resol.reduc, mc.cores=27)

```

```

dist_OM_rep <- d_OM + as.matrix(ms$dist)

mcor <- match(seqconc(seq_mis_last), seqconc(unique(seq_mis_last))) # finding unique sequences
uni <- (!duplicated(mcor))
dist_unique_OM_rep <- list("dist"=dist_OM_rep[uni,uni], "weight"= table(mcor), "mcor"= mcov)

```

Weighted MDS is performed assigning weights of  $10^{-20}$  to replication sample:

```

dt_rep[,mcov2:=dist_unique_OM_rep$mcov]
case_cor <- dt_rep[case==1,sum(case2012,na.rm = T)+1e-20,mcov2]
setkey(case_cor, mcov2)
w <- case_cor$V1

```

Showing that MDS of Case2012 in unaltered

```

wmd2012 <- wcmdscale(dist_unique_OM_rep$dist[w>=1,w>=1], w = w[w>=1], k=7)
wmd_rep <- wcmdscale(dist_unique_OM_rep$dist, w = case_cor$V1, k=7,)

sum(abs(scale(wmd2012[,1])-scale(wmd_rep[w>=1,1])))
## [1] 6.01e-13

specify_decimal <- function(x, k) trimws(format(round(x, k), nsmall=k))

specify_decimal(diag(cor(scale(wmd2012)-scale(wmd_rep[w>=1,]))),15)
## [1] "1.000000000000000" "0.999999999999993" "0.999999999999998"
## [4] "1.000000000000000" "1.000000000000000" "1.000000000000000"
## [7] "1.000000000000000"

dt_rep[,paste0("Dim",1:7):=as.data.frame(wmd_rep[dt_rep$mcov2,])]
```

Thus producing a projection of Case2016 unto the MDS og Case2012:

### 1.9.2 Merging with putative risk variable data

Testing which significant associations replicate:

```
data_rep <- merge(dt_rep,additional,by="pid")
```

Modify as above:

```

data_rep$apgar5=factor(data_rep$apgar5,levels=c("0","1","2","3","4","5","6","7","8","9","10","A"))
data_rep$apgar5=as.character(data_rep$apgar5)
data_rep$apgar5[data_rep$apgar5=="A"]=NA
data_rep$apgar5=as.numeric(data_rep$apgar5)
langde=data_rep$langde
langde[langde=="A"]=NA
langde[langde=="1"]=NA
langde[langde=="10"]=NA
langde=as.numeric(as.character(langde))
data_rep$langde=langde;rm(langde)
```

```

data_rep$C_RYGER[data_rep$C_RYGER==99]=NA
data_rep$RYGER <- data_rep$B_RYGER
data_rep$RYGER[!is.na(data_rep$C_RYGER)] <-1

data_rep$wave[is.na(data_rep$wave)]="Other"
data_rep[,wave:=factor(wave)]

nms = grep("d2100",names(data_rep))
nms <- names(data_rep)[nms];nms
data_rep[,contacts:= Reduce('+', .SD), .SDcols=nms[c(1,3,7)]]
data_rep[,visits:= Reduce('+', .SD), .SDcols=nms[4:5]]

```

Standardize PRS scores

```

prs = names(data_rep)[grep("THRESHOLDS_05", names(data_rep))]
data_rep[!is.na(C1),c(prs):= lapply(.SD,scale),.SDcols=prs]

```

### 1.9.3 Association analysis in replication data:

Test replication of associations with  $p < 0.05/17$ :

```

rep_geno <- data.table(which(p.vals[1:3,1:3]<0.05/nrow(betas),arr.ind = T),keep.rownames = T )
rep_geno[,dim:=paste0("Dim",col)]
rep_geno[,2:3:=NULL,with=F]

p_rep <- lapply(1:nrow(rep_geno), function(i) {
  tmp <- data_rep[is.na(case2012)&!is.na(C1)]
  tmp$x <-tmp[,rep_geno[i,rn],with=F]
  tmp$y <- tmp[,rep_geno[i,dim],with=F]
  mf <- y~x+gender*birthdate+C1+C2+C3+C4+C5+C6+C7+C8+C9+C10+wave
  list(anova= anova(lm(mf, tmp)),coef=coef(lm(mf, tmp)))
})

rep_geno[,N:=unlist(lapply(p_rep,function(x) sum(x[[1]][,1])+1))]
rep_geno[,Fstat:=unlist(lapply(p_rep,function(x) x[[1]][1,4]))]
rep_geno[,p.val:=unlist(lapply(p_rep,function(x) x[[1]][1,5]))]
rep_geno[,coef:=unlist(lapply(p_rep,function(x) x[[2]][2]))]

options(digits=3, scipen=T)
kable(rep_geno,digits=c(1,1,1,1,32,2))

```

rn		dim	N	Fstat	p.val	coef
disruptive_and_damage		Dim1	154	2.3	0.1320	0.46
EduYears_2016_SCORES_AT_ALL_THRESHOLDS_05		Dim3	650	4.8	0.0284	-0.17
SCZ_2015_SCORES_AT_ALL_THRESHOLDS_05		Dim3	650	6.6	0.0102	0.21

```

rep_register <- data.table(which(p.vals[5:16,1:3]<.01,arr.ind = T),keep.rownames = T )
rep_register[,dim:=paste0("Dim",col)]
rep_register[,2:3:=NULL,with=F]

p_rep <- lapply(1:nrow(rep_register), function(i) {
  tmp <- data_rep[is.na(case2012)]
  tmp$x <-tmp[,rep_register[i,rn],with=F]
  tmp$y <- tmp[,rep_register[i,dim],with=F]
  mf <- y~x+gender*birthdate
  list(anova= anova(lm(mf, tmp)),coef=coef(lm(mf, tmp)))
})

rep_register[,N:=unlist(lapply(p_rep,function(x) sum(x[[1]][,1])+1))]
rep_register[,Fstat:=unlist(lapply(p_rep,function(x) x[[1]][1,4]))]
rep_register[,p.val:=unlist(lapply(p_rep,function(x) x[[1]][1,5]))]
rep_register[,coef:=unlist(lapply(p_rep,function(x) x[[2]][2]))]

kable(rep_register,digits=c(1,1,1,1,32,2))

```

rn	dim	N	Fstat	p.val	coef
langde	Dim1	851	1.4	2.33e-01	0.01
ind_Bacterial_infection	Dim1	870	3.2	7.25e-02	-0.52
ind_Viral_infection	Dim1	870	6.0	1.45e-02	-0.51
contacts	Dim1	857	30.8	3.84e-08	-0.08
visits	Dim1	857	1.5	2.16e-01	0.00
ind_Fxxxx_m	Dim1	858	3.1	7.92e-02	-0.31
ind_Fxxxx_f	Dim1	858	1.5	2.23e-01	-0.15
bw_score	Dim2	834	1.8	1.75e-01	0.00
contacts	Dim2	857	3.9	4.94e-02	-0.03
visits	Dim2	857	14.2	1.80e-04	-0.01
ind_F2100_m	Dim2	858	1.0	3.25e-01	-0.68
ind_F2100_f	Dim2	858	0.6	4.37e-01	0.43
maternal_age	Dim3	856	20.2	8.13e-06	-0.04
paternal_age	Dim3	843	2.5	1.12e-01	0.00
langde	Dim3	851	1.5	2.24e-01	-0.02
bw_score	Dim3	834	3.4	6.37e-02	0.00
contacts	Dim3	857	0.1	8.03e-01	0.01
ind_Fxxxx_m	Dim3	858	1.3	2.62e-01	0.12
ind_Fxxxx_f	Dim3	858	0.2	6.41e-01	-0.08
ind_F2100_f	Dim3	858	0.9	3.53e-01	-0.60

## 1.10 Robustness

Selecting subset of data that does not require imputation:

```

#seq25 <- seq[as.numeric(dt_mds[case2012==1,year]) < 1991,]
seq <- seq[rownames(seq) %in% dt[case2012==1,pid],]
seq25 <- seq[as.numeric(dt_mds[case2012==1,year]) < 1991 & as.numeric(dt_mds[case2012==1,SCZ]) < 25,]
seq25 <- seq25[,seq(1,25)]

```

Sequence Analysis with 3 different substitution cost settings: Jaccard Distance:

```
sub.cost <- seqsubm(seqdata = seq25, method = "CONSTANT")
alp <- seqdef(as.data.frame(alphabet(seq25)), stsep="[.]")
rownames(alp) <- alphabet(seq25)
sub.cost_jacc <- matrix(0, nrow(alp), nrow(alp))
alp <- alp[order(seqlength(alp)),]
for(i in 1:nrow(alp))
  for(j in 1:nrow(alp))
    if(i>=j){
      sub.cost_jacc[i,j] <- 1 - sum(unlist(alp[i,1:seqlength(alp[i])])) %in% unlist(alp[j,1:seqlength(alp[j])])
      length(unique(c(unlist(alp[i,1:seqlength(alp[i])]), unlist(alp[j,1:seqlength(alp[j])]))))} else
sub.cost_jacc <- as.matrix(as.dist(sub.cost_jacc))
rownames(sub.cost_jacc) <- colnames(sub.cost_jacc) <- rownames(alp)
sub.cost_jacc <- sub.cost_jacc[order(rownames(sub.cost_jacc)), order(rownames(sub.cost_jacc))]
rownames(sub.cost_jacc) <- rownames(sub.cost)
colnames(sub.cost_jacc) <- colnames(sub.cost)
```

1- Simple matching coefficient (Equivalent to Multichannel Sequence Analysis)

```
sub.cost_smc <- matrix(0, nrow(alp), nrow(alp))
for(i in 1:nrow(alp))
  for(j in 1:nrow(alp))
    if(i>=j){
      sub.cost_smc[i,j] <- sum(!unlist(alp[i,1:seqlength(alp[i])])) %in% unlist(alp[j,1:seqlength(alp[j])])
    } else sub.cost_smc[i,j] <- NA
sub.cost_smc <- as.matrix(as.dist(sub.cost_smc))
rownames(sub.cost_smc) <- colnames(sub.cost_smc) <- rownames(alp)
sub.cost_smc <- sub.cost_smc[order(rownames(sub.cost_smc)), order(rownames(sub.cost_smc))]
rownames(sub.cost_smc) <- rownames(sub.cost)
colnames(sub.cost_smc) <- colnames(sub.cost)
```

Computing dissimilarities

```
dist_list <- list(
OM_jacc_.5 = seqdist(seq25, method = "OM", sm = sub.cost_jacc, indel = .5*max(sub.cost_jacc)),
OM_jacc_1 = seqdist(seq25, method = "OM", sm = sub.cost_jacc, indel = max(sub.cost_jacc)),
HAM_jacc = seqdist(seq25, method = "HAM", sm = sub.cost_jacc),

OM_smc_.5 = seqdist(seq25, method = "OM", sm = sub.cost_smc, indel = .5*max(sub.cost_smc)),
OM_smc_1 = seqdist(seq25, method = "OM", sm = sub.cost_smc, indel = max(sub.cost_smc)),
HAM_smc = seqdist(seq25, method = "HAM", sm = sub.cost_smc),

OM_const_.5 = seqdist(seq25, method = "OM", sm = "CONSTANT", indel = 1),
OM_const_1 = seqdist(seq25, method = "OM", sm = "CONSTANT", indel = 2),
HAM_const = seqdist(seq25, method = "HAM", sm = "CONSTANT"),

eucl_2 = seqdist(seq25[,2:25], method = "EUCLID", step = 2),
chi2_2 = seqdist(seq25[,2:25], method = "CHI2", step = 2),

eucl_5 = seqdist(seq25, method = "EUCLID", step = 5),
```

```

chi2_5 = seqdist(seq25, method = "CHI2", step = 5 ),

eucl_12 = seqdist(seq25[,2:25], method = "EUCLID", step = 12),
chi2_12 = seqdist(seq25[,2:25], method = "CHI2", step = 12 )
)

```

Multidimensional Scaling and MANCOVA:

```

MD_list <- mclapply(dist_list, function(x) cmdscale(x, k = 7), mc.cores = 16)

l <- length(MD_list)
setkey(data,pid)
data25 <- data[pid %in% rownames(seq25)]
man_list <- list(
  scz_pgs = lapply(1:l, function(x) {
    tmp<- as.data.frame(data25)
    tmp$y <- MD_list[[x]]
    summary(manova(y ~ SCZ_2015_SCORES_AT_ALL_THRESHOLDS_05 + birthdate +gender+wave+C1+C2+C3+C4+C5+C6+C7+C8+C9+C10, data=tmp))
  }),
  edu_pgs = lapply(1:l, function(x) {
    tmp<- as.data.frame(data25)
    tmp$y <- MD_list[[x]]
    summary(manova(y ~ EduYears_2016_SCORES_AT_ALL_THRESHOLDS_05+ birthdate +gender+wave+C1+C2+C3+C4+C5+C6+C7+C8+C9+C10, data=tmp))
  }),
  disr_and_damage = lapply(1:l, function(x) {
    tmp<- as.data.frame(data25)
    tmp$y <- MD_list[[x]]
    summary(manova(y ~ disruptive_and_damaging+ birthdate +gender+wave+C1+C2+C3+C4+C5+C6+C7+C8+C9+C10, data=tmp))
  }),
  mat_age = lapply(1:l, function(x) {
    tmp<- as.data.frame(data25)
    tmp$y <- MD_list[[x]]
    summary(manova(y ~ paternal_age+ birthdate +gender,data=tmp ))}),
  pat_age = lapply(1:l, function(x) {
    tmp<- as.data.frame(data25)
    tmp$y <- MD_list[[x]]
    summary(manova(y ~ maternal_age+ birthdate +gender,data=tmp ))}),
  langde =lapply(1:l, function(x) {
    tmp<- as.data.frame(data25)
    tmp$y <- MD_list[[x]]
    summary(manova(y ~ langde+ birthdate +gender,data=tmp ))}),
  bw_score = lapply(1:l, function(x) {
    tmp<- as.data.frame(data25)
    tmp$y <- MD_list[[x]]
    summary(manova(y ~ bw_score+ birthdate +gender,data=tmp ))}),
  bac_inf = lapply(1:l, function(x) {
    tmp<- as.data.frame(data25)
    tmp$y <- MD_list[[x]]
    summary(manova(y ~ ind_Bacterial_infection+ birthdate +gender,data=tmp ))}),
  vir_inf = lapply(1:l, function(x) {
    tmp<- as.data.frame(data25)
    tmp$y <- MD_list[[x]]
    summary(manova(y ~ ind_Viral_infection+ birthdate +gender,data=tmp ))}),
  contacts =lapply(1:l, function(x) {
    tmp<- as.data.frame(data25)
    tmp$y <- MD_list[[x]]
    summary(manova(y ~ contacts+ birthdate +gender,data=tmp ))})
)

```

```

tmp$y <- MD_list[[x]]
summary(manova(y ~ contacts+ birthdate +gender,data=tmp ))}),
visits =lapply(1:1, function(x) {
  tmp<- as.data.frame(data25)
  tmp$y <- MD_list[[x]]
  summary(manova(y ~ visits+ birthdate +gender,data=tmp ))}),
ind_Fxxxx_m =lapply(1:1, function(x) {
  tmp<- as.data.frame(data25)
  tmp$y <- MD_list[[x]]
  summary(manova(y ~ ind_Fxxxx_m+ birthdate +gender,data=tmp ))}),
ind_F2100_m =lapply(1:1, function(x) {
  tmp<- as.data.frame(data25)
  tmp$y <- MD_list[[x]]
  summary(manova(y ~ ind_F2100_m+ birthdate +gender,data=tmp ))}),
ind_Fxxxx_f =lapply(1:1, function(x) {
  tmp<- as.data.frame(data25)
  tmp$y <- MD_list[[x]]
  summary(manova(y ~ ind_Fxxxx_f+ birthdate +gender,data=tmp ))}),
ind_F2100_f =lapply(1:1, function(x) {
  tmp<- as.data.frame(data25)
  tmp$y <- MD_list[[x]]
  summary(manova(y ~ ind_F2100_f+ birthdate +gender,data=tmp ))})
})

p.mat <- Reduce("rbind",lapply(man_list, function(x) sapply(x, function(y) y$stats[1,6])))

row.names(p.mat) <- names(man_list)
colnames(p.mat) <- names(dist_list)

options(digits=2, scipen=T)
kable(p.mat,digits=32)

```

	OM_jacc_.5	OM_jacc_1	HAM_jacc	OM_smc_.5	OM_smc_1	HAM_smc	OM_const_5
scz_pgs	1.6e-02	1.7e-02	1.7e-02	1.8e-02	1.8e-02	1.8e-02	4.7e-02
edu_pgs	7.1e-02	5.6e-02	5.6e-02	6.7e-02	6.7e-02	6.7e-02	2.1e-01
disr_and_damage	6.5e-03	9.6e-03	9.6e-03	5.9e-02	5.9e-02	5.9e-02	7.1e-02
mat_age	1.1e-01	1.9e-01	1.9e-01	1.3e-01	1.3e-01	1.3e-01	1.3e-01
pat_age	8.3e-05	1.2e-04	1.2e-04	1.2e-03	1.2e-03	1.2e-03	1.3e-04
langde	1.5e-03	1.8e-03	1.8e-03	6.1e-06	6.1e-06	6.1e-06	1.2e-01
bw_score	4.7e-03	6.4e-03	6.4e-03	1.2e-01	1.2e-01	1.2e-01	9.3e-02
bac_inf	5.2e-13	1.3e-12	1.3e-12	2.8e-10	2.8e-10	2.8e-10	1.6e-10
vir_inf	1.1e-02	1.5e-02	1.5e-02	4.2e-02	4.2e-02	4.2e-02	1.9e-02
contacts	1.6e-15	3.8e-15	3.8e-15	6.8e-19	6.8e-19	6.8e-19	1.9e-11
visits	6.9e-09	9.6e-09	9.6e-09	2.7e-10	2.7e-10	2.7e-10	1.0e-07
ind_Fxxxx_m	5.6e-08	1.1e-08	1.1e-08	3.3e-09	3.3e-09	3.3e-09	2.4e-08
ind_F2100_m	5.6e-05	4.6e-05	4.6e-05	2.0e-06	2.0e-06	2.0e-06	6.6e-04
ind_Fxxxx_f	5.5e-05	2.2e-05	2.2e-05	1.1e-02	1.1e-02	1.1e-02	3.5e-05
ind_F2100_f	2.3e-02	1.7e-02	1.7e-02	9.7e-02	9.7e-02	9.7e-02	2.4e-02

## 1.11 Post-hoc Regressions

```
#data2 <- merge(data, dt_mds[,.(pid,age_min,n_dia)],by="pid")
data2 <- data
dim(data2)
```

```
## [1] 5432 94
```

Dimension 1 and Rare Mutations:

Adjusting for count of synonomous mutations:

```
anova(lm(Dim1~disruptive_and_damaging+synonymous+gender+birthdate+C1+C2+C3+C4+C5+C6+C7+C8+C9+C10+wave,da
```

```
## [1] 3.6e-09
```

Adjusting for Schizophrenia diagnosis before/after Jan 1, 1995.

```
anova(lm(Dim1~disruptive_and_damaging+synonymous+gender+birthdate+C1+C2+C3+C4+C5+C6+C7+C8+C9+C10+wave+(
```

```
## [1] 3.6e-09
```

```
anova(lm(Dim1~disruptive_and_damaging+synonymous+gender+birthdate+C1+C2+C3+C4+C5+C6+C7+C8+C9+C10+wave+(
```

```
## [1] 3.6e-09
```

Adjusting for family history

```
anova(lm(Dim1~disruptive_and_damaging+gender+synonymous+birthdate+ind_Fxxxx_f+ind_F2100_f+ind_F2100_m+in
```

```
## [1] 2.7e-09
```

Adjusting for age at first diagnosis and total number of diagnoses:

```
summary(lm(Dim1~disruptive_and_damaging+synonymous+gender+birthdate+C1+C2+C3+C4+C5+C6+C7+C8+C9+C10+wave
```

```
## [1] 0.0056
```

```
summary(lm(Dim1~disruptive_and_damaging+age_min+synonymous+gender+birthdate+C1+C2+C3+C4+C5+C6+C7+C8+C9+C10+wave
```

```
## [1] 0.0066
```

```
summary(lm(Dim1~disruptive_and_damaging+n_dia+synonymous+gender+birthdate+C1+C2+C3+C4+C5+C6+C7+C8+C9+C10+wave
```

```
## [1] 0.23
```

Rare mutations ~ number of diagnoses:

```

summary(glm(disruptive_and_damaging~n_dia+gender+synonymous+birthdate+C1+C2+C3+C4+C5+C6+C7+C8+C9+C10+wa))

## [1] 0.0054

exp(coef(glm(disruptive_and_damaging~n_dia+gender+synonymous+birthdate+C1+C2+C3+C4+C5+C6+C7+C8+C9+C10+wa))

## n_dia
## 0.95

exp(confint(glm(disruptive_and_damaging~n_dia+gender+synonymous+birthdate+C1+C2+C3+C4+C5+C6+C7+C8+C9+C10+wa))

## Waiting for profiling to be done...

## 2.5 % 97.5 %
## 0.92 0.99

```

Dimension 3 and Polygenic Scores for Schizophrenia and Educational Attainment:

```

anova(lm(Dim3~EduYears_2016_SCORES_AT_ALL_THRESHOLDS_05+gender+birthdate+C1+C2+C3+C4+C5+C6+C7+C8+C9+C10+wave))

## [1] 0.0022

anova(lm(Dim3~SCZ_2015_SCORES_AT_ALL_THRESHOLDS_05+gender+birthdate+C1+C2+C3+C4+C5+C6+C7+C8+C9+C10+wave))

## [1] 0.0022

```

Substance Abuse ~ Education Polygenic scores:

```

summary(glm(!is.na(F10)~scale(EduYears_2016_SCORES_AT_ALL_THRESHOLDS_05)+gender+birthdate+C1+C2+C3+C4+C5+C6+C7+C8+C9+C10+wave))

## [1] 0.016

exp(coef(glm(!is.na(F10)~scale(EduYears_2016_SCORES_AT_ALL_THRESHOLDS_05)+gender+birthdate+C1+C2+C3+C4+C5+C6+C7+C8+C9+C10+wave))

## scale(EduYears_2016_SCORES_AT_ALL_THRESHOLDS_05)
## 0.89

exp(confint(glm(!is.na(F10)~scale(EduYears_2016_SCORES_AT_ALL_THRESHOLDS_05)+gender+birthdate+C1+C2+C3+C4+C5+C6+C7+C8+C9+C10+wave))

## Waiting for profiling to be done...

## 2.5 % 97.5 %
## 0.81 0.98

```

Substance Abuse ~ Schizophrenia Polygenic scores:

```

summary(glm(!is.na(F10)~SCZ_2015_SCORES_AT_ALL_THRESHOLDS_05+gender+birthdate+C1+C2+C3+C4+C5+C6+C7+C8+C9))

## [1] 0.094

exp(coef(glm(!is.na(F10)~scale(SCZ_2015_SCORES_AT_ALL_THRESHOLDS_05)+gender+birthdate+C1+C2+C3+C4+C5+C6+C7+C8+C9))

## scale(SCZ_2015_SCORES_AT_ALL_THRESHOLDS_05)
##                                     1.1

exp(confint(glm(!is.na(F10)~scale(SCZ_2015_SCORES_AT_ALL_THRESHOLDS_05)+gender+birthdate+C1+C2+C3+C4+C5+C6+C7+C8+C9))

## Waiting for profiling to be done...

##   2.5 % 97.5 %
##   0.99    1.19

Early Substance Abuse ~ Schizophrenia Polygenic scores

data2[,median(F10,na.rm = T)]

## [1] 21

data2[,earl_F10 := 0]

##          pid year gender birthdate case case2012 random F10 F20 F30 F40
## 1: 1000166 1990      F 1990-12-13    1      1     NA     NA     NA 21  NA
## 2: 1002947 1989      M 1989-07-15    1      1     NA    17     NA     NA
## 3: 1009068 1987      M 1987-03-03    1      1     NA     NA     NA 21  NA
## 4: 1013507 1984      F 1984-11-21    1      1     NA     NA     NA 16  NA
## 5: 1014243 1983      F 1983-10-20    1      1     NA     NA    21  20  NA
##   ---
## 5428: 9315714 1993      M 1993-09-19    1      1      1    17    18  NA 23
## 5429: 9316160 1985      M 1985-09-08    1      1     NA    18    18  NA  NA
## 5430: 9319643 1991      F 1991-03-02    1      1     NA    20     NA    19
## 5431: 9320666 1998      F 1998-09-09    1      1     NA     NA     NA  NA
## 5432: 9323646 1982      M 1982-08-08    1      1     NA    24    27  NA  NA
##          F50 F60 F70 F84 F90 SCZ missing censored N pop_weights mcov Dim1
## 1:    NA  NA  NA  NA  NA  21    26    26 634           1    2 -2.45
## 2:    NA  18  24  NA  17  19    27    27 617           1   12  6.02
## 3:    NA  21  NA  NA  NA  21    30    30 585           1   19  0.85
## 4:   16  16  NA  NA  NA  18    32    32 525           1   25  6.36
## 5:   20  20  NA  NA  NA  32    22    33 480           1   26  3.52
##   ---
## 5428:  NA  NA  NA  NA  23  18    23    23 607           1 5682  3.16
## 5429:  NA  NA  NA  14  18  20    31    31 592           1 5683  6.96
## 5430:  NA  NA  NA  NA  NA  19    26    26 613           1 5687  1.21
## 5431:  NA  NA  NA  NA  NA  14    18    18 613           1   17 -1.61
## 5432:  NA  NA  NA  NA  NA  27    34    34 492           1 322 -5.19
##          Dim2 Dim3 Dim4 Dim5 Dim6 Dim7 n_dia age_min dia_year fit
## 1:  1.70 2.558 -1.86 -0.18 -1.764 -0.616    1     21    2012  0.70

```

```

##      2:  1.00 -3.387  1.89 -0.67  0.355 -1.845    4     17   2008  0.20
##      3:  2.86  1.243 -1.62 -2.22  0.956  1.532    2     21   2008 -0.64
##      4:  3.86  2.045  0.73 -0.92  1.240 -1.178    3     16   2003 -1.44
##      5:  3.92  0.121  1.00 -1.39  0.152 -0.140    4     20   2005 -1.83
##      ---
## 5428: -0.41 -3.712 -0.61  1.25 -0.391  0.241    7     17   2011  1.67
## 5429: -3.39 -1.517  0.46  3.53  0.487 -0.078    3     14   2005 -1.16
## 5430:  0.77 -2.695 -1.11  1.50  1.454  2.479    2     19   2010  0.77
## 5431:  0.52 -0.012  0.83 -0.37  0.045  0.584    0     14   2012  3.43
## 5432: -0.93 -2.280 -0.28  1.19 -0.237 -0.908    1     24   2009 -2.25
##      maternal_age paternal_age langde apgar5 gest_age bw_score C_RYGER
##      1:          16            23   50    10     38    75.9    NA
##      2:          22            33   54    10     39    98.4    NA
##      3:          32            51   54    10     42    15.0    NA
##      4:          30            29   50    10     38    63.3    NA
##      5:          31            34   49    NA     37    57.3    NA
##      ---
## 5428:          32            25   52    10     39    46.4    NA
## 5429:          18            21   54    10     40    64.7    NA
## 5430:          32            24   46    10     35    45.7    NA
## 5431:          28            36   52    10     38    77.3     0
## 5432:          18           NA   45    10     36     8.9    NA
##      B_RYGER ind_mat_preg_Bacterial_infection
##      1:       NA                      0
##      2:       NA                      0
##      3:       NA                      0
##      4:       NA                      0
##      5:       NA                      0
##      ---
## 5428:        0                      0
## 5429:       NA                      0
## 5430:        1                      0
## 5431:       NA                      0
## 5432:       NA                      0
##      ind_mat_preg_Viral_infection AN_2016_SCORES_AT_ALL_THRESHOLDS_05
##      1:                      0                  -0.059
##      2:                      0                  -0.059
##      3:                      0                      NA
##      4:                      0                      NA
##      5:                      0                      NA
##      ---
## 5428:        0                  -0.059
## 5429:        0                      NA
## 5430:        0                  -0.059
## 5431:        0                  -0.059
## 5432:        0                  -0.059
##      Anxiety_SCORES_AT_ALL_THRESHOLDS_05
##      1:        -0.00045
##      2:        -0.00051
##      3:                      NA
##      4:                      NA
##      5:                      NA
##      ---
## 5428:        -0.00058

```

```

## 5429:                               NA
## 5430:                           -0.00050
## 5431:                           -0.00054
## 5432:                           -0.00046
##      BMI_2015_SCORES_AT_ALL_THRESHOLDS_05
## 1:                               -0.00045
## 2:                           -0.00042
## 3:                               NA
## 4:                               NA
## 5:                               NA
## ---
## 5428:                           -0.00042
## 5429:                               NA
## 5430:                           -0.00034
## 5431:                           -0.00041
## 5432:                           -0.00033
##      BIP_2016_SCORES_AT_ALL_THRESHOLDS_05 BW_SCORES_AT_ALL_THRESHOLDS_05
## 1:                               0.423          0.0029
## 2:                           -0.779          0.0032
## 3:                               NA           NA
## 4:                               NA           NA
## 5:                               NA           NA
## ---
## 5428:                           0.031          0.0031
## 5429:                               NA           NA
## 5430:                           0.532          0.0030
## 5431:                           -2.547          0.0031
## 5432:                           1.358          0.0031
##      Cannib SCORES_AT_ALL_THRESHOLDS_05
## 1:                               -0.10
## 2:                           -2.71
## 3:                               NA
## 4:                               NA
## 5:                               NA
## ---
## 5428:                           -0.81
## 5429:                               NA
## 5430:                           -0.69
## 5431:                           -1.37
## 5432:                           0.91
##      DS_2016_SCORES_AT_ALL_THRESHOLDS_05
## 1:                           -0.201
## 2:                           0.598
## 3:                               NA
## 4:                               NA
## 5:                               NA
## ---
## 5428:                           -0.107
## 5429:                               NA
## 5430:                           0.534
## 5431:                           0.714
## 5432:                           -0.081
##      EduYears_2016_SCORES_AT_ALL_THRESHOLDS_05
## 1:                           -0.682

```

```

##      2:          -0.624
##      3:             NA
##      4:             NA
##      5:             NA
##      ---
## 5428:          -0.417
## 5429:             NA
## 5430:          -0.324
## 5431:         -1.766
## 5432:          0.065
##      GPC_Ext_SCORES_AT_ALL_THRESHOLDS_05
##      1:          -0.000071
##      2:         -0.000149
##      3:             NA
##      4:             NA
##      5:             NA
##      ---
## 5428:          -0.000068
## 5429:             NA
## 5430:          -0.000102
## 5431:          -0.000082
## 5432:          -0.000069
##      NEURO_2016_SCORES_AT_ALL_THRESHOLDS_05
##      1:            0.60
##      2:          -1.48
##      3:             NA
##      4:             NA
##      5:             NA
##      ---
## 5428:            1.86
## 5429:             NA
## 5430:          -0.14
## 5431:          -1.20
## 5432:            0.96
##      SCZ_2015_SCORES_AT_ALL_THRESHOLDS_05
##      1:          -0.80
##      2:            0.35
##      3:             NA
##      4:             NA
##      5:             NA
##      ---
## 5428:            1.14
## 5429:             NA
## 5430:          -0.37
## 5431:          -0.19
## 5432:            1.51
##      SWB_2016_SCORES_AT_ALL_THRESHOLDS_05      C1      C2      C3      C4
##      1:          -0.40 -0.0027 -0.0009  0.0034  0.0073
##      2:           0.66 -0.0035  0.0085 -0.0021 -0.0029
##      3:             NA             NA             NA             NA
##      4:             NA             NA             NA             NA
##      5:             NA             NA             NA             NA
##      ---
## 5428:          -0.13 -0.0051 -0.0007  0.0051 -0.0001

```

```

## 5429:                               NA      NA      NA      NA      NA
## 5430:                           -0.66  0.0067 -0.0022 -0.0004 -0.0033
## 5431:                            0.30  0.0075 -0.0035  0.0027  0.0038
## 5432:                            1.66  0.0002 -0.0014  0.0043 -0.0019
##          C5      C6      C7      C8      C9      C10   wave
## 1: -0.0030  0.0000  0.0065 -0.0011 -0.0004  0.0005 Wave13
## 2: -0.0010 -0.0049  0.0020  0.0064  0.0046  0.0027 Wave14
## 3:     NA      NA      NA      NA      NA      NA Other
## 4:     NA      NA      NA      NA      NA      NA Other
## 5:     NA      NA      NA      NA      NA      NA Other
## ---
## 5428:  0.0031 -0.0008  0.0060  0.0078  0.0010  0.0034 Wave17
## 5429:     NA      NA      NA      NA      NA      NA Other
## 5430:  0.0048 -0.0028 -0.0016  0.0001  0.0043  0.0046 Wave10
## 5431: -0.0105 -0.0024 -0.0025 -0.0050 -0.0038 -0.0037 Wave4_
## 5432:  0.0000  0.0019 -0.0006  0.0054  0.0001 -0.0057 Wave20
##          d2100_ptype0_contacts d2100_ptype0_days d2100_ptype1_contacts
## 1:           7             234                  0
## 2:            0              0                  0
## 3:            0              0                  0
## 4:            32             645                  0
## 5:            2              65                  0
## ---
## 5428:           9             433                  0
## 5429:            0              0                  0
## 5430:            8             120                  0
## 5431:            0              0                  0
## 5432:            4              65                  0
##          d2100_ptype1_visits d2100_ptype2_visits d2100_ptype2_novisits
## 1:            0              69                  0
## 2:            0              0                  0
## 3:            0             114                  0
## 4:            0              69                  0
## 5:            0              0                  0
## ---
## 5428:            0              80                  0
## 5429:            0              77                  0
## 5430:            0             103                  0
## 5431:            0              89                  0
## 5432:            0              90                  1
##          d2100_ptype3_contacts ind_Otitis_infection ind_Bacterial_infection
## 1:            1                  0                  0
## 2:            0                  0                  0
## 3:            0                  0                  0
## 4:            12                 0                  0
## 5:            4                  0                  1
## ---
## 5428:            4                  0                  0
## 5429:            0                  0                  0
## 5430:            0                  1                  0
## 5431:            0                  1                  0
## 5432:            5                  1                  1
##          ind_CNS_infection ind_Viral_infection disruptpive damaging
## 1:            0                  0                  0

```

```

##      2:          0          0          0          1
##      3:          0          0          0          2
##      4:          0          0          NA         NA
##      5:          0          0          NA         NA
##      ---
## 5428:          0          0          0          0
## 5429:          0          0          NA         NA
## 5430:          0          0          0          0
## 5431:          0          1          0          0
## 5432:          0          0          0          0
##      disruptive_and_damaging synonymous age_Otitis_infection
##      1:          0          0          22.1
##      2:          1          2          23.5
##      3:          2          2          25.8
##      4:          NA         NA          28.1
##      5:          NA         NA          29.2
##      ---
## 5428:          0          0          19.3
## 5429:          NA        NA          27.3
## 5430:          0          2          8.9
## 5431:          0          3          4.4
## 5432:          0          0          1.3
##      age_Bacterial_infection age_CNS_infection age_Viral_infection
##      1:          22.1         22         22.1
##      2:          23.5         23         23.5
##      3:          25.8         26         25.8
##      4:          28.1         28         28.1
##      5:          24.5         29         29.2
##      ---
## 5428:          19.3         19         19.3
## 5429:          27.3         27         27.3
## 5430:          21.8         22         21.8
## 5431:          14.3         14         4.2
## 5432:          1.3          30         30.4
##      B_SECTIOU B_I11 ind_Fxxxx_m ind_F2100_m ind_Fxxxx_f ind_F2100_f
##      1:       NA     0      1      0      0      0
##      2:       NA     0      1      0      0      0
##      3:       NA     0      0      0      1      0
##      4:       NA     0      0      0      0      0
##      5:       NA     0      1      0      0      0
##      ---
## 5428:       NA    NA     0      0      0      0
## 5429:       NA    0      0      0      0      0
## 5430:       1    NA     0      0      0      0
## 5431:       NA    NA     0      0      1      0
## 5432:       NA    0     NA     NA     NA     NA
##      RYGER contacts visits earl_F10
##      1:    NA     8    69     0
##      2:    NA     0     0     0
##      3:    NA     0   114     0
##      4:    NA    44    69     0
##      5:    NA     6     0     0
##      ---
## 5428:    0    13    80     0

```

```

## 5429:   NA      0     77      0
## 5430:   1      8    103      0
## 5431:   1      0     89      0
## 5432:   NA     9     90      0

data2[F10<median(F10,na.rm = T) ,earl_F10 := 1]

##          pid year gender birthdate case case2012 random F10 F20 F30 F40
## 1: 1000166 1990       F 1990-12-13     1      1     NA     NA     NA 21  NA
## 2: 1002947 1989       M 1989-07-15     1      1     NA     17     NA     NA 21  NA
## 3: 1009068 1987       M 1987-03-03     1      1     NA     NA     NA 21  NA
## 4: 1013507 1984       F 1984-11-21     1      1     NA     NA     NA 16  NA
## 5: 1014243 1983       F 1983-10-20     1      1     NA     NA 21  20  NA
## ---
## 5428: 9315714 1993       M 1993-09-19     1      1     1    17    18  NA 23
## 5429: 9316160 1985       M 1985-09-08     1      1     NA    18    18  NA  NA
## 5430: 9319643 1991       F 1991-03-02     1      1     NA    20    NA  NA 19
## 5431: 9320666 1998       F 1998-09-09     1      1     NA     NA     NA  NA  NA
## 5432: 9323646 1982       M 1982-08-08     1      1     NA    24    27  NA  NA
##          F50 F60 F70 F84 F90 SCZ missing censored N pop_weights mcov Dim1
## 1:   NA  NA  NA  NA  NA 21     26     26 634      1     2 -2.45
## 2:   NA 18  24  NA 17 19     27     27 617      1    12  6.02
## 3:   NA 21  NA  NA  NA 21     30     30 585      1    19  0.85
## 4:  16 16  NA  NA  NA 18     32     32 525      1    25  6.36
## 5:  20 20  NA  NA 32 22     33     33 480      1    26  3.52
## ---
## 5428:  NA  NA  NA  NA 23 18     23     23 607      1 5682  3.16
## 5429:  NA  NA  NA 14 18 20     31     31 592      1 5683  6.96
## 5430:  NA  NA  NA  NA 19 26     26     26 613      1 5687  1.21
## 5431:  NA  NA  NA  NA 14 18     18     18 613      1   17 -1.61
## 5432:  NA  NA  NA  NA 27 34     34     34 492      1 322 -5.19
##          Dim2 Dim3 Dim4 Dim5 Dim6 Dim7 n_dia age_min dia_year fit
## 1:  1.70 2.558 -1.86 -0.18 -1.764 -0.616     1    21 2012  0.70
## 2:  1.00 -3.387  1.89 -0.67  0.355 -1.845     4    17 2008  0.20
## 3:  2.86 1.243 -1.62 -2.22  0.956  1.532     2    21 2008 -0.64
## 4:  3.86 2.045  0.73 -0.92  1.240 -1.178     3    16 2003 -1.44
## 5:  3.92 0.121  1.00 -1.39  0.152 -0.140     4    20 2005 -1.83
## ---
## 5428: -0.41 -3.712 -0.61  1.25 -0.391  0.241     7    17 2011  1.67
## 5429: -3.39 -1.517  0.46  3.53  0.487 -0.078     3    14 2005 -1.16
## 5430:  0.77 -2.695 -1.11  1.50  1.454  2.479     2    19 2010  0.77
## 5431:  0.52 -0.012  0.83 -0.37  0.045  0.584     0    14 2012  3.43
## 5432: -0.93 -2.280 -0.28  1.19 -0.237 -0.908     1    24 2009 -2.25
##          maternal_age paternal_age langde apgar5 gest_age bw_score C_RYGER
## 1:           16            23     50     10     38    75.9     NA
## 2:            22            33     54     10     39    98.4     NA
## 3:            32            51     54     10     42    15.0     NA
## 4:            30            29     50     10     38    63.3     NA
## 5:            31            34     49     NA     37    57.3     NA
## ---
## 5428:           32            25     52     10     39    46.4     NA
## 5429:           18            21     54     10     40    64.7     NA
## 5430:           32            24     46     10     35    45.7     NA
## 5431:           28            36     52     10     38    77.3      0

```

```

## 5432:          18          NA         45        10        36      8.9       NA
##           B_RYGER ind_mat_preg_Bacterial_infection
##   1:      NA                      0
##   2:      NA                      0
##   3:      NA                      0
##   4:      NA                      0
##   5:      NA                      0
##   --- 
## 5428:      0                      0
## 5429:     NA                     0
## 5430:      1                     0
## 5431:     NA                     0
## 5432:     NA                     0
##           ind_mat_preg_Viral_infection AN_2016_SCORES_AT_ALL_THRESHOLDS_05
##   1:                      0                  -0.059
##   2:                      0                  -0.059
##   3:                      0                   NA
##   4:                      0                   NA
##   5:                      0                   NA
##   --- 
## 5428:                      0                  -0.059
## 5429:                      0                   NA
## 5430:                      0                  -0.059
## 5431:                      0                  -0.059
## 5432:                      0                  -0.059
##           Anxiety_SCORES_AT_ALL_THRESHOLDS_05
##   1:                 -0.00045
##   2:                 -0.00051
##   3:                   NA
##   4:                   NA
##   5:                   NA
##   --- 
## 5428:                 -0.00058
## 5429:                   NA
## 5430:                 -0.00050
## 5431:                 -0.00054
## 5432:                 -0.00046
##           BMI_2015_SCORES_AT_ALL_THRESHOLDS_05
##   1:                 -0.00045
##   2:                 -0.00042
##   3:                   NA
##   4:                   NA
##   5:                   NA
##   --- 
## 5428:                 -0.00042
## 5429:                   NA
## 5430:                 -0.00034
## 5431:                 -0.00041
## 5432:                 -0.00033
##           BIP_2016_SCORES_AT_ALL_THRESHOLDS_05 BW_SCORES_AT_ALL_THRESHOLDS_05
##   1:                      0.423                0.0029
##   2:                     -0.779                0.0032
##   3:                   NA                   NA
##   4:                   NA                   NA

```

```

##      5:                               NA                               NA
##    ---
## 5428:                               0.031                           0.0031
## 5429:                               NA                             NA
## 5430:                               0.532                           0.0030
## 5431:                               -2.547                          0.0031
## 5432:                               1.358                           0.0031
##      Cannib_SCORES_AT_ALL_THRESHOLDS_05
##      1:                               -0.10
##      2:                               -2.71
##      3:                               NA
##      4:                               NA
##      5:                               NA
##    ---
## 5428:                               -0.81
## 5429:                               NA
## 5430:                               -0.69
## 5431:                               -1.37
## 5432:                               0.91
##      DS_2016_SCORES_AT_ALL_THRESHOLDS_05
##      1:                               -0.201
##      2:                               0.598
##      3:                               NA
##      4:                               NA
##      5:                               NA
##    ---
## 5428:                               -0.107
## 5429:                               NA
## 5430:                               0.534
## 5431:                               0.714
## 5432:                               -0.081
##      EduYears_2016_SCORES_AT_ALL_THRESHOLDS_05
##      1:                               -0.682
##      2:                               -0.624
##      3:                               NA
##      4:                               NA
##      5:                               NA
##    ---
## 5428:                               -0.417
## 5429:                               NA
## 5430:                               -0.324
## 5431:                               -1.766
## 5432:                               0.065
##      GPC_Ext_SCORES_AT_ALL_THRESHOLDS_05
##      1:                               -0.000071
##      2:                               -0.000149
##      3:                               NA
##      4:                               NA
##      5:                               NA
##    ---
## 5428:                               -0.000068
## 5429:                               NA
## 5430:                               -0.000102
## 5431:                               -0.000082

```

```

## 5432: -0.000069
##      NEURO_2016_SCORES_AT_ALL_THRESHOLDS_05
## 1: 0.60
## 2: -1.48
## 3: NA
## 4: NA
## 5: NA
## ---
## 5428: 1.86
## 5429: NA
## 5430: -0.14
## 5431: -1.20
## 5432: 0.96
##      SCZ_2015_SCORES_AT_ALL_THRESHOLDS_05
## 1: -0.80
## 2: 0.35
## 3: NA
## 4: NA
## 5: NA
## ---
## 5428: 1.14
## 5429: NA
## 5430: -0.37
## 5431: -0.19
## 5432: 1.51
##      SWB_2016_SCORES_AT_ALL_THRESHOLDS_05 C1     C2     C3     C4
## 1: -0.40 -0.0027 -0.0009 0.0034 0.0073
## 2: 0.66 -0.0035 0.0085 -0.0021 -0.0029
## 3: NA     NA     NA     NA     NA
## 4: NA     NA     NA     NA     NA
## 5: NA     NA     NA     NA     NA
## ---
## 5428: -0.13 -0.0051 -0.0007 0.0051 -0.0001
## 5429: NA     NA     NA     NA     NA
## 5430: -0.66 0.0067 -0.0022 -0.0004 -0.0033
## 5431: 0.30 0.0075 -0.0035 0.0027 0.0038
## 5432: 1.66 0.0002 -0.0014 0.0043 -0.0019
##      C5     C6     C7     C8     C9     C10    wave
## 1: -0.0030 0.0000 0.0065 -0.0011 -0.0004 0.0005 Wave13
## 2: -0.0010 -0.0049 0.0020 0.0064 0.0046 0.0027 Wave14
## 3: NA     NA     NA     NA     NA     NA     Other
## 4: NA     NA     NA     NA     NA     NA     Other
## 5: NA     NA     NA     NA     NA     NA     Other
## ---
## 5428: 0.0031 -0.0008 0.0060 0.0078 0.0010 0.0034 Wave17
## 5429: NA     NA     NA     NA     NA     NA     Other
## 5430: 0.0048 -0.0028 -0.0016 0.0001 0.0043 0.0046 Wave10
## 5431: -0.0105 -0.0024 -0.0025 -0.0050 -0.0038 -0.0037 Wave4_
## 5432: 0.0000 0.0019 -0.0006 0.0054 0.0001 -0.0057 Wave20
##      d2100_ptype0_contacts d2100_ptype0_days d2100_ptype1_contacts
## 1: 7             234                0
## 2: 0             0                  0
## 3: 0             0                  0
## 4: 32            645                0

```

```

##      5:          2          65          0
##      ---
## 5428:          9         433          0
## 5429:          0          0          0
## 5430:          8         120          0
## 5431:          0          0          0
## 5432:          4         65          0
##      d2100_ptype1_visits d2100_ptype2_visits d2100_ptype2_novisits
##      1:          0          69          0
##      2:          0          0          0
##      3:          0         114          0
##      4:          0          69          0
##      5:          0          0          0
##      ---
## 5428:          0          80          0
## 5429:          0          77          0
## 5430:          0         103          0
## 5431:          0          89          0
## 5432:          0         90          1
##      d2100_ptype3_contacts ind_Otitis_infection ind_Bacterial_infection
##      1:          1          0          0
##      2:          0          0          0
##      3:          0          0          0
##      4:         12          0          0
##      5:          4          0          1
##      ---
## 5428:          4          0          0
## 5429:          0          0          0
## 5430:          0          1          0
## 5431:          0          1          0
## 5432:          5          1          1
##      ind_CNS_infection ind_Viral_infection disruptpive damaging
##      1:          0          0          0          0
##      2:          0          0          0          1
##      3:          0          0          0          2
##      4:          0          0          NA          NA
##      5:          0          0          NA          NA
##      ---
## 5428:          0          0          0          0
## 5429:          0          0          NA          NA
## 5430:          0          0          0          0
## 5431:          0          1          0          0
## 5432:          0          0          0          0
##      disruptive_and_damaging synonymous age_Otitis_infection
##      1:          0          0        22.1
##      2:          1          2        23.5
##      3:          2          2        25.8
##      4:         NA         NA        28.1
##      5:         NA         NA        29.2
##      ---
## 5428:          0          0        19.3
## 5429:         NA         NA        27.3
## 5430:          0          2         8.9
## 5431:          0          3         4.4

```

```

## 5432:          0          0          1.3
##      age_Bacterial_infection age_CNS_infection age_Viral_infection
## 1:            22.1            22            22.1
## 2:            23.5            23            23.5
## 3:            25.8            26            25.8
## 4:            28.1            28            28.1
## 5:            24.5            29            29.2
## ---
## 5428:          19.3           19           19.3
## 5429:          27.3           27           27.3
## 5430:          21.8           22           21.8
## 5431:          14.3           14            4.2
## 5432:          1.3            30           30.4
##      B_SECTIOU B_I11 ind_Fxxxx_m ind_F2100_m ind_Fxxxx_f ind_F2100_f
## 1:       NA     0         1         0         0         0
## 2:       NA     0         1         0         0         0
## 3:       NA     0         0         0         1         0
## 4:       NA     0         0         0         0         0
## 5:       NA     0         1         0         0         0
## ---
## 5428:       NA    NA         0         0         0         0
## 5429:       NA    0         0         0         0         0
## 5430:        1    NA         0         0         0         0
## 5431:       NA    NA         0         0         1         0
## 5432:       NA    0        NA        NA        NA        NA
##      RYGER contacts visits earl_F10
## 1:   NA      8     69      0
## 2:   NA      0      0      1
## 3:   NA      0    114      0
## 4:   NA     44     69      0
## 5:   NA      6      0      0
## ---
## 5428:    0     13     80      1
## 5429:   NA     0     77      1
## 5430:    1      8    103      1
## 5431:    1      0     89      0
## 5432:   NA     9     90      0

summary(glm(earl_F10~SCZ_2015_SCORES_AT_ALL_THRESHOLDS_05+gender+birthdate+C1+C2+C3+C4+C5+C6+C7+C8+C9+C10))

## [1] 0.0083

exp(confint(glm(earl_F10~scale(SCZ_2015_SCORES_AT_ALL_THRESHOLDS_05)+gender+birthdate+C1+C2+C3+C4+C5+C6+C7+C8+C9+C10)))

## Waiting for profiling to be done...

## 2.5 % 97.5 %
## 1.0    1.3

```

Dim3 ~ Maternal and Paternal Age

```
summary(lm(Dim3~maternal_age+paternal_age+gender+birthdate,data=data2))$coefficients[2:3,4]
```

```
## maternal_age paternal_age  
##      0.00076      0.26220
```

Dim3 ~ Maternal comparing to random sample:

```
summary(lm(maternal_age~as.numeric(!is.na(case))+gender+birthdate,data=dt_pg[pid %in% dt_mds[Dim3<median
```

```
##  
## Call:  
## lm(formula = maternal_age ~ as.numeric(!is.na(case)) + gender +  
##     birthdate, data = dt_pg[pid %in% dt_mds[Dim3 < median(Dim3),  
##     pid] | random == 1])  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -14.553  -3.381  -0.288   3.093  22.421  
##  
## Coefficients:  
##                               Estimate Std. Error t value Pr(>|t|)  
## (Intercept)            24.7252244  0.1205287 205.14    <2e-16 ***  
## as.numeric(!is.na(case)) -0.8962732  0.1062991  -8.43    <2e-16 ***  
## genderM                -0.0564889  0.0583207   -0.97    0.33  
## birthdate               0.0004220  0.0000139   30.33    <2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 4.8 on 27132 degrees of freedom  
## (538 observations deleted due to missingness)  
## Multiple R-squared:  0.04,   Adjusted R-squared:  0.0399  
## F-statistic:  377 on 3 and 27132 DF,  p-value: <2e-16
```

```
summary(lm(maternal_age~as.numeric(!is.na(case))+gender+birthdate,data=dt_pg[pid %in% dt_mds[Dim3>quantile(dt
```

```
##  
## Call:  
## lm(formula = maternal_age ~ as.numeric(!is.na(case)) + gender +  
##     birthdate, data = dt_pg[pid %in% dt_mds[Dim3 > quantile(Dim3,  
##     0.75), pid] | random == 1])  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -14.543  -3.364  -0.275   3.095  20.145  
##  
## Coefficients:  
##                               Estimate Std. Error t value Pr(>|t|)  
## (Intercept)            24.7567024  0.1204704 205.50    <2e-16 ***  
## as.numeric(!is.na(case)) -0.1423132  0.1310627  -1.09    0.28  
## genderM                -0.0436566  0.0588447   -0.74    0.46  
## birthdate               0.0004173  0.0000139   29.97    <2e-16 ***
```

```

## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.8 on 26247 degrees of freedom
##   (104 observations deleted due to missingness)
## Multiple R-squared: 0.0341, Adjusted R-squared: 0.034
## F-statistic: 309 on 3 and 26247 DF, p-value: <2e-16

Dim1 ~ Infections

anova(lm(Dim1~ind_Bacterial_infection+gender+birthdate,data=data2))$"Pr(>F)"[1]

## [1] 6.1e-07

anova(lm(Dim1~ind_Bacterial_infection+gender+birthdate+contacts+visits,data=data2))$"Pr(>F)"[1]

## [1] 3.7e-07

anova(lm(Dim1~ind_Viral_infection+gender+birthdate,data=data2))$"Pr(>F)"[1]

## [1] 0.000012

summary(lm(Dim1~ind_Viral_infection+gender+birthdate+contacts+visits,data=data2))$"Pr(>F)"[1]

## NULL

```

Excluding infections after onset of first psychiatric:

```

data2[,Bac_after := 0]

##          pid year gender birthdate case case2012 random F10 F20 F30 F40
## 1: 1000166 1990      F 1990-12-13    1       1     NA  NA  NA 21  NA
## 2: 1002947 1989      M 1989-07-15    1       1     NA 17  NA  NA  NA
## 3: 1009068 1987      M 1987-03-03    1       1     NA  NA  NA 21  NA
## 4: 1013507 1984      F 1984-11-21    1       1     NA  NA  NA 16  NA
## 5: 1014243 1983      F 1983-10-20    1       1     NA  NA 21 20  NA
##   ---
## 5428: 9315714 1993      M 1993-09-19    1       1     1 17 18  NA 23
## 5429: 9316160 1985      M 1985-09-08    1       1     NA 18 18  NA  NA
## 5430: 9319643 1991      F 1991-03-02    1       1     NA 20  NA  NA 19
## 5431: 9320666 1998      F 1998-09-09    1       1     NA  NA  NA  NA  NA
## 5432: 9323646 1982      M 1982-08-08    1       1     NA 24 27  NA  NA
##          F50 F60 F70 F84 F90 SCZ missing censored N pop_weights mcov Dim1
## 1:    NA  NA  NA  NA  NA 21    26     26 634           1    2 -2.45
## 2:    NA 18 24  NA 17 19    27     27 617           1   12  6.02
## 3:    NA 21  NA  NA  NA 21    30     30 585           1   19  0.85
## 4:   16 16  NA  NA  NA 18    32     32 525           1   25  6.36
## 5:   20 20  NA  NA 32 22    33     33 480           1   26  3.52
##   ---
## 5428:   NA  NA  NA  NA 23 18    23     23 607           1 5682  3.16

```

```

## 5429: NA NA NA 14 18 20 31 31 592 1 5683 6.96
## 5430: NA NA NA NA NA 19 26 26 613 1 5687 1.21
## 5431: NA NA NA NA NA 14 18 18 613 1 17 -1.61
## 5432: NA NA NA NA NA 27 34 34 492 1 322 -5.19
## Dim2 Dim3 Dim4 Dim5 Dim6 Dim7 n_dia age_min dia_year fit
## 1: 1.70 2.558 -1.86 -0.18 -1.764 -0.616 1 21 2012 0.70
## 2: 1.00 -3.387 1.89 -0.67 0.355 -1.845 4 17 2008 0.20
## 3: 2.86 1.243 -1.62 -2.22 0.956 1.532 2 21 2008 -0.64
## 4: 3.86 2.045 0.73 -0.92 1.240 -1.178 3 16 2003 -1.44
## 5: 3.92 0.121 1.00 -1.39 0.152 -0.140 4 20 2005 -1.83
## ---
## 5428: -0.41 -3.712 -0.61 1.25 -0.391 0.241 7 17 2011 1.67
## 5429: -3.39 -1.517 0.46 3.53 0.487 -0.078 3 14 2005 -1.16
## 5430: 0.77 -2.695 -1.11 1.50 1.454 2.479 2 19 2010 0.77
## 5431: 0.52 -0.012 0.83 -0.37 0.045 0.584 0 14 2012 3.43
## 5432: -0.93 -2.280 -0.28 1.19 -0.237 -0.908 1 24 2009 -2.25
## maternal_age paternal_age langde apgar5 gest_age bw_score C_RYGER
## 1: 16 23 50 10 38 75.9 NA
## 2: 22 33 54 10 39 98.4 NA
## 3: 32 51 54 10 42 15.0 NA
## 4: 30 29 50 10 38 63.3 NA
## 5: 31 34 49 NA 37 57.3 NA
## ---
## 5428: 32 25 52 10 39 46.4 NA
## 5429: 18 21 54 10 40 64.7 NA
## 5430: 32 24 46 10 35 45.7 NA
## 5431: 28 36 52 10 38 77.3 0
## 5432: 18 NA 45 10 36 8.9 NA
## B_RYGER ind_mat_preg_Bacterial_infection
## 1: NA 0
## 2: NA 0
## 3: NA 0
## 4: NA 0
## 5: NA 0
## ---
## 5428: 0 0
## 5429: NA 0
## 5430: 1 0
## 5431: NA 0
## 5432: NA 0
## ind_mat_preg_Viral_infection AN_2016_SCORES_AT_ALL_THRESHOLDS_05
## 1: 0 -0.059
## 2: 0 -0.059
## 3: 0 NA
## 4: 0 NA
## 5: 0 NA
## ---
## 5428: 0 -0.059
## 5429: 0 NA
## 5430: 0 -0.059
## 5431: 0 -0.059
## 5432: 0 -0.059
## Anxiety_SCORES_AT_ALL_THRESHOLDS_05
## 1: -0.00045

```

```

##      2:           -0.00051
##      3:             NA
##      4:             NA
##      5:             NA
##    ---
## 5428:           -0.00058
## 5429:             NA
## 5430:           -0.00050
## 5431:           -0.00054
## 5432:           -0.00046
##      BMI_2015_SCORES_AT_ALL_THRESHOLDS_05
##      1:           -0.00045
##      2:           -0.00042
##      3:             NA
##      4:             NA
##      5:             NA
##    ---
## 5428:           -0.00042
## 5429:             NA
## 5430:           -0.00034
## 5431:           -0.00041
## 5432:           -0.00033
##      BIP_2016_SCORES_AT_ALL_THRESHOLDS_05 BW_SCORES_AT_ALL_THRESHOLDS_05
##      1:           0.423            0.0029
##      2:           -0.779           0.0032
##      3:             NA             NA
##      4:             NA             NA
##      5:             NA             NA
##    ---
## 5428:           0.031            0.0031
## 5429:             NA             NA
## 5430:           0.532            0.0030
## 5431:           -2.547           0.0031
## 5432:           1.358            0.0031
##      Cannib SCORES_AT_ALL_THRESHOLDS_05
##      1:           -0.10
##      2:           -2.71
##      3:             NA
##      4:             NA
##      5:             NA
##    ---
## 5428:           -0.81
## 5429:             NA
## 5430:           -0.69
## 5431:           -1.37
## 5432:           0.91
##      DS_2016_SCORES_AT_ALL_THRESHOLDS_05
##      1:           -0.201
##      2:           0.598
##      3:             NA
##      4:             NA
##      5:             NA
##    ---
## 5428:           -0.107

```

```

## 5429:                               NA
## 5430:                           0.534
## 5431:                           0.714
## 5432:                          -0.081
##      EduYears_2016_SCORES_AT_ALL_THRESHOLDS_05
## 1:                                -0.682
## 2:                                -0.624
## 3:                                 NA
## 4:                                 NA
## 5:                                 NA
##    ---
## 5428:                           -0.417
## 5429:                               NA
## 5430:                           -0.324
## 5431:                          -1.766
## 5432:                           0.065
##      GPC_Ext_SCORES_AT_ALL_THRESHOLDS_05
## 1:                         -0.000071
## 2:                         -0.000149
## 3:                                 NA
## 4:                                 NA
## 5:                                 NA
##    ---
## 5428:                           -0.000068
## 5429:                               NA
## 5430:                         -0.000102
## 5431:                         -0.000082
## 5432:                         -0.000069
##      NEURO_2016_SCORES_AT_ALL_THRESHOLDS_05
## 1:                                0.60
## 2:                               -1.48
## 3:                                 NA
## 4:                                 NA
## 5:                                 NA
##    ---
## 5428:                               1.86
## 5429:                               NA
## 5430:                           -0.14
## 5431:                           -1.20
## 5432:                           0.96
##      SCZ_2015_SCORES_AT_ALL_THRESHOLDS_05
## 1:                                -0.80
## 2:                                 0.35
## 3:                                 NA
## 4:                                 NA
## 5:                                 NA
##    ---
## 5428:                               1.14
## 5429:                               NA
## 5430:                           -0.37
## 5431:                           -0.19
## 5432:                           1.51
##      SWB_2016_SCORES_AT_ALL_THRESHOLDS_05      C1      C2      C3      C4
## 1:          -0.40 -0.0027 -0.0009  0.0034  0.0073

```

```

##      2:          0.66 -0.0035  0.0085 -0.0021 -0.0029
##      3:             NA      NA      NA      NA      NA
##      4:             NA      NA      NA      NA      NA
##      5:             NA      NA      NA      NA      NA
##      ---
## 5428:          -0.13 -0.0051 -0.0007  0.0051 -0.0001
## 5429:             NA      NA      NA      NA      NA
## 5430:          -0.66  0.0067 -0.0022 -0.0004 -0.0033
## 5431:           0.30  0.0075 -0.0035  0.0027  0.0038
## 5432:           1.66  0.0002 -0.0014  0.0043 -0.0019
##      C5      C6      C7      C8      C9      C10    wave
##      1: -0.0030  0.0000  0.0065 -0.0011 -0.0004  0.0005 Wave13
##      2: -0.0010 -0.0049  0.0020  0.0064  0.0046  0.0027 Wave14
##      3:     NA     NA     NA     NA     NA     NA Other
##      4:     NA     NA     NA     NA     NA     NA Other
##      5:     NA     NA     NA     NA     NA     NA Other
##      ---
## 5428:  0.0031 -0.0008  0.0060  0.0078  0.0010  0.0034 Wave17
## 5429:     NA     NA     NA     NA     NA     NA Other
## 5430:  0.0048 -0.0028 -0.0016  0.0001  0.0043  0.0046 Wave10
## 5431: -0.0105 -0.0024 -0.0025 -0.0050 -0.0038 -0.0037 Wave4_
## 5432:  0.0000  0.0019 -0.0006  0.0054  0.0001 -0.0057 Wave20
##      d2100_ptype0_contacts d2100_ptype0_days d2100_ptype1_contacts
##      1:                 7        234          0
##      2:                 0         0          0
##      3:                 0         0          0
##      4:                32        645          0
##      5:                 2         65          0
##      ---
## 5428:                 9        433          0
## 5429:                 0         0          0
## 5430:                 8        120          0
## 5431:                 0         0          0
## 5432:                 4         65          0
##      d2100_ptype1_visits d2100_ptype2_visits d2100_ptype2_novisits
##      1:                 0         69          0
##      2:                 0         0          0
##      3:                 0        114          0
##      4:                 0         69          0
##      5:                 0         0          0
##      ---
## 5428:                 0         80          0
## 5429:                 0         77          0
## 5430:                 0        103          0
## 5431:                 0         89          0
## 5432:                 0         90          1
##      d2100_ptype3_contacts ind_Otitis_infection ind_Bacterial_infection
##      1:                 1          0          0
##      2:                 0          0          0
##      3:                 0          0          0
##      4:                12          0          0
##      5:                 4          0          1
##      ---
## 5428:                 4          0          0

```

```

## 5429:          0          0          0
## 5430:          0          1          0
## 5431:          0          1          0
## 5432:         5          1          1
##      ind_CNS_infection ind_Viral_infection disruptvive damaging
## 1:              0          0          0          0
## 2:              0          0          0          1
## 3:              0          0          0          2
## 4:              0          0          NA          NA
## 5:              0          0          NA          NA
## ---
## 5428:          0          0          0          0
## 5429:          0          0          NA          NA
## 5430:          0          0          0          0
## 5431:          0          1          0          0
## 5432:          0          0          0          0
##      disruptive_and_damaging synonymous age_Otitis_infection
## 1:              0          0        22.1
## 2:              1          2        23.5
## 3:              2          2        25.8
## 4:             NA          NA        28.1
## 5:             NA          NA        29.2
## ---
## 5428:          0          0        19.3
## 5429:         NA          NA        27.3
## 5430:          0          2         8.9
## 5431:          0          3         4.4
## 5432:          0          0         1.3
##      age_Bacterial_infection age_CNS_infection age_Viral_infection
## 1:        22.1          22        22.1
## 2:        23.5          23        23.5
## 3:        25.8          26        25.8
## 4:        28.1          28        28.1
## 5:        24.5          29        29.2
## ---
## 5428:        19.3          19        19.3
## 5429:        27.3          27        27.3
## 5430:        21.8          22        21.8
## 5431:        14.3          14         4.2
## 5432:        1.3           30        30.4
##      B_SECTIOU B_I11 ind_Fxxxx_m ind_F2100_m ind_Fxxxx_f ind_F2100_f
## 1:     NA     0      1      0      0      0
## 2:     NA     0      1      0      0      0
## 3:     NA     0      0      0      1      0
## 4:     NA     0      0      0      0      0
## 5:     NA     0      1      0      0      0
## ---
## 5428:    NA    NA     0     0     0     0
## 5429:    NA    0     0     0     0     0
## 5430:    1    NA     0     0     0     0
## 5431:   NA   NA     0     0     1     0
## 5432:   NA    0    NA    NA    NA    NA
##      RYGER contacts visits earl_F10 Bac_after
## 1:    NA      8     69      0      0

```

```

##   2:    NA     0     0     1     0
##   3:    NA     0    114     0     0
##   4:    NA    44     69     0     0
##   5:    NA     6     0     0     0
##   ---
## 5428:    0    13    80     1     0
## 5429:   NA     0    77     1     0
## 5430:    1     8   103     1     0
## 5431:    1     0    89     0     0
## 5432:   NA     9    90     0     0

data2[ind_Bacterial_infection==1, Bac_after := as.numeric(age_Bacterial_infection>= age_min)]
```

	pid	year	gender	birthdate	case	case2012	random	F10	F20	F30	F40	
## 1:	1000166	1990	F	1990-12-13	1	1	NA	NA	NA	21	NA	
## 2:	1002947	1989	M	1989-07-15	1	1	NA	17	NA	NA	NA	
## 3:	1009068	1987	M	1987-03-03	1	1	NA	NA	NA	21	NA	
## 4:	1013507	1984	F	1984-11-21	1	1	NA	NA	NA	16	NA	
## 5:	1014243	1983	F	1983-10-20	1	1	NA	NA	21	20	NA	
## ---												
## 5428:	9315714	1993	M	1993-09-19	1	1	1	17	18	NA	23	
## 5429:	9316160	1985	M	1985-09-08	1	1	NA	18	18	NA	NA	
## 5430:	9319643	1991	F	1991-03-02	1	1	NA	20	NA	NA	19	
## 5431:	9320666	1998	F	1998-09-09	1	1	NA	NA	NA	NA	NA	
## 5432:	9323646	1982	M	1982-08-08	1	1	NA	24	27	NA	NA	
##												
##	F50	F60	F70	F84	F90	SCZ	missing	censored	N	pop_weights	mcov	Dim1
## 1:	NA	NA	NA	NA	NA	21	26	26	634		1	2 -2.45
## 2:	NA	18	24	NA	17	19	27	27	617		1	12 6.02
## 3:	NA	21	NA	NA	NA	21	30	30	585		1	19 0.85
## 4:	16	16	NA	NA	NA	18	32	32	525		1	25 6.36
## 5:	20	20	NA	NA	32	22	33	33	480		1	26 3.52
## ---												
## 5428:	NA	NA	NA	NA	23	18	23	23	607		1	5682 3.16
## 5429:	NA	NA	NA	14	18	20	31	31	592		1	5683 6.96
## 5430:	NA	NA	NA	NA	NA	19	26	26	613		1	5687 1.21
## 5431:	NA	NA	NA	NA	NA	14	18	18	613		1	17 -1.61
## 5432:	NA	NA	NA	NA	NA	27	34	34	492		1	322 -5.19
##												
##	Dim2	Dim3	Dim4	Dim5	Dim6	Dim7	n_dia	age_min	dia_year	fit		
## 1:	1.70	2.558	-1.86	-0.18	-1.764	-0.616	1	21	2012	0.70		
## 2:	1.00	-3.387	1.89	-0.67	0.355	-1.845	4	17	2008	0.20		
## 3:	2.86	1.243	-1.62	-2.22	0.956	1.532	2	21	2008	-0.64		
## 4:	3.86	2.045	0.73	-0.92	1.240	-1.178	3	16	2003	-1.44		
## 5:	3.92	0.121	1.00	-1.39	0.152	-0.140	4	20	2005	-1.83		
## ---												
## 5428:	-0.41	-3.712	-0.61	1.25	-0.391	0.241	7	17	2011	1.67		
## 5429:	-3.39	-1.517	0.46	3.53	0.487	-0.078	3	14	2005	-1.16		
## 5430:	0.77	-2.695	-1.11	1.50	1.454	2.479	2	19	2010	0.77		
## 5431:	0.52	-0.012	0.83	-0.37	0.045	0.584	0	14	2012	3.43		
## 5432:	-0.93	-2.280	-0.28	1.19	-0.237	-0.908	1	24	2009	-2.25		
##												
##	maternal_age	paternal_age	langde	apgar5	gest_age	bw_score	C_RYGER					
## 1:		16		23	50	10	38	75.9		NA		
## 2:		22		33	54	10	39	98.4		NA		
## 3:		32		51	54	10	42	15.0		NA		
## 4:		30		29	50	10	38	63.3		NA		

```

##      5:          31          34          49        NA        37      57.3        NA
##      ---
## 5428:          32          25          52        10        39      46.4        NA
## 5429:          18          21          54        10        40      64.7        NA
## 5430:          32          24          46        10        35      45.7        NA
## 5431:          28          36          52        10        38      77.3         0
## 5432:          18         NA          45        10        36       8.9        NA
##      B_RYGER ind_mat_preg_Bacterial_infection
##      1:        NA          0
##      2:        NA          0
##      3:        NA          0
##      4:        NA          0
##      5:        NA          0
##      ---
## 5428:          0          0
## 5429:        NA          0
## 5430:          1          0
## 5431:        NA          0
## 5432:        NA          0
##      ind_mat_preg_Viral_infection AN_2016_SCORES_AT_ALL_THRESHOLDS_05
##      1:          0        -0.059
##      2:          0        -0.059
##      3:          0          NA
##      4:          0          NA
##      5:          0          NA
##      ---
## 5428:          0        -0.059
## 5429:          0          NA
## 5430:          0        -0.059
## 5431:          0        -0.059
## 5432:          0        -0.059
##      Anxiety_SCORES_AT_ALL_THRESHOLDS_05
##      1:      -0.00045
##      2:      -0.00051
##      3:          NA
##      4:          NA
##      5:          NA
##      ---
## 5428:      -0.00058
## 5429:          NA
## 5430:      -0.00050
## 5431:      -0.00054
## 5432:      -0.00046
##      BMI_2015_SCORES_AT_ALL_THRESHOLDS_05
##      1:      -0.00045
##      2:      -0.00042
##      3:          NA
##      4:          NA
##      5:          NA
##      ---
## 5428:      -0.00042
## 5429:          NA
## 5430:      -0.00034
## 5431:      -0.00041

```

```

## 5432: -0.00033
##      BIP_2016_SCORES_AT_ALL_THRESHOLDS_05 BW_SCORES_AT_ALL_THRESHOLDS_05
## 1:          0.423          0.0029
## 2:         -0.779          0.0032
## 3:           NA            NA
## 4:           NA            NA
## 5:           NA            NA
## ---
## 5428:        0.031          0.0031
## 5429:           NA            NA
## 5430:        0.532          0.0030
## 5431:       -2.547          0.0031
## 5432:        1.358          0.0031
##      Cannib_SCORES_AT_ALL_THRESHOLDS_05
## 1:        -0.10
## 2:       -2.71
## 3:           NA
## 4:           NA
## 5:           NA
## ---
## 5428:        -0.81
## 5429:           NA
## 5430:        -0.69
## 5431:       -1.37
## 5432:        0.91
##      DS_2016_SCORES_AT_ALL_THRESHOLDS_05
## 1:        -0.201
## 2:        0.598
## 3:           NA
## 4:           NA
## 5:           NA
## ---
## 5428:        -0.107
## 5429:           NA
## 5430:        0.534
## 5431:        0.714
## 5432:       -0.081
##      EduYears_2016_SCORES_AT_ALL_THRESHOLDS_05
## 1:        -0.682
## 2:        -0.624
## 3:           NA
## 4:           NA
## 5:           NA
## ---
## 5428:        -0.417
## 5429:           NA
## 5430:        -0.324
## 5431:       -1.766
## 5432:        0.065
##      GPC_Ext_SCORES_AT_ALL_THRESHOLDS_05
## 1:       -0.000071
## 2:      -0.000149
## 3:           NA
## 4:           NA

```

```

##      5:                               NA
##    ---
## 5428:                           -0.000068
## 5429:                               NA
## 5430:                           -0.000102
## 5431:                           -0.000082
## 5432:                           -0.000069
##      NEURO_2016_SCORES_AT_ALL_THRESHOLDS_05
##      1:                               0.60
##      2:                           -1.48
##      3:                               NA
##      4:                               NA
##      5:                               NA
##    ---
## 5428:                           1.86
## 5429:                               NA
## 5430:                           -0.14
## 5431:                           -1.20
## 5432:                           0.96
##      SCZ_2015_SCORES_AT_ALL_THRESHOLDS_05
##      1:                           -0.80
##      2:                           0.35
##      3:                               NA
##      4:                               NA
##      5:                               NA
##    ---
## 5428:                           1.14
## 5429:                               NA
## 5430:                           -0.37
## 5431:                           -0.19
## 5432:                           1.51
##      SWB_2016_SCORES_AT_ALL_THRESHOLDS_05      C1      C2      C3      C4
##      1:           -0.40 -0.0027 -0.0009  0.0034  0.0073
##      2:           0.66 -0.0035  0.0085 -0.0021 -0.0029
##      3:           NA     NA     NA     NA     NA
##      4:           NA     NA     NA     NA     NA
##      5:           NA     NA     NA     NA     NA
##    ---
## 5428:                           -0.13 -0.0051 -0.0007  0.0051 -0.0001
## 5429:                               NA     NA     NA     NA     NA
## 5430:                           -0.66  0.0067 -0.0022 -0.0004 -0.0033
## 5431:                           0.30  0.0075 -0.0035  0.0027  0.0038
## 5432:                           1.66  0.0002 -0.0014  0.0043 -0.0019
##      C5      C6      C7      C8      C9      C10     wave
##      1: -0.0030  0.0000  0.0065 -0.0011 -0.0004  0.0005 Wave13
##      2: -0.0010 -0.0049  0.0020  0.0064  0.0046  0.0027 Wave14
##      3:     NA     NA     NA     NA     NA     NA Other
##      4:     NA     NA     NA     NA     NA     NA Other
##      5:     NA     NA     NA     NA     NA     NA Other
##    ---
## 5428:  0.0031 -0.0008  0.0060  0.0078  0.0010  0.0034 Wave17
## 5429:     NA     NA     NA     NA     NA     NA Other
## 5430:  0.0048 -0.0028 -0.0016  0.0001  0.0043  0.0046 Wave10
## 5431: -0.0105 -0.0024 -0.0025 -0.0050 -0.0038 -0.0037 Wave4_

```

```

## 5432: 0.0000 0.0019 -0.0006 0.0054 0.0001 -0.0057 Wave20
##      d2100_ptype0_contacts d2100_ptype0_days d2100_ptype1_contacts
## 1:          7           234           0
## 2:          0           0           0
## 3:          0           0           0
## 4:         32           645           0
## 5:          2           65           0
## ---
## 5428:          9           433           0
## 5429:          0           0           0
## 5430:          8           120           0
## 5431:          0           0           0
## 5432:          4           65           0
##      d2100_ptype1_visits d2100_ptype2_visits d2100_ptype2_novisits
## 1:          0           69           0
## 2:          0           0           0
## 3:          0           114           0
## 4:          0           69           0
## 5:          0           0           0
## ---
## 5428:          0           80           0
## 5429:          0           77           0
## 5430:          0           103           0
## 5431:          0           89           0
## 5432:          0           90           1
##      d2100_ptype3_contacts ind_Otitis_infection ind_Bacterial_infection
## 1:          1           0           0
## 2:          0           0           0
## 3:          0           0           0
## 4:         12           0           0
## 5:          4           0           1
## ---
## 5428:          4           0           0
## 5429:          0           0           0
## 5430:          0           1           0
## 5431:          0           1           0
## 5432:          5           1           1
##      ind_CNS_infection ind_Viral_infection disruptpive damaging
## 1:          0           0           0           0
## 2:          0           0           0           1
## 3:          0           0           0           2
## 4:          0           0           NA          NA
## 5:          0           0           NA          NA
## ---
## 5428:          0           0           0           0
## 5429:          0           0           NA          NA
## 5430:          0           0           0           0
## 5431:          0           1           0           0
## 5432:          0           0           0           0
##      disruptive_and_damaging synonymous age_Otitis_infection
## 1:          0           0           22.1
## 2:          1           2           23.5
## 3:          2           2           25.8
## 4:         NA          NA           28.1

```

```

##      5:                      NA          NA          29.2
##      ---
## 5428:                      0          0          19.3
## 5429:                     NA         NA          27.3
## 5430:                      0          2           8.9
## 5431:                      0          3           4.4
## 5432:                      0          0           1.3
##      age_Bacterial_infection age_CNS_infection age_Viral_infection
##      1:                  22.1                  22          22.1
##      2:                  23.5                  23          23.5
##      3:                  25.8                  26          25.8
##      4:                  28.1                  28          28.1
##      5:                  24.5                  29          29.2
##      ---
## 5428:                  19.3                  19          19.3
## 5429:                  27.3                  27          27.3
## 5430:                  21.8                  22          21.8
## 5431:                  14.3                  14           4.2
## 5432:                  1.3                  30          30.4
##      B_SECTIOU B_I11 ind_Fxxxx_m ind_F2100_m ind_Fxxxx_f ind_F2100_f
##      1:       NA     0        1        0        0        0
##      2:       NA     0        1        0        0        0
##      3:       NA     0        0        0        1        0
##      4:       NA     0        0        0        0        0
##      5:       NA     0        1        0        0        0
##      ---
## 5428:       NA    NA        0        0        0        0
## 5429:       NA    0        0        0        0        0
## 5430:       1    NA        0        0        0        0
## 5431:       NA    NA        0        0        1        0
## 5432:       NA    0       NA       NA       NA       NA
##      RYGER contacts visits earl_F10 Bac_after
##      1:    NA      8     69      0      0
##      2:    NA      0      0      1      0
##      3:    NA      0    114      0      0
##      4:    NA     44     69      0      0
##      5:    NA      6      0      0      1
##      ---
## 5428:     0     13     80      1      0
## 5429:    NA      0     77      1      0
## 5430:     1      8    103      1      0
## 5431:     1      0     89      0      0
## 5432:    NA      9     90      0      0

```

```
anova(lm(Dim1~ind_Bacterial_infection+gender+birthdate,data=data2 [Bac_after!=1]))
```

```

## Analysis of Variance Table
##
## Response: Dim1
##                               Df Sum Sq Mean Sq F value Pr(>F)
## ind_Bacterial_infection   1     6      6   0.38   0.54
## gender                     1 1421    1421  96.92 <2e-16 ***
## birthdate                  1 10030   10030  684.27 <2e-16 ***
## Residuals                 4966 72789     15

```

```

## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

data2[,Vir_after := 0]

##          pid year gender birthdate case case2012 random F10 F20 F30 F40
## 1: 1000166 1990      F 1990-12-13    1      1 NA NA NA 21 NA
## 2: 1002947 1989      M 1989-07-15    1      1 NA 17 NA NA NA
## 3: 1009068 1987      M 1987-03-03    1      1 NA NA NA 21 NA
## 4: 1013507 1984      F 1984-11-21    1      1 NA NA NA 16 NA
## 5: 1014243 1983      F 1983-10-20    1      1 NA NA 21 20 NA
## ---
## 5428: 9315714 1993      M 1993-09-19    1      1 17 18 NA 23
## 5429: 9316160 1985      M 1985-09-08    1      1 NA 18 18 NA NA
## 5430: 9319643 1991      F 1991-03-02    1      1 NA 20 NA NA 19
## 5431: 9320666 1998      F 1998-09-09    1      1 NA NA NA NA NA
## 5432: 9323646 1982      M 1982-08-08    1      1 NA 24 27 NA NA
##          F50 F60 F70 F84 F90 SCZ missing censored N pop_weights mcov Dim1
## 1: NA NA NA NA NA 21 26 26 634 1 2 -2.45
## 2: NA 18 24 NA 17 19 27 27 617 1 12 6.02
## 3: NA 21 NA NA NA 21 30 30 585 1 19 0.85
## 4: 16 16 NA NA NA 18 32 32 525 1 25 6.36
## 5: 20 20 NA NA 32 22 33 33 480 1 26 3.52
## ---
## 5428: NA NA NA NA 23 18 23 23 607 1 5682 3.16
## 5429: NA NA NA 14 18 20 31 31 592 1 5683 6.96
## 5430: NA NA NA NA NA 19 26 26 613 1 5687 1.21
## 5431: NA NA NA NA NA 14 18 18 613 1 17 -1.61
## 5432: NA NA NA NA NA 27 34 34 492 1 322 -5.19
##          Dim2 Dim3 Dim4 Dim5 Dim6 Dim7 n_dia age_min dia_year fit
## 1: 1.70 2.558 -1.86 -0.18 -1.764 -0.616 1 21 2012 0.70
## 2: 1.00 -3.387 1.89 -0.67 0.355 -1.845 4 17 2008 0.20
## 3: 2.86 1.243 -1.62 -2.22 0.956 1.532 2 21 2008 -0.64
## 4: 3.86 2.045 0.73 -0.92 1.240 -1.178 3 16 2003 -1.44
## 5: 3.92 0.121 1.00 -1.39 0.152 -0.140 4 20 2005 -1.83
## ---
## 5428: -0.41 -3.712 -0.61 1.25 -0.391 0.241 7 17 2011 1.67
## 5429: -3.39 -1.517 0.46 3.53 0.487 -0.078 3 14 2005 -1.16
## 5430: 0.77 -2.695 -1.11 1.50 1.454 2.479 2 19 2010 0.77
## 5431: 0.52 -0.012 0.83 -0.37 0.045 0.584 0 14 2012 3.43
## 5432: -0.93 -2.280 -0.28 1.19 -0.237 -0.908 1 24 2009 -2.25
##          maternal_age paternal_age langde apgar5 gest_age bw_score C_RYGER
## 1:           16            23   50    10    38  75.9    NA
## 2:            22            33   54    10    39  98.4    NA
## 3:            32            51   54    10    42  15.0    NA
## 4:            30            29   50    10    38  63.3    NA
## 5:            31            34   49    NA    37  57.3    NA
## ---
## 5428:           32            25   52    10    39  46.4    NA
## 5429:           18            21   54    10    40  64.7    NA
## 5430:           32            24   46    10    35  45.7    NA
## 5431:           28            36   52    10    38  77.3     0
## 5432:           18            NA   45    10    36   8.9    NA
##          B_RYGER ind_mat_preg_Bacterial_infection

```

```

##   1:    NA          0
##   2:    NA          0
##   3:    NA          0
##   4:    NA          0
##   5:    NA          0
##   ---
## 5428:    0          0
## 5429:    NA         0
## 5430:    1          0
## 5431:    NA         0
## 5432:    NA         0
##      ind_mat_preg_Viral_infection AN_2016_SCORES_AT_ALL_THRESHOLDS_05
##   1:          0        -0.059
##   2:          0        -0.059
##   3:          0           NA
##   4:          0           NA
##   5:          0           NA
##   ---
## 5428:          0        -0.059
## 5429:          0           NA
## 5430:          0        -0.059
## 5431:          0        -0.059
## 5432:          0        -0.059
##      Anxiety_SCORES_AT_ALL_THRESHOLDS_05
##   1:      -0.00045
##   2:      -0.00051
##   3:           NA
##   4:           NA
##   5:           NA
##   ---
## 5428:      -0.00058
## 5429:           NA
## 5430:      -0.00050
## 5431:      -0.00054
## 5432:      -0.00046
##      BMI_2015_SCORES_AT_ALL_THRESHOLDS_05
##   1:      -0.00045
##   2:      -0.00042
##   3:           NA
##   4:           NA
##   5:           NA
##   ---
## 5428:      -0.00042
## 5429:           NA
## 5430:      -0.00034
## 5431:      -0.00041
## 5432:      -0.00033
##      BIP_2016_SCORES_AT_ALL_THRESHOLDS_05 BW_SCORES_AT_ALL_THRESHOLDS_05
##   1:          0.423       0.0029
##   2:         -0.779       0.0032
##   3:           NA           NA
##   4:           NA           NA
##   5:           NA           NA
##   ---

```

```

## 5428:          0.031      0.0031
## 5429:          NA        NA
## 5430:          0.532      0.0030
## 5431:         -2.547      0.0031
## 5432:          1.358      0.0031
##   Cannib_SCORES_AT_ALL_THRESHOLDS_05
## 1:          -0.10
## 2:          -2.71
## 3:          NA
## 4:          NA
## 5:          NA
##   ---
## 5428:          -0.81
## 5429:          NA
## 5430:          -0.69
## 5431:          -1.37
## 5432:          0.91
##   DS_2016_SCORES_AT_ALL_THRESHOLDS_05
## 1:          -0.201
## 2:          0.598
## 3:          NA
## 4:          NA
## 5:          NA
##   ---
## 5428:          -0.107
## 5429:          NA
## 5430:          0.534
## 5431:          0.714
## 5432:          -0.081
##   EduYears_2016_SCORES_AT_ALL_THRESHOLDS_05
## 1:          -0.682
## 2:          -0.624
## 3:          NA
## 4:          NA
## 5:          NA
##   ---
## 5428:          -0.417
## 5429:          NA
## 5430:          -0.324
## 5431:          -1.766
## 5432:          0.065
##   GPC_Ext_SCORES_AT_ALL_THRESHOLDS_05
## 1:          -0.000071
## 2:          -0.000149
## 3:          NA
## 4:          NA
## 5:          NA
##   ---
## 5428:          -0.000068
## 5429:          NA
## 5430:          -0.000102
## 5431:          -0.000082
## 5432:          -0.000069
##   NEURO_2016_SCORES_AT_ALL_THRESHOLDS_05

```

```

##    1:                      0.60
##    2:                     -1.48
##    3:                       NA
##    4:                       NA
##    5:                       NA
##    ---
## 5428:                      1.86
## 5429:                         NA
## 5430:                     -0.14
## 5431:                     -1.20
## 5432:                      0.96
##      SCZ_2015_SCORES_AT_ALL_THRESHOLDS_05
##    1:                     -0.80
##    2:                      0.35
##    3:                         NA
##    4:                         NA
##    5:                         NA
##    ---
## 5428:                      1.14
## 5429:                         NA
## 5430:                     -0.37
## 5431:                     -0.19
## 5432:                      1.51
##      SWB_2016_SCORES_AT_ALL_THRESHOLDS_05      C1      C2      C3      C4
##    1:                  -0.40 -0.0027 -0.0009  0.0034  0.0073
##    2:                  0.66 -0.0035  0.0085 -0.0021 -0.0029
##    3:                     NA      NA      NA      NA      NA
##    4:                     NA      NA      NA      NA      NA
##    5:                     NA      NA      NA      NA      NA
##    ---
## 5428:                  -0.13 -0.0051 -0.0007  0.0051 -0.0001
## 5429:                     NA      NA      NA      NA      NA
## 5430:                  -0.66  0.0067 -0.0022 -0.0004 -0.0033
## 5431:                   0.30  0.0075 -0.0035  0.0027  0.0038
## 5432:                   1.66  0.0002 -0.0014  0.0043 -0.0019
##      C5      C6      C7      C8      C9      C10     wave
##    1: -0.0030  0.0000  0.0065 -0.0011 -0.0004  0.0005 Wave13
##    2: -0.0010 -0.0049  0.0020  0.0064  0.0046  0.0027 Wave14
##    3:      NA      NA      NA      NA      NA      NA Other
##    4:      NA      NA      NA      NA      NA      NA Other
##    5:      NA      NA      NA      NA      NA      NA Other
##    ---
## 5428:  0.0031 -0.0008  0.0060  0.0078  0.0010  0.0034 Wave17
## 5429:      NA      NA      NA      NA      NA      NA Other
## 5430:  0.0048 -0.0028 -0.0016  0.0001  0.0043  0.0046 Wave10
## 5431: -0.0105 -0.0024 -0.0025 -0.0050 -0.0038 -0.0037 Wave4_
## 5432:  0.0000  0.0019 -0.0006  0.0054  0.0001 -0.0057 Wave20
##      d2100_ptype0_contacts d2100_ptype0_days d2100_ptype1_contacts
##    1:                      7                  234                  0
##    2:                      0                  0                  0
##    3:                      0                  0                  0
##    4:                     32                 645                  0
##    5:                      2                  65                  0
##    ---

```

```

## 5428:         9        433        0
## 5429:         0         0        0
## 5430:         8       120        0
## 5431:         0         0        0
## 5432:         4        65        0
##      d2100_ptype1_visits d2100_ptype2_visits d2100_ptype2_novisits
## 1:             0           69        0
## 2:             0           0        0
## 3:             0          114        0
## 4:             0           69        0
## 5:             0           0        0
## ---
## 5428:         0           80        0
## 5429:         0           77        0
## 5430:         0          103        0
## 5431:         0           89        0
## 5432:         0           90        1
##      d2100_ptype3_contacts ind_Otitis_infection ind_Bacterial_infection
## 1:             1             0        0
## 2:             0             0        0
## 3:             0             0        0
## 4:             12            0        0
## 5:             4             0        1
## ---
## 5428:         4             0        0
## 5429:         0             0        0
## 5430:         0             1        0
## 5431:         0             1        0
## 5432:         5             1        1
##      ind_CNS_infection ind_Viral_infection disruptpive damaging
## 1:             0             0        0        0
## 2:             0             0        0        1
## 3:             0             0        0        2
## 4:             0             0        NA        NA
## 5:             0             0        NA        NA
## ---
## 5428:         0             0        0        0
## 5429:         0             0        NA        NA
## 5430:         0             0        0        0
## 5431:         0             1        0        0
## 5432:         0             0        0        0
##      disruptive_and_damaging synonymous age_Otitis_infection
## 1:                 0        0        22.1
## 2:                 1        2        23.5
## 3:                 2        2        25.8
## 4:                NA        NA        28.1
## 5:                NA        NA        29.2
## ---
## 5428:         0        0        19.3
## 5429:        NA        NA        27.3
## 5430:         0        2        8.9
## 5431:         0        3        4.4
## 5432:         0        0        1.3
##      age_Bacterial_infection age_CNS_infection age_Viral_infection

```

```

##      1:          22.1          22          22.1
##      2:          23.5          23          23.5
##      3:          25.8          26          25.8
##      4:          28.1          28          28.1
##      5:          24.5          29          29.2
##      ---
## 5428:          19.3          19          19.3
## 5429:          27.3          27          27.3
## 5430:          21.8          22          21.8
## 5431:          14.3          14          4.2
## 5432:          1.3          30          30.4
##      B_SECTIOU B_I11 ind_Fxxxx_m ind_F2100_m ind_Fxxxx_f ind_F2100_f
##      1:       NA   0       1   0       0   0
##      2:       NA   0       1   0       0   0
##      3:       NA   0       0   0       1   0
##      4:       NA   0       0   0       0   0
##      5:       NA   0       1   0       0   0
##      ---
## 5428:       NA  NA   0   0       0   0
## 5429:       NA   0   0   0       0   0
## 5430:       1  NA   0   0       0   0
## 5431:       NA  NA   0   0       1   0
## 5432:       NA   0  NA  NA       NA   NA
##      RYGER contacts visits earl_F10 Bac_after Vir_after
##      1:    NA     8    69     0     0     0
##      2:    NA     0     0     1     0     0
##      3:    NA     0   114     0     0     0
##      4:    NA    44    69     0     0     0
##      5:    NA     6     0     0     1     0
##      ---
## 5428:     0    13    80     1     0     0
## 5429:    NA     0    77     1     0     0
## 5430:     1     8   103     1     0     0
## 5431:     1     0    89     0     0     0
## 5432:    NA     9    90     0     0     0

```

```
data2[ind_Viral_infection==1, Vir_after := as.numeric(age_Viral_infection>= age_min)]
```

```

##      pid year gender birthdate case case2012 random F10 F20 F30 F40
##      1: 1000166 1990      F 1990-12-13    1      1   NA   NA   NA  21   NA
##      2: 1002947 1989      M 1989-07-15    1      1   NA  17   NA   NA   NA
##      3: 1009068 1987      M 1987-03-03    1      1   NA   NA   NA  21   NA
##      4: 1013507 1984      F 1984-11-21    1      1   NA   NA   NA  16   NA
##      5: 1014243 1983      F 1983-10-20    1      1   NA   NA  21  20   NA
##      ---
## 5428: 9315714 1993      M 1993-09-19    1      1     1  17  18   NA  23
## 5429: 9316160 1985      M 1985-09-08    1      1   NA  18  18   NA   NA
## 5430: 9319643 1991      F 1991-03-02    1      1   NA  20   NA   NA  19
## 5431: 9320666 1998      F 1998-09-09    1      1   NA   NA   NA   NA   NA
## 5432: 9323646 1982      M 1982-08-08    1      1   NA  24  27   NA   NA
##      F50 F60 F70 F84 F90 SCZ missing censored N pop_weights mcov Dim1
##      1:  NA  NA  NA  NA  NA  21    26    26  634      1     2 -2.45
##      2:  NA  18  24  NA  17  19    27    27  617      1    12  6.02
##      3:  NA  21  NA  NA  NA  21    30    30  585      1    19  0.85

```

```

##    4: 16 16 NA NA NA 18      32      32 525      1 25 6.36
##    5: 20 20 NA NA 32 22      33      33 480      1 26 3.52
##    ---
## 5428: NA NA NA NA 23 18      23      23 607      1 5682 3.16
## 5429: NA NA NA 14 18 20      31      31 592      1 5683 6.96
## 5430: NA NA NA NA NA 19      26      26 613      1 5687 1.21
## 5431: NA NA NA NA NA 14      18      18 613      1 17 -1.61
## 5432: NA NA NA NA NA 27      34      34 492      1 322 -5.19
##          Dim2   Dim3   Dim4   Dim5   Dim6   Dim7 n_dia age_min dia_year fit
##    1: 1.70 2.558 -1.86 -0.18 -1.764 -0.616  1     21  2012  0.70
##    2: 1.00 -3.387  1.89 -0.67  0.355 -1.845  4     17  2008  0.20
##    3: 2.86 1.243 -1.62 -2.22  0.956  1.532  2     21  2008 -0.64
##    4: 3.86 2.045  0.73 -0.92  1.240 -1.178  3     16  2003 -1.44
##    5: 3.92 0.121  1.00 -1.39  0.152 -0.140  4     20  2005 -1.83
##    ---
## 5428: -0.41 -3.712 -0.61  1.25 -0.391  0.241  7     17  2011  1.67
## 5429: -3.39 -1.517  0.46  3.53  0.487 -0.078  3     14  2005 -1.16
## 5430:  0.77 -2.695 -1.11  1.50  1.454  2.479  2     19  2010  0.77
## 5431:  0.52 -0.012  0.83 -0.37  0.045  0.584  0     14  2012  3.43
## 5432: -0.93 -2.280 -0.28  1.19 -0.237 -0.908  1     24  2009 -2.25
##          maternal_age paternal_age langde apgar5 gest_age bw_score C_RYGER
##    1:           16            23     50     10     38    75.9    NA
##    2:           22            33     54     10     39    98.4    NA
##    3:           32            51     54     10     42    15.0    NA
##    4:           30            29     50     10     38    63.3    NA
##    5:           31            34     49     NA     37    57.3    NA
##    ---
## 5428:           32            25     52     10     39    46.4    NA
## 5429:           18            21     54     10     40    64.7    NA
## 5430:           32            24     46     10     35    45.7    NA
## 5431:           28            36     52     10     38    77.3     0
## 5432:           18           NA     45     10     36     8.9    NA
##          B_RYGER ind_mat_preg_Bacterial_infection
##    1:       NA                      0
##    2:       NA                      0
##    3:       NA                      0
##    4:       NA                      0
##    5:       NA                      0
##    ---
## 5428:       0                      0
## 5429:     NA                      0
## 5430:       1                      0
## 5431:     NA                      0
## 5432:     NA                      0
##          ind_mat_preg_Viral_infection AN_2016_SCORES_AT_ALL_THRESHOLDS_05
##    1:                     0                  -0.059
##    2:                     0                  -0.059
##    3:                     0                      NA
##    4:                     0                      NA
##    5:                     0                      NA
##    ---
## 5428:                     0                  -0.059
## 5429:                     0                      NA
## 5430:                     0                  -0.059

```

```

## 5431:          0          -0.059
## 5432:          0         -0.059
##      Anxiety_SCORES_AT_ALL_THRESHOLDS_05
## 1:          -0.00045
## 2:          -0.00051
## 3:             NA
## 4:             NA
## 5:             NA
##    ---
## 5428:          -0.00058
## 5429:             NA
## 5430:          -0.00050
## 5431:          -0.00054
## 5432:          -0.00046
##      BMI_2015_SCORES_AT_ALL_THRESHOLDS_05
## 1:          -0.00045
## 2:          -0.00042
## 3:             NA
## 4:             NA
## 5:             NA
##    ---
## 5428:          -0.00042
## 5429:             NA
## 5430:          -0.00034
## 5431:          -0.00041
## 5432:          -0.00033
##      BIP_2016_SCORES_AT_ALL_THRESHOLDS_05 BW_SCORES_AT_ALL_THRESHOLDS_05
## 1:            0.423          0.0029
## 2:           -0.779          0.0032
## 3:             NA             NA
## 4:             NA             NA
## 5:             NA             NA
##    ---
## 5428:            0.031          0.0031
## 5429:             NA             NA
## 5430:            0.532          0.0030
## 5431:           -2.547          0.0031
## 5432:            1.358          0.0031
##      Cannib SCORES_AT_ALL_THRESHOLDS_05
## 1:           -0.10
## 2:           -2.71
## 3:             NA
## 4:             NA
## 5:             NA
##    ---
## 5428:           -0.81
## 5429:             NA
## 5430:           -0.69
## 5431:           -1.37
## 5432:            0.91
##      DS_2016_SCORES_AT_ALL_THRESHOLDS_05
## 1:           -0.201
## 2:            0.598
## 3:             NA

```

```

##      4:                      NA
##      5:                      NA
##      ---
## 5428:                  -0.107
## 5429:                      NA
## 5430:                  0.534
## 5431:                  0.714
## 5432:                 -0.081
##      EduYears_2016_SCORES_AT_ALL_THRESHOLDS_05
##      1:                  -0.682
##      2:                  -0.624
##      3:                      NA
##      4:                      NA
##      5:                      NA
##      ---
## 5428:                  -0.417
## 5429:                      NA
## 5430:                  -0.324
## 5431:                 -1.766
## 5432:                  0.065
##      GPC_Ext_SCORES_AT_ALL_THRESHOLDS_05
##      1:                 -0.000071
##      2:                 -0.000149
##      3:                      NA
##      4:                      NA
##      5:                      NA
##      ---
## 5428:                 -0.000068
## 5429:                      NA
## 5430:                 -0.000102
## 5431:                 -0.000082
## 5432:                 -0.000069
##      NEURO_2016_SCORES_AT_ALL_THRESHOLDS_05
##      1:                  0.60
##      2:                 -1.48
##      3:                      NA
##      4:                      NA
##      5:                      NA
##      ---
## 5428:                  1.86
## 5429:                      NA
## 5430:                 -0.14
## 5431:                 -1.20
## 5432:                  0.96
##      SCZ_2015_SCORES_AT_ALL_THRESHOLDS_05
##      1:                 -0.80
##      2:                  0.35
##      3:                      NA
##      4:                      NA
##      5:                      NA
##      ---
## 5428:                  1.14
## 5429:                      NA
## 5430:                 -0.37

```

```

## 5431: -0.19
## 5432: 1.51
## SWB_2016_SCORES_AT_ALL_THRESHOLDS_05 C1 C2 C3 C4
## 1: -0.40 -0.0027 -0.0009 0.0034 0.0073
## 2: 0.66 -0.0035 0.0085 -0.0021 -0.0029
## 3: NA NA NA NA NA
## 4: NA NA NA NA NA
## 5: NA NA NA NA NA
## ---
## 5428: -0.13 -0.0051 -0.0007 0.0051 -0.0001
## 5429: NA NA NA NA NA
## 5430: -0.66 0.0067 -0.0022 -0.0004 -0.0033
## 5431: 0.30 0.0075 -0.0035 0.0027 0.0038
## 5432: 1.66 0.0002 -0.0014 0.0043 -0.0019
## C5 C6 C7 C8 C9 C10 wave
## 1: -0.0030 0.0000 0.0065 -0.0011 -0.0004 0.0005 Wave13
## 2: -0.0010 -0.0049 0.0020 0.0064 0.0046 0.0027 Wave14
## 3: NA NA NA NA NA NA Other
## 4: NA NA NA NA NA NA Other
## 5: NA NA NA NA NA NA Other
## ---
## 5428: 0.0031 -0.0008 0.0060 0.0078 0.0010 0.0034 Wave17
## 5429: NA NA NA NA NA NA Other
## 5430: 0.0048 -0.0028 -0.0016 0.0001 0.0043 0.0046 Wave10
## 5431: -0.0105 -0.0024 -0.0025 -0.0050 -0.0038 -0.0037 Wave4_
## 5432: 0.0000 0.0019 -0.0006 0.0054 0.0001 -0.0057 Wave20
## d2100_ptype0_contacts d2100_ptype0_days d2100_ptype1_contacts
## 1: 7 234 0
## 2: 0 0 0
## 3: 0 0 0
## 4: 32 645 0
## 5: 2 65 0
## ---
## 5428: 9 433 0
## 5429: 0 0 0
## 5430: 8 120 0
## 5431: 0 0 0
## 5432: 4 65 0
## d2100_ptype1_visits d2100_ptype2_visits d2100_ptype2_novisits
## 1: 0 69 0
## 2: 0 0 0
## 3: 0 114 0
## 4: 0 69 0
## 5: 0 0 0
## ---
## 5428: 0 80 0
## 5429: 0 77 0
## 5430: 0 103 0
## 5431: 0 89 0
## 5432: 0 90 1
## d2100_ptype3_contacts ind_Otitis_infection ind_Bacterial_infection
## 1: 1 0 0
## 2: 0 0 0
## 3: 0 0 0

```

```

##    4:          12          0          0
##    5:          4          0          1
##    ---
## 5428:          4          0          0
## 5429:          0          0          0
## 5430:          0          1          0
## 5431:          0          1          0
## 5432:          5          1          1
##      ind_CNS_infection ind_Viral_infection disruptpive damaging
##    1:          0          0          0          0
##    2:          0          0          0          1
##    3:          0          0          0          2
##    4:          0          0          NA         NA
##    5:          0          0          NA         NA
##    ---
## 5428:          0          0          0          0
## 5429:          0          0          NA         NA
## 5430:          0          0          0          0
## 5431:          0          1          0          0
## 5432:          0          0          0          0
##      disruptive_and_damaging synonymous age_Otitis_infection
##    1:          0          0        22.1
##    2:          1          2        23.5
##    3:          2          2        25.8
##    4:          NA         NA        28.1
##    5:          NA         NA        29.2
##    ---
## 5428:          0          0        19.3
## 5429:          NA        NA        27.3
## 5430:          0          2        8.9
## 5431:          0          3        4.4
## 5432:          0          0        1.3
##      age_Bacterial_infection age_CNS_infection age_Viral_infection
##    1:        22.1        22        22.1
##    2:        23.5        23        23.5
##    3:        25.8        26        25.8
##    4:        28.1        28        28.1
##    5:        24.5        29        29.2
##    ---
## 5428:        19.3        19        19.3
## 5429:        27.3        27        27.3
## 5430:        21.8        22        21.8
## 5431:        14.3        14        4.2
## 5432:        1.3         30        30.4
##      B_SECTIOU B_I11 ind_Fxxxx_m ind_F2100_m ind_Fxxxx_f ind_F2100_f
##    1:     NA     0      1      0      0      0
##    2:     NA     0      1      0      0      0
##    3:     NA     0      0      0      1      0
##    4:     NA     0      0      0      0      0
##    5:     NA     0      1      0      0      0
##    ---
## 5428:     NA    NA     0      0      0      0
## 5429:     NA    0      0      0      0      0
## 5430:     1    NA     0      0      0      0

```

```

## 5431:      NA    NA      0      0      1      0
## 5432:      NA    0    NA      NA      NA      NA
##          RYGER contacts visits earl_F10 Bac_after Vir_after
## 1:      NA        8     69      0      0      0
## 2:      NA        0     01      1      0      0
## 3:      NA        0    114      0      0      0
## 4:      NA       44     69      0      0      0
## 5:      NA        6     00      0      1      0
## ---
## 5428:      0       13     80      1      0      0
## 5429:     NA       0     77      1      0      0
## 5430:      1       8    103      1      0      0
## 5431:      1       0     89      0      0      0
## 5432:     NA       9     90      0      0      0

```

```
anova(lm(Dim1~ind_Viral_infection+gender+birthdate,data=data2[Vir_after!=1]))
```

```

## Analysis of Variance Table
##
## Response: Dim1
##                         Df Sum Sq Mean Sq F value Pr(>F)
## ind_Viral_infection   1    90     90   6.03  0.014 *
## gender                  1  1696   1696 113.42 <2e-16 ***
## birthdate                1 10030   10030 670.75 <2e-16 ***
## Residuals            5221 78071      15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```
coef(lm(Dim1~ind_Viral_infection+gender+birthdate,data=data2[Vir_after!=1]))
```

```

##              (Intercept) ind_Viral_infection           genderM
##                 -6.28210             0.19644             -0.92048
##               birthdate
##                 0.00097

```

Family history:

## 1.12 Browser

Summarizing data

```

library(shiny)
library(ggplot2)
library(TraMineR)

mds<- read.csv("/data/projects/IBP/gonthe/sequence_analysis/all/output/mds_trajectories_k7_n31127_18102"
mds_scores <- mds[which(mds$case2012==1),]
mds_norm = as.data.frame(scale(mds_scores[,4:10]))
mds <- cbind(pid=mds_scores$pid, case=mds_scores$case, case2012=mds_scores$case2012, mds_norm)

load("/data/projects/IBP/gonthe/sequence_analysis/all/dist/seq_case-cohort_180314_thin.Rda")

```

```

seq <- seq[rownames(seq) %in% mds$pid,]

mds[, "pid"] <- rownames(mds) <- rownames(seq) <- sample(1:nrow(mds))

hclustlist <- lapply(1:7, function(x) lapply(x:7, function(y) hclust(dist(mds[, c(x+3, y+3)]), "ward.D2")))
tree <- lapply(hclustlist, function(x) lapply(x, cutree, k=1:100))
min <- lapply(tree, function(x) lapply(x, function(y) min(which(sapply(apply(y, 2, table), min)<5))-1))
tree2 <- lapply(1:length(tree), function(x) lapply(1:length(tree[[x]]), function(y) tree[[x]][[y]][, 1]))
tree3 <- lapply(1:length(tree2), function(x) lapply(1:length(tree2[[x]]), function(y) tree2[[x]][[y]][![

rare <- lapply(alphabet(seq), function(x) apply(seq, 1, function(y) sum(y %in% x)))
nseq <- lapply(rare, function(x) lapply(x, function(y) sum(y>0)))
nseq2 <- lapply(nseq, function(x) sum(unlist(x)))
rare_states <- alphabet(seq)[which(nseq2<5)]
rare_seqs <- rare[which(nseq2<5)]
rare_list <- Reduce('+', rare_seqs)

comb <- levels(seq$a1)
comb[levels(seq$a1) %in% rare_states] <- "Other"
for(i in 1:36) levels(seq[[i]]) <- comb

attributes(seq)$alphabet <- unique(comb)
attributes(seq)$labels <- unique(comb)

seq <- seqdef(seq[, seq(3, 36, 3)])

rare_times <- lapply(seq, function(x) which(table(x)<4 & table(x)>0))

for(i in 1:length(rare_times))
  seq[which(as.numeric(seq[, i]) %in% unlist(rare_times[i])), i] <- "Other"

freq_bl <- lapply(1:length(tree), function(x) lapply(1:length(tree[[x]]), function(y) lapply(1:min[[x]][[y]], 1)))
n_bl <- lapply(1:length(tree), function(x) lapply(1:length(tree[[x]]), function(y) lapply(1:min[[x]][[y]], 1)))

cols <- c("#E41A1C", "#377EB8", "#4DAF4A", "#984EA3",
          "#FF7F00", "#FFFF33", "#A65628", "#F781BF", "#999999")
palette(cols)

alpa <- alphabet(seq)

setwd("shiny")

save(tree3, tree2, freq_bl, n_bl, cols, mds, alpa, file="data.Rda")
rm(list=ls())

```

App Script:

```

library(shiny)
library(ggplot2)
library(TraMineR)

```

```

load("data.Rda")
source("seqmodst2.R")

ui <- fluidPage(
  headerPanel('Schizophrenia Trajectories'),
  sidebarPanel(
    selectInput('xcol', 'X Variable', names(mds[4:10])),
    selectInput('ycol', 'Y Variable', names(mds[4:10])),
    selected = names(mds[4:10])[2]),
    numericInput('clusters', 'Groups', 30,
                min = 1, max = 300, ),
    selectInput('type', 'Plot type', c("Modal state", "Chronogram"))

  ),
  mainPanel(
    p("All states observed in <5 individuals are labeled as", span("other.",style="color:brown"),
      "All states observed in <4 individuals at a given age are also labeled as",
      span("other.",style="color:brown")),
    verbatimTextOutput("info"),
    plotOutput('plot1', click = "plot_click"),
    plotOutput('plot2',height = 200),
    plotOutput('plot3')
  )
)

server <- function(input, output) {

  selectedData <- reactive({
    mds[, c(input$xcol, input$ycol)]
  })

  clusters <- reactive({
    x <- which(names(mds[4:10]) %in% input$xcol)
    y <- which(names(mds[4:10]) %in% input$ycol)
    ind1 <- min(x,y)
    ind2 <- abs(x-y)+1
    if(input$clusters>ncol(tree3[[ind1]][[ind2]])) F else tree2[[ind1]][[ind2]][,input$clusters]
  })

  df <-reactive({
    dft <- cbind(selectedData(),factor(clusters()))
    colnames(dft) <- c("x","y","cl")
    dft})
  }

  output$info <- renderPrint({
    x <- which(names(mds[4:10]) %in% input$xcol)
    y <- which(names(mds[4:10]) %in% input$ycol)
    ind1 <- min(x,y)
    ind2 <- abs(x-y)+1
    if(clusters()==F) cat("maximum number of clusters for these dimensions is: ", paste(ncol(tree3[[ind1]])))
  })
}

```

```

click_saved <- reactiveValues(singleclick =NULL)
observeEvent(eventExpr = input$plot_click, handlerExpr = { click_saved$singleclick <- input$plot_click
np <- reactive({
  # if(!is.null(click_saved$singleclick))
  nearPoints(df(),click_saved$singleclick, threshold = 100,
             maxpoints = 1,
             addDist = TRUE) #else nearPoints(df(),list(x=0,y=0,col=1), threshold = 100,maxpoints = 1
})

output$plot1 <- renderPlot({
  par(mar = c(5.1, 4.1, 0, 1))
  g <- ggplot(df(),aes(x=x,y= y,col = cl))+geom_point()+
  theme(legend.position="none")
  #if(nrow(np())<1) np <- matrix(rep(0,4),nrow = 1) else
  np <- np()
  print(np)

  #if(nrow(np)>0){
    cla <-factor(clusters())
    np1 <- cla[which(rownames(mds) %in% rownames(np))]
    print(np1)
    if(nrow(np())<1) cl <- cla %in% 1 else cl <- cla %in% np1
    df <- df()
    g <- g+stat_ellipse(data=df[,cl])
  #}
  g
})

cl <- reactive({
  if(nrow(np())<1) df()$cl[which(df()$cl %in% 1)][1] else df()$cl[which(rownames(mds) %in% rownames(np))]
})

seq1 <- reactive({
  x <- which(names(mds[4:10]) %in% input$xcol)
  y <- which(names(mds[4:10]) %in% input$ycol)
  ind1 <- min(x,y)
  ind2 <- abs(x-y)+1
  c <-which(tree3[[ind1]][[ind2]][,input$clusters]==cl())
  #freq_x <- Reduce('+',freq_bl[[ind1]][[ind2]][c])/sum(c)
  freq_x <- Reduce('+',lapply(c, function(x) freq_bl[[ind1]][[ind2]][[x]]* n_bl[[ind1]][[ind2]][[x]]))
  n_x <- Reduce('+',n_bl[[ind1]][[ind2]][c])
  print("cp")
  cpal <- rep("grey", length(alpa))
  stat<-seqstat1(seqmodst2(freq_x,n=n_x,weighted=F))
  cpal[alpa %in% stat] <- c(cols[1:length(stat)])
  cpal[alpa=="None"]<-"lightgray"
  cpal[alpa=="Other"]<-"brown"
  cpal[alpa=="censored"]<-"white"
  list(freq=freq_x,n=n_x,stat=stat,cpal=cpal)
})

```

```

})

output$plot2 <- renderPlot({
  par(mar = c(5.1, 4.1, 0, 1))
  if(input$type!="Chronogram") plot(seqmodst2(seq1()$freq,n=seq1()$n,weighted=F,freq.mean=T),cpal=)
  res<- list(seq1()$freq,n=seq1()$n)
  names(res) <- c("Frequencies", "ValidStates")
  class(res) <- c("stslist.statd", "list")
  attr(res, "nbseq") <- seq1()$n
  attr(res, "cpal") <- seq1()$cpal
  attr(res, "xtlab") <- colnames(seq1()$freq)
  attr(res, "weighted") <- F
  plot(res,weighted=F)}

})

output$plot3 <- renderPlot({
  par(mar = c(5.1, 4.1, 0, 1))
  if(input$type!="Chronogram"){ alp <- seq1()$stat
  col1 <- seq1()$cpal[alpa %in% seq1()$stat] }else{
    alp <- alpa[apply(seq1()$freq,1,max)>.1]
    col1 <- seq1()$cpal[apply(seq1()$freq,1,max)>.1]}
  plot(0,type='n',axes=FALSE,ann=FALSE)
  legend( 1,1, alp ,fill = col1,cex=2)

})

}

shinyApp(ui = ui, server = server)

```