Data export and wrangling

Andreas Gammelgaard Damsbo

Knitted: 06 January, 2022

Contents


```
library(haven)
library(plyr)
library(dplyr)
library(reshape2)
```

```
dta<-read.csv("/Volumes/Data/exercise/source/background.csv",
              na.strings = c("NA",""),colClasses = "character")
# dta_b<-dta
```

Variables

List of variables included in dataset

```
dput(names(dta))
```

```
## c("cpr", "rnumb", "rdate", "rtreat", "compliant", "wants_out",
## "side_effect", "open_treat", "side_effect2", "eos_done", "protocol",
## "enddate", "eos_early", "mors_d", "mors_p", "inc_time", "completed",
## "mors_180", "intention_t", "who5_score_0", "who5_score_1", "who5_score_6",
## "who5_cut_0", "who5_cut_1", "who5_cut_6", "sdmt_1_c", "sdmt_6_c",
## "mmse_6", "mmse_single", "mmse_range", "mmse_severity", "mmse_clin",
## "cprdash", "debut_time", "deb_adm", "time_diff", "nihss_0", "mrs_0",
## "mrs_1", "mrs_6", "visit_1", "visit_6", "mrs_0_cut2", "mrs_1_cut2",
## "mrs_6_cut2", "mrs_0_cut1", "mrs_1_cut1", "mrs_6_cut1", "mrs_0_cut0",
## "mrs_1_cut0", "mrs_6_cut0", "dob", "age", "sex", "pase_0_date",
## "pase_0", "pase_6_date", "pase_6", "pase_0_q", "pase_0_m", "pase_6_q",
## "pase_06_q", "pase_6_m", "pase_time", "mdi_1", "mdi_6", "mfi_gen_1",
## "mfi_phys_1", "mfi_act_1", "mfi_mot_1", "mfi_men_1", "mfi_gen_bin_1",
## "mfi_phys_bin_1", "mfi_act_bin_1", "mfi_mot_bin_1", "mfi_men_bin_1",
## "mfi_gen_6", "mfi_phys_6", "mfi_act_6", "mfi_mot_6", "mfi_men_6",
## "mfi_gen_bin_6", "mfi_phys_bin_6", "mfi_act_bin_6", "mfi_mot_bin_6",
## "mfi_men_bin_6", "diabetes", "hypertension", "claud", "pad",
## "height", "weight", "weight_est", "afli", "smoker", "alc", "civil",
## "bolig", "ami", "tci", "thrombolysis", "thrombechtomy", "dap_nihss",
## "dap_nihss_24", "smoke_ever", "nihss_c", "nihss_clin5", "rep_any",
## "vasc_dis", "age_std", "nihss_std", "pase_0_std", "est_weight",
## "bmi", "bmi_cut", "pase_change", "pase_qfall", "pase_dcln", "pase_drop",
## "pase_extreme", "pase_0_q_n", "pase_06_q_n", "delta_pase", "pase_change_num"
## )
```

New additions and formatting

```
dta$mors_delay<-difftime(as.Date(dta$mors_d),as.Date(dta$rdate),units = "days")
dta$mors_v1<-factor(ifelse(dta$mors_delay<=38&
                             (dta$mors_delay-as.numeric(dta$inc_time))<=1,
                           "yes","no"))
# Tæller som død hvis død inden 38 dage og dødsdato og EOS ligger indenfor 1 døgn.
dta$mors_v1[is.na(dta$mors_v1)]<-"no"
```

```
dta$mors_v16<-factor(ifelse(dta$mors_v1=="no"&
                              (dta$mors_delay-as.numeric(dta$inc_time))<=1,
                            "yes","no"))
# Tæller som død mellem 1 til 6 mdr, hvis ikke død inden 1 mdr,
# og dødsdato og EOS ligger indenfor 1 døgn.
dta$mors_v16[is.na(dta$mors_v16)]<-"no"
```

PASE score dichotomisation at median score.

```
dta$pase_0<-as.numeric(dta$pase_0)
dta$pase_0_bin<-cut(dta$pase_0,
                    c(min(dta$pase_0,na.rm = T),median(dta$pase_0,na.rm = T),
                      max(dta$pase_0,na.rm = T)),include.lowest = T,
                    labels = c("lower","higher"))
quantile(dta$pase_0,na.rm = T)
```

```
##     0%    25%    50%    75%   100%
##   0.00  76.40 132.50 197.00 574.26
```

Formatting

```
dta$inc_time<-as.numeric(dta$inc_time)
```

Cleaning MDI scores

The following contains a serious bit of data wrangling. Reasons are the occasional recording of visit 1 data at 6 months due to LOCF approach. Additionally some patients have data recorded at 6 months, but later end date has been defined as prior to the visit 6. Additionally the definition of when to define a MDI recording as 1 month or 6 months have added a bit of extra work..

This work should be applied for all endpoint data. If needed, a general script or function should be written.

Steps used for the correction:

1. If the inc_time is 38 days or less MDI 6 scores are moved to MDI 1 and visit 6 is defined as visit 1.
2. If both visit 1 and 6 dates are NA, use enddate as visit 1 date. This is the case if patients were excluded early.
3. If visit 6 is recorded later than enddate, use enddate instead. MDI 6 score is dropped.
4. If visit delay is 7 days or less, and inclusion time is more than 38, MDI 1 is moved to MDI 6 and dropped. If MDI 1 and 6 are different both are kept. Enddate is moved to visit 6 date.
5. Defining the visit 6 date as same as enddate if visit delay is <7.

```
summary(inc196<-dta$inc_time>196)
```

```
##    Mode   FALSE    TRUE
## logical     612      30
```

```
dt1<-dta[inc196,c("rnumb","rdate","visit_1","visit_6","enddate","inc_time","mdi_1","mdi_6","mors_delay"
```

Step 1

```
summary(inc38<-dta$inc_time<=38)
```

```
##    Mode   FALSE    TRUE
## logical     549      93
```

```
dt1<-dta[inc38,c("rnumb","rdate","visit_1","visit_6","inc_time","mdi_1","mdi_6")]
dta$visit_1<-ifelse(inc38&!is.na(dta$visit_6),dta$visit_6,dta$visit_1)
dta$mdi_1<-ifelse(inc38&is.na(dta$mdi_1),dta$mdi_6,dta$mdi_1)
dta$mdi_6[inc38]<-NA
dta$visit_6[inc38]<-NA
# If the inc_time is 38 days or less MDI 6 scores are moved to MDI 1 and visit 6 is defined as visit 1.
# LOCF correction.
```

Step 2

```
summary(na16enddate<-is.na(dta$visit_1)&is.na(dta$visit_6))
```

```
##    Mode   FALSE    TRUE
## logical     570      72
```

```
dt2<-dta[na16enddate,c("rnumb","rdate","visit_1","visit_6","inc_time","mdi_1","mdi_6")]
dta$visit_1<-ifelse(na16enddate,dta$enddate,dta$visit_1)
# If both visit 1 and 6 dates are NA, use enddate as visit 1 date. This is the case if patients were ex
```

Step 3

```
summary(late61<-as.Date(dta$visit_6)>as.Date(dta$enddate)&difftime(as.Date(dta$visit_6),as.Date(dta$end
```

```
##    Mode   FALSE    TRUE    NA's
## logical     537       2     103
```

```
summary(late62<-as.Date(dta$visit_6)>as.Date(dta$enddate)&difftime(as.Date(dta$visit_6),as.Date(dta$end
```

```
##    Mode   FALSE    TRUE    NA's
## logical     530       9     103
```

```
late61[is.na(late61)]<-FALSE
late62[is.na(late62)]<-FALSE

# dt5<-dta[late61,c("rnumb","rdate","visit_1","visit_6","enddate","inc_time","mdi_1","mdi_6")]
# dt6<-dta[late62,c("rnumb","rdate","visit_1","visit_6","enddate","inc_time","mdi_1","mdi_6")]

dta$visit_6<-ifelse(late61,dta$enddate,dta$visit_6)
dta$visit_6<-ifelse(late62,dta$enddate,dta$visit_6)
dta$mdi_6[late62]<-NA
# If visit 6 is recorded later than enddate, use enddate instead
# A group of patients have visit 6 and MDI 6 recorded, but enddate is before visit 6 data.
# After manual lookups, this is likely due to some patients coming for visit 6, but the
# interviewer later realizing, that the patients should have been excluded earlier on.
# Due to this, patients with enddate more than 1 day (leaving room for simple recording errors) prior t
# Patients with 1 day difference the enddate is moved to visit 6 date.
```

Step 4

```
summary(locflate<-(difftime(as.Date(dta$visit_6),as.Date(dta$visit_1))<=7|is.na(dta$visit_1))&dta$inc_t
```

```
##    Mode   FALSE    TRUE    NA's
## logical     620      12      10
```

```
locflate[is.na(locflate)]<-FALSE

dt2<-dta[locflate,c("rnumb","rdate","visit_1","visit_6","inc_time","mdi_1","mdi_6")]

dta$mdi_6<-ifelse(locflate&is.na(dta$mdi_6),dta$mdi_1,dta$mdi_6)

dta$mdi_1[locflate&dta$mdi_1==dta$mdi_6]<-NA
dta$visit_1[locflate&is.na(dta$mdi_1)]<-NA

dta$visit_6<-ifelse(locflate,dta$enddate,dta$visit_6)

# If visit delay is 7 days or less, and inclusion time is more than 38, MDI 1 is moved to MDI 6 and dro
```

Step 5

```
summary(same1n6date<-difftime(as.Date(dta$visit_6),as.Date(dta$visit_1),units = "days")<7)
```

```
##    Mode   FALSE    TRUE    NA's
## logical     527       1     114
```

```
same1n6date[is.na(same1n6date)]<-FALSE
# dt5<-dta[same1n6date,c("rnumb","rdate","visit_1","visit_6","enddate","inc_time","mdi_1","mdi_6",drops
dta$visit_6<-ifelse(same1n6date,dta$enddate,dta$visit_6)
# Defining the visit 6 date as same as enddate if visit delay is <7.
```

Visit delay

```
dta$visit_delay<-difftime(as.Date(dta$visit_6),as.Date(dta$visit_1),units = "days")
# Final calculation of days between visits
summary(as.numeric(dta$visit_delay))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    13.0   145.0   152.0   143.6   155.0   286.0     114
```

newobs definition - DEPRECATED

The definition of a truly new observation is a recorded score at least 7 days after the first score. This was relevant prior to the work of redefining time points for scoring.

```
dta$mdi_6_newobs<-dta$mdi_6
# The newobs variable is later used, but is obsolete due to the previous change in definitions. The pre
```

```
# dta$mdi_6_166<-ifelse(dta$inc_time>166,NA,dta$mdi_6)
# dta$mdi_6_80<-ifelse(dta$inc_time>80,NA,dta$mdi_6)
# dta$mdi_6_protocol<-ifelse(dta$protocol=="2",NA,ifelse(is.na(dta$mdi_6),dta$mdi_1,dta$mdi_6))
# dta$mdi_6_locf<-ifelse(is.na(dta$mdi_6),dta$mdi_1,dta$mdi_6)
```

Drops

Streamlining drop out data to avoid NA's.

```
drops<-c("side_effect2","side_effect","wants_out","open_treat")
for (i in drops) {
  dta[i]<-ifelse(dta[i]=="1. Ja","yes","no")
  dta[i][is.na(dta[i])]<-"no"
}
```

Defining a common all cause drop variable

```
dta$drop<-ifelse(dta$side_effect2=="yes"|dta$side_effect=="yes"|dta$wants_out=="yes"|dta$open_treat=="y
```

Defining drop before or at day 38 (Following protocol design) as drop before 1 month and drop after day 38 as drop between 1 and 6 months

```
cut_line<-38
dta$inc_time<-as.numeric(dta$inc_time)
dta$drop1<-ifelse(dta$drop=="yes"&dta$inc_time<=cut_line,"yes","no")
summary(factor(dta$drop1))
```

```
##  no yes
## 568  74
```

```
# dt3<-dta[,c("rnumb","rdate","visit_1","visit_6","inc_time","mdi_1","mdi_6","mdi_6_newobs","drop1","dr

dta$drop16<-ifelse(dta$drop=="yes"&dta$inc_time>cut_line,"yes","no")
summary(factor(dta$drop16))
```

```
##  no yes
## 567  75
```

```
summary(factor(dta$drop))
```

```
##  no yes
## 493 149
```

```
# dtf<-dta[dta$drop1=="yes",c("mdi_6_newobs","inc_time")]
# dtf<-dta[,c("mdi_1","mdi_6_newobs","inc_time","drop","drop1","drop16")]
```

Enriching

With patients excluded due to open treatment need and defining populations to include/exclude

```
summary(sel_enr_1<-dta$open_treat=="yes"&is.na(dta$mdi_1)&dta$drop1=="yes")
```

```
##  open_treat
##  Mode :logical
##  FALSE:633
##  TRUE :9
```

```
dta$mdi_1_enr<-ifelse(sel_enr_1,21,dta$mdi_1)  # Per agreement, patients excluded due to open treatment
```

Vectorising ex/inclusions at 1 month, to keep patients with data or with later data.

```
summary(dta$excluded_1<-factor(case_when(dta$mors_v1=="yes"|
                                           is.na(dta$mdi_1_enr)&
                                           dta$drop1=="yes"~"ex_1", # Excluded
                         is.na(dta$mdi_1_enr)&!is.na(dta$mdi_6_newobs)~"ca_1", # Missing, but carried t
                         is.na(dta$mdi_1_enr)~"mi_1", # Missing,
                         is.na(dta$mdi_1)&!is.na(dta$mdi_1_enr)~"en_1",
                         TRUE ~ "dt_1"))) # Data available
```

```
## ca_1 dt_1 en_1 ex_1 mi_1
##   17  550    9   55   11
```

```
summary(sel_enr_6<-dta$open_treat=="yes"&dta$drop16=="yes"&is.na(dta$mdi_6_newobs)&dta$excluded_1%in%c(
```

```
##  open_treat
##  Mode :logical
##  FALSE:633
##  TRUE :9
```

```
# Entries to be enriched are entries with need for open treatment after 1 month, with missing mdi_6_new

dta$mdi_6_newobs_enr<-as.numeric(ifelse(sel_enr_6,21,dta$mdi_6_newobs))  # Per agreement, patients excl
```

```
summary(dta$excluded_6<-factor(case_when(is.na(dta$mdi_6_newobs_enr)&dta$excluded_1%in%c("ca_1","dt_1",
                                         is.na(dta$mdi_6_newobs_enr)~"mi_6", # Missing data due to excl
                                         is.na(dta$mdi_6_newobs)&!is.na(dta$mdi_6_newobs_enr)~"en_6", #
                                         dta$excluded_1%in%c("ca_1","dt_1")~"dt_6" # Organic data avail
                                         ))) # Data available
```

```
## dt_6 en_6 ex_6 mi_6
##  505    9   62   66
```

```
# dtf<-cbind(dta[,c("rnumb","mdi_1","mdi_6_newobs","inc_time","drop","drop1","drop16","mdi_1_enr","mdi_
#
# summary(dtf %>% filter(excluded==TRUE))
```

Main Dataset export

```
variable_namebits<-c("rnumb","rtreat","age","sex",
                "bmi",
                "smoke_ever",
                "civil",
                "diabetes",
                "hypertension",
                "pad",
                "afli",
                "ami",
                "tci",
                "nihss_0",
                "thrombolysis",
                "thrombechtomy",
                "rep_any",
                "pase_0",
                "pase_6",
                "mrs_0","mrs_1","mrs_6",
#                "who5_score",
                "mdi",
#                "ham_score_1","ham_score_6",
                "mors",
                "drop",
                "wants_out",
                "side_effect",
                "open_treat",
                "side_effect2",
                "excluded",
                "protocol","eos_early","inc_time",
                "rdate","visit","enddate"
                )
```

```
export<-dta %>% select(contains(variable_namebits))
```

```
write.csv(export,"/Volumes/Data/depression/dep_dataset.csv",row.names = FALSE)
```