Data preprocessing

Bénédicte Colnet*and Imke Mayer†

November 2020

Abstract

This notebook accompanies the review article *Causal inference methods for combining randomizedtrials and observational studies: a review* (2020) and performs the data preprocessing for the joint analysis of CRASH-3 and the Traumabase. It takes as an entry the raw data from each data sets and bind them with proper covariates. The output is the combined data with the raw Traumabase data (with missing values kept). Another similar data frame but with the imputed Traumabase is also produced.

Contents

Load libraries

```
library(readxl)  # Read xlsx
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
```

*Inria, benedicte.colnet@inria.fr
†EHESS, imke.mayer@ehess.fr

```
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(reshape2)
library(ggplot2)

# Set random generator seed for
# reproducible results
set.seed(123)
```

```
# Define data directory for
# loading raw data
# traumabase_rawdata_dir <-
# '~/Documents/phd/Traumabase/'
# crash3_rawdata_dir <-
# '~/Documents/phd/crash-data/crash3/'
traumabase_rawdata_dir <- "~/Documents/TraumaMatrix/TraumaMatrixPipeline/"
crash3_rawdata_dir <- "~/Documents/TraumaMatrix/CausalInference/SAMSI/CRASH/Data/"

data_dir <- "./Data/"  # where to store output
```

CRASH-3

Data loading

In this part we load the CRASH-3 data and merge it with the code of treated or placebo that comes from a separate file. These two files correspond to what we received from the CRASH-3 principal investigators.

```
# Load CRASH3 data
rawData_CRASH3 <- read_xlsx(paste0(crash3_rawdata_dir,
    "CRASH-3_dataset_anonymised_for_Freebird.xlsx"),
    na = c("", "NR", "NA", "NF",
        "IMP", "ND"), )
```

```
print(paste0("Raw CRASH-3 data contains following number of observations: ",
    nrow(rawData_CRASH3)))
```

```
## [1] "Raw CRASH-3 data contains following number of observations: 12743"
```

```
print(paste0("Raw CRASH-3 data contains following number of eligible observations: ",
    nrow(rawData_CRASH3[rawData_CRASH3$eligible ==
        "Yes", ])))
```

```
## [1] "Raw CRASH-3 data contains following number of eligible observations: 12663"
```

We observe that a column precise which patients are eligible or not.

```
# Load treatment information
placebo_treatment_keys <- read.csv(paste0(crash3_rawdata_dir,
    "CRASH-3_unblinded code list.csv"))
print(paste0("Treatment code for CRASH-3 contains following number of observations: ",
    nrow(placebo_treatment_keys)))

# prepare the common code to
# merge the two tables
```

```
placebo_treatment_keys$code <- paste(placebo_treatment_keys$box,
    placebo_treatment_keys$pack,
    sep = "-")

# Merge tables
CRASH3 <- merge(rawData_CRASH3,
    placebo_treatment_keys, by = "code")
print(paste0("Merged data table number of observations: ",
    nrow(CRASH3)))
```

Outcome and treatment

Note that the outcome is the 28-day death due to brain injury (and not all deaths).

```
# Death to binary (1=death)
CRASH3$Death <- ifelse(is.na(CRASH3$timerandtodeath),
    0, 1)

# Brain injury death related to
# binary (1=tbi-death) ---> The
# outcome of interest
CRASH3$TBI_Death <- ifelse((is.na(CRASH3$causeDeath) |
    CRASH3$causeDeath != "Head injury"),
    0, 1)

# Treatment as a binary
# variable
CRASH3$treatment <- ifelse(CRASH3$treatment ==
    "Placebo", 0, 1)
```

Traumabase

Data loading

```
rawData_Traumabase <- read.csv(paste0(traumabase_rawdata_dir,
    "4_computed_dataset.csv"),
    na.strings = c("", "NR", "NA",
        "NF", "IMP", "ND"), sep = ",")

print(paste0("Raw traumabase data contains following number of observations: ",
    nrow(rawData_Traumabase)))
```

```
## [1] "Raw traumabase data contains following number of observations: 20037"
```

Outcome and treatment

We also define treatment and outcome on the Traumabase.

The treatment is considered given when the column *Acide.tranexamique* is equal to “Oui”, and if “No” or missing value is present it is considered as no-treatment.

The outcome in the Traumabase is brain injury related death, with column *Cause.du.décès* equals to “LATA”, or “Mort encéphalique”, or “Trauma cranien”, or “Défaillance multi-viscérale”. Note that it is not all death, and this outcome matches the definition of the CRASH-3 outcome.

```r
# Traumabase outcome
rawData_Traumabase$Death <- rawData_Traumabase$Décès
rawData_Traumabase$Death <- ifelse(rawData_Traumabase$Death ==
    "Oui", 1, 0)

# TBI-related deth definition
# according to the doctors
rawData_Traumabase$TBI_Death <- ifelse(rawData_Traumabase$Cause.du.décès %in%
    c("LATA", "Mort encéphalique",
        "Trauma cranien", "Défaillance multi-viscérale"),
    1, 0)


# Traumabase treatment
rawData_Traumabase$treatment <- ifelse(is.na(rawData_Traumabase$Acide.tranexamique) |
    rawData_Traumabase$Acide.tranexamique ==
        "Non", 0, 1)
```

Common set of covariates

Covariates accounting for patient inclusion into CRASH-3 trial

Extra-cranial bleeding

In CRASH3, one of the eligibility criteria is no major extra-cranial bleeding. The feature is called "majorExtracranial" in the CRASH3 trial with a Yes/No answer. We binarize this data with Yes corresponding to 1, and No to 0 (this is the standard proceadure we apply all along this part for binary covariate).

The equivalent variable in the Traumabase is chosen based on $CGR.6h > 3$ or if variable *colloides* ou *cristallides* > 0 (corresponding to a major extracranial bleeding). These conditions determining presence or absence of an major extracranial bleeding have been decided with the Traumbase doctors.

Age

Only adults are said to be eligible in CRASH3, but we observe that children are included. We record 58 values with age lower than 18 years. Some of them are eligible, others are not. Note that we also record 12737 observations with missing data in the age column.

TBI

The Traumabase contains this feature, we just rename it and binarize it (1 for TBI, and 0 for no TBI). In the CRASH3 trial we made it correspond with intraCranialbleeding feature which as Yes, No and, No CT scan available. To conclude on an intracranial bleeding with no CT scan, we consider there is a TBI since the patient is said to be eligible in CRASH3.

GSC

The Traumabase contains the *Glasgow.initial* covariate (a discrete, range: [3, 15]), and corresponds to Initial Glasgow Coma Scale (GCS) on arrival on scene of enhanced care team and on arrival at the hospital (GCS = 3: deep coma; GCS = 15: conscious and alert). In CRASH 3 data it corresponds to 3 variables that have to be summed. It is also important to note that some Glasgow score are taken after intubation, so not initially. As only one GSC values is mentioned per observation, we keep all the values and consider it initial value.

Other covariates

In this part we also include other covariates that are in the baseline (so that probably have an impact on the treatment effect and the outcome), and other "easy" covariates to map. We include systolic blood pressure, sex, and also pupils reactivity.

```
# the vector that stores the
# variable name relative to
# trial inclusion
trial_eligibility <- c()

# the vector that stores
# additional common variables
# not said to be relative to
# the trial inclusion criteria,
# but still mentioned in the
# CRASH-3 table 1 results
outcome_impact <- c()

# Extracranial bleeding ->
# majorExtracranial
rawData_Traumabase$majorExtracranial <- ifelse((!is.na(rawData_Traumabase$CGR.6h) &
    rawData_Traumabase$CGR.6h >
        3) | (!is.na(rawData_Traumabase$Cristalloïdes) &
    rawData_Traumabase$Cristalloïdes >
        0) | (!is.na(rawData_Traumabase$Colloïdes) &
    rawData_Traumabase$Colloïdes >
        0), 1, 0)

CRASH3$majorExtracranial <- ifelse(CRASH3$majorExtracranial ==
    "Yes", 1, 0)

# store majorExtracranial
# component
trial_eligibility <- c(trial_eligibility,
    "majorExtracranial")

# Age
rawData_Traumabase$age <- rawData_Traumabase$Age.du.patient..ans

# Note that there are two
# outliers with age>120 years.
# By manual inspection, we can
# correct these observations
rawData_Traumabase$age[which(rawData_Traumabase$age ==
    721)] <- 72
rawData_Traumabase$age[which(rawData_Traumabase$age ==
    184)] <- 18

# store age component
trial_eligibility <- c(trial_eligibility,
    "age")

# TBI (1 for TBI, 0 if not TBI)
CRASH3$TBI <- ifelse(CRASH3$intraCranialBleeding ==
```

```
    "Yes" | (CRASH3$intraCranialBleeding ==
    "No CT scan available" & CRASH3$eligible ==
    "Yes"), 1, 0)
rawData_Traumabase$TBI <- ifelse(rawData_Traumabase$Trauma.crânien..lésion.cérébrale.TDM. ==
    "Oui" | rawData_Traumabase$ISS....Head_neck >=
    2, 1, 0)

# store TBI component
trial_eligibility <- c(trial_eligibility,
    "TBI")

# GSC
CRASH3$Glasgow.initial <- as.numeric(substring(CRASH3$gcsEyeOpening,
    1, 1)) + as.numeric(substring(CRASH3$gcsMotorResponse,
    1, 1)) + as.numeric(substring(CRASH3$gcsVerbalResponse,
    1, 1))
trial_eligibility <- c(trial_eligibility,
    "Glasgow.initial")

# Systolic blood pressure
rawData_Traumabase$systolicBloodPressure <- rawData_Traumabase$Pression.Artérielle.Systolique..PAS..à.l

outcome_impact <- c(outcome_impact,
    "systolicBloodPressure")

# Women (1) and men (0)
CRASH3$sexe <- ifelse(CRASH3$sex ==
    "Female", 1, 0)
rawData_Traumabase$sexe <- ifelse(rawData_Traumabase$Sexe ==
    "Féminin", 1, 0)

outcome_impact <- c(outcome_impact,
    "sexe")

# Pupil reactivity
x <- rawData_Traumabase[, "Anomalie.pupillaire..Pré.hospitalier."]
rawData_Traumabase$pupilReact <- case_when(x ==
    "Non" ~ "Both React", x ==
    "Anisocorie (unilatérale)" ~
    "One Reacts", x == "Mydriase Bilatérale" ~
    "None React", x == "Pas précisé" ~
    "Unable to assess")
rawData_Traumabase$pupilReact_num <- case_when(rawData_Traumabase$pupilReact ==
    "Both React" ~ 2, rawData_Traumabase$pupilReact ==
    "One Reacts" ~ 1, rawData_Traumabase$pupilReact ==
    "None React" ~ 0, rawData_Traumabase$pupilReact ==
    "Unable to assess" ~ -1)

CRASH3$pupilReact_num <- case_when(CRASH3$pupilReact ==
    "Both React" ~ 2, CRASH3$pupilReact ==
    "One Reacts" ~ 1, CRASH3$pupilReact ==
    "None React" ~ 0, CRASH3$pupilReact ==
    "Unable to assess" ~ -1)
```

```
outcome_impact <- c(outcome_impact,
    "pupilReact_num")
```

Merge and store data

Note that in CRASH3, first patient could be treated in a 8h window after injury, and then finally 3h. In the final data frame we only keep these patients.

In CRASH-3 it corresponds to timeSinceInjury in hours. In France recommandations for doctors already state to use the tranexamic acid as soon as possible, and in a 3h window after injury.

Also, we only consider patient in the Traumabase that have TBI, as it is the criteria on which inclusion was done in CRASH3.

```
# Time between injury and
# treatment --> keep only
# patients treated within 3h
# data treatement from string
# to numeric hours and minutes
CRASH3$timeSinceInjury_h = format(as.POSIXct(CRASH3$timeSinceInjury,
    format = "%Y-%m-%d %H:%M"),
    format = "%H")
CRASH3$timeSinceInjury_h <- as.numeric(CRASH3$timeSinceInjury_h)
CRASH3$timeSinceInjury_m = format(as.POSIXct(CRASH3$timeSinceInjury,
    format = "%Y-%m-%d %H:%M"),
    format = "%M")
CRASH3$timeSinceInjury_m <- as.numeric(CRASH3$timeSinceInjury_m)

## selection of the pertinent
## subtable for the rest of the
## analysis as the CRASH3
## investigators change the
## protocol to keep only patient
## treated before 3h after
## injury
CRASH3_3h <- CRASH3[CRASH3$timeSinceInjury_h <
    3 | (CRASH3$timeSinceInjury_h ==
    3 & CRASH3$timeSinceInjury_m ==
    0), ]


# only patients from the
# Traumabase with TBI
rawData_Traumabase_tbionly <- rawData_Traumabase[which(rawData_Traumabase$TBI ==
    1), ]

# a few patients have no TBI in
# CRASH-3, to compare similar
# quantity we exclude them
CRASH3_3h_tbionly <- CRASH3_3h[CRASH3_3h$TBI ==
    1, ]

# drop this variable as it
trial_eligibility <- setdiff(trial_eligibility,
```

```
    "TBI")

# additionally, we only
# consider patients from
# centers with sufficiently
# many trauma patients
df <- rawData_Traumabase_tbionly %>%
    dplyr::select(c("Numéro.de.centre")) %>%
    group_by(Numéro.de.centre) %>%
    summarise(n = n()) %>% mutate(effectifs = paste(n,
    "TBI \n patients"))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
centers.too.small <- df[which(df$n <
    20), "Numéro.de.centre"]
rawData_Traumabase_tbionly_goodcenters <- rawData_Traumabase_tbionly[which(!(rawData_Traumabase_tbionly
    "Numéro.de.centre"] %in% c(centers.too.small$Numéro.de.centre))),
    ]


# indicator for RCT and RWD
rawData_Traumabase_tbionly_goodcenters$V <- rep(0,
    nrow(rawData_Traumabase_tbionly_goodcenters))
CRASH3_3h_tbionly$V <- rep(1, nrow(CRASH3_3h_tbionly))

# total data frame
total <- rbind(CRASH3_3h_tbionly[,
    c(trial_eligibility, outcome_impact,
        "TBI_Death", "treatment",
        "V")], rawData_Traumabase_tbionly_goodcenters[,
    c(trial_eligibility, outcome_impact,
        "TBI_Death", "treatment",
        "V")])
```

```
path_to_output <- paste0(data_dir,
    "output_preprocess_combined_crash3_TB.csv")
write.csv(total, path_to_output)
```

Short analysis of time to treatement

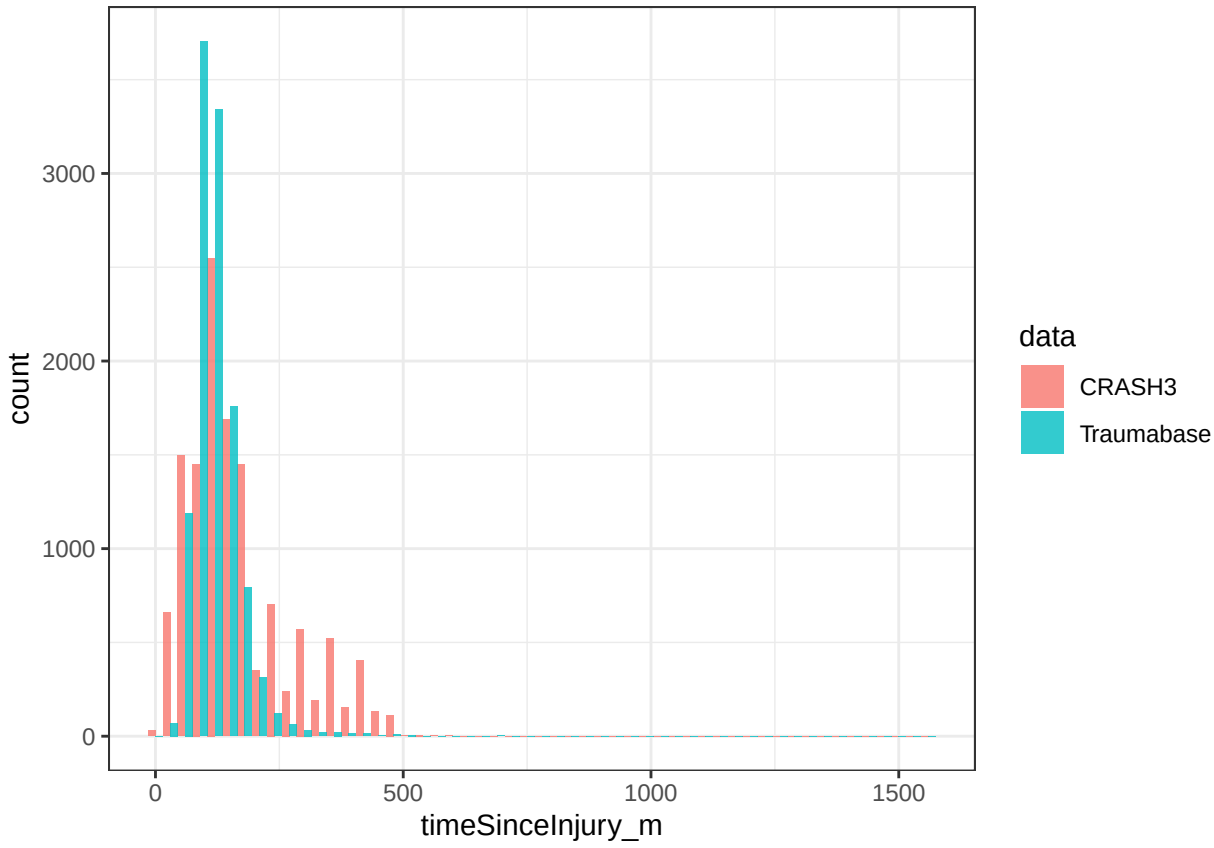
```
# proxy for CRASH-3 data
# treatement from string to
# numeric hours and minutes
rawData_CRASH3$timeSinceInjury_h = format(as.POSIXct(rawData_CRASH3$timeSinceInjury,
    format = "%Y-%m-%d %H:%M"),
    format = "%H")
rawData_CRASH3$timeSinceInjury_h <- as.numeric(rawData_CRASH3$timeSinceInjury_h)
rawData_CRASH3$timeSinceInjury_m = format(as.POSIXct(rawData_CRASH3$timeSinceInjury,
    format = "%Y-%m-%d %H:%M"),
    format = "%M")
rawData_CRASH3$timeSinceInjury_m <- as.numeric(rawData_CRASH3$timeSinceInjury_m)
rawData_CRASH3$timeSinceInjury_m <- rawData_CRASH3$timeSinceInjury_m +
    60 * rawData_CRASH3$timeSinceInjury_h
```

```r
# proxy for Traumabase
rawData_Traumabase$timeSinceInjury_m <- as.numeric(rawData_Traumabase$Délai...arrivée.sur.les.lieux...a
    as.numeric(rawData_Traumabase$Délai...départ.base...arrivée.sur.les.lieux..) +
    30

#+ as.numeric(rawData_Traumabase$Délai.DVE...arrivée.hôpital...pose.de.DVE)
```

```r
rawData_Traumabase$data <- rep("Traumabase",
    nrow(rawData_Traumabase))
rawData_CRASH3$data <- rep("CRASH3",
    nrow(rawData_CRASH3))
comparison_time <- rbind(rawData_Traumabase[,
    c("timeSinceInjury_m", "data")],
    rawData_CRASH3[, c("timeSinceInjury_m",
        "data")])

ggplot(comparison_time, aes(x = timeSinceInjury_m,
    group = data, fill = data)) +
    geom_histogram(binwidth = 30,
        alpha = 0.8, position = "dodge") +
    theme_bw()
```



Imputed data for the Traumbase

The same proceadure is performed with the imputed Traumabase.

Perform imputation

Imputation is performed on the already filtered Traumabase data set.

```
# Recode values of imputed
# categorical variables and
# recast some numericals
# variables into integers

cast_types = function(i, df, data.num) {
    if (is.factor(df[, i])) {
        df[, i] = plyr::mapvalues(df[,
            i], from = levels(df[,
            i]), to = gsub(paste(i,
            "_", sep = ""), "",
            levels(df[, i])))
    } else {
        if (i %in% data.num) {
            df[, i] <- round(df[,
                i], digits = 1)
        } else {
            df[, i] <- as.integer(round(df[,
                i], digits = 0))
        }
    }
    return(df[, i])
}

vars.for.imputation <- c("Numéro.de.centre",
    "Traitement.anticoagulant",
    "Traitement.antiagrégants",
    "Glasgow.initial", "Glasgow.moteur.initial",
    "Mannitol...SSH", "Régression.mydriase.sous.osmothérapie",
    "Arrêt.cardio.respiratoire..massage.",
    "Fréquence.cardiaque..FC..à.l.arrivée.du.SMUR",
    "Cristalloïdes", "Colloïdes",
    "Hémocue.initial", "Delta.Hémocue",
    "Catécholamines", "SpO2.min",
    "Délai...arrivée.sur.les.lieux...arrivée.hôpital..",
    "Score.de.Glasgow.en.phase.hospitalière",
    "Glasgow.moteur", "Anomalie.pupillaire..Phase.hospitalière.",
    "FC.en.phase.hospitalière",
    "Doppler.TransCrânien..DTC...Index.de.Pulsatilité..IP..max",
    "FiO2", "Bloc.dans.les.premières.24h....Neurochirurgie..ex....Craniotomie.ou.DVE.",
    "Total.Score.IGS", "Osmothérapie",
    "HTIC...25.PIC.simple.sédation.",
    "Dérivation.ventriculaire.externe..DVE.",
    "Craniectomie.dé.compressive",
    "ISS....Head_neck", "ISS....Face",
    "ISS....External", "Score.ISS",
    "Activation.procédure.choc.hémorragique",
    "ISS....Selection", "age",
    "TBI", "majorExtracranial",
    "systolicBloodPressure", "pupilReact_num",
    "sexe", "treatment")
```

```
if (file.exists(paste0(data_dir,
    "traumabase_tbideath_tbi_imputed_mice.RData"))) {
    load(file = paste0(data_dir,
        "traumabase_tbideath_tbi_imputed_mice.RData"))
} else {
    m = 5

    DF_tbi <- rawData_Traumabase_tbionly_goodcenters

    df.tmp <- DF_tbi[, vars.for.imputation]
    df.tmp$treatment <- as.factor(df.tmp$treatment)
    imp.mice.mids <- mice::mice(df.tmp,
        m = m, printFlag = F)
    df.imp.mice <- list()
    for (k in 1:m) {
        df.imp.mice[[k]] <- mice::complete(imp.mice.mids,
            k)
        df.imp.mice[[k]]$TBI_Death <- DF_tbi$TBI_Death
        df.imp.mice[[k]]$treatment <- as.numeric(as.character(df.imp.mice[[k]]$treatment))
        df.imp.mice[[k]] <- df.imp.mice[[k]][,
            c(trial_eligibility,
                outcome_impact,
                "TBI_Death", "treatment",
                "Numéro.de.centre",
                "ISS....Head_neck")]
    }
    save(df.imp.mice, imp.mice.mids,
        file = paste0(data_dir,
            "traumabase_tbideath_tbi_imputed_mice.RData"))
}
```

Merge imputed data

```
imputed_traumabase <- df.imp.mice[[1]]
imputed_traumabase$V <- rep(0,
    nrow(imputed_traumabase))
imputed_traumabase <- imputed_traumabase[,
    names(total)]
total_with_imputations <- total[total$V ==
    1, ]
total_with_imputations <- rbind(total_with_imputations,
    imputed_traumabase)

path_to_imputed <- paste0(data_dir,
    "output_preprocess_combined_crash3_TB_imputed.csv")
write.csv(total_with_imputations,
    path_to_imputed)
```