Combined analysis of CRASH-2 and Traumabase for patients with Traumatic Brain Injury, part of 'Treatment effec estimation with missing attributes'

Imke Mayer* Other contributors†

March 2021

Abstract

This notebook performs separate and joint analyses of the CRASH-2 RCT and the observational Traumabase registry. The input data are merged to form a single table ofthe randomized controlled trial and the observational registry (corresponding to the output of `preprocessCrash2Crash3Traumabase.Rmd`). The key functions to perform the analysis below come from the script `estimators.R` (or `estimators_wo_cw.R`).

Contents

*EHESS, imke.mayer@ehess.fr

†Other contributors to this notebook through previous collaborations or active discussions, Bénédicte Colnet, Julie Josse, Françoix-Xavier Ageron, Tobias Gauss, Jean-Denis Moyer.

Preliminaries

Load libraries

```
library(cobalt)  # for balance plots
library(ggplot2)  # for plots
library(forcats)  # for factor handling
library(mice)  # for multiple imputation
library(boot)  # for bootstrap methods
library(naniar)  # for missing values plots
library(FactoMineR)  # for catdes
library(assertthat)  # for assert_that
library(devtools)  # for source_url load
library(nleqslv)  # for searchZeros function required in genRCT package
library(misaem)  # for glm (linear and logistic) with missing data
library(grf)  # for generalized random forests, option for incomplete data
library(pracma)
library(micemd)  # for mice on multilevel data
library(purrr)  # for the map function used in data.frame handling

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

# Set data path Define data
# directory for loading
# pre-processed data
data_dir <- "./data/"
# Define figure directory to
# save figures
fig_dir <- "./figures/"
# Define results directory to
# save computation results
# (bootstrap)
results_dir <- "./results/"

# Load estimators and auxiliary
# functions
source("./catdes_redefined.R")

# If not installed yet, you
# need to un-comment the
# following line once to
# install the genRCT package
# that allows to use the
# calibration weighting (CW)
```

```
# estimator
# install.packages('genRCT_0.1.0.tar.gz',
# repos = NULL)
access_genRCT <- require(genRCT)  # calibration weighting estimator, implementation by Dong et al.

# Load implemented estimation
# functions from GitLab
# repository
if (access_genRCT) {
    source("estimators_and_simulations.R")
} else {
    source("estimators_and_simulations_wo_cw.R")
}

source_url("https://raw.githubusercontent.com/imkemayer/causal-inference-missing/master/Helper/helper_ca
source_url("https://raw.githubusercontent.com/imkemayer/causal-inference-missing/master/Helper/helper_in
```

Choose analysis parameters (outcome, stratum, target population, methods, number of bootstrap samples)

```
outcome_name <- "Death"  # outcome used in all analyses (either 'Death' or 'TBI_Death',
# corresponding to 28day
# all-cause mortality and 28day
# TBI related mortality
# respectively)
outcome_name_string <- "death28d"
stratum_name <- "all"  # stratum to consider (either 'all', 'mild_moderate', 'severe', 'any_non_react',
rct_name <- "CRASH-2"
source_population <- "all"  # either 'all' patients from RCT, or only subset of '3h' patients or 'tbi'
```

Recall observational results for the Traumabase

```
results_rwe
```

```
##     Context         Model Stratum        ATE        STD      CI_inf    CI_sup
## 19      RWD MICE_AIPW_glm     all 0.11097006 0.06187617 -0.01030724 0.2322474
## 20      RWD MICE_AIPW_grf     all 0.03442870 0.03519385 -0.03455125 0.1034087
## 37      RWD  MIA_AIPW_grf     all 0.06153907 0.04415771 -0.02501003 0.1480882
## 191     RWD  MICE_IPW_glm     all 0.28302595 0.12970166  0.02881068 0.5372412
## 201     RWD  MICE_IPW_grf     all 0.16805604 0.04277203  0.08422286 0.2518892
## 33      RWD   MIA_IPW_grf     all 0.17262484 0.04770000  0.07913283 0.2661168
```

Separate analyses

Load the pre-processed CRASH-2 and Traumabase data

To pre-process the CRASH-2 and the Traumabase data, first run the notebook `preprocessCrash2Crash3Traumabase.Rmd`.

```
# Load names of relevant
# variables
load(paste0(data_dir, "crash2_crash3_variables.RData"))

# Load incomplete combined RCT
# data
```

```r
total_allPatients <- read.csv(paste0(data_dir,
    "output_preprocess_combined_allPatients_crash2_crash3_TB.csv"),
    row.names = 1)

# Load incomplete combined data
total_tbi <- read.csv(paste0(data_dir,
    "output_preprocess_combined_crash2_crash3_TB.csv"),
    row.names = 1)

# Load incomplete combined data
total_tbi_3h <- read.csv(paste0(data_dir,
    "output_preprocess_combined_crash2_3h_crash3_3h_TB.csv"),
    row.names = 1)
```

Depending on the value of `source_population`, we keep

- all patients if `source_population=="all"`
- only patients from CRASH-2 who were randomized within 3 hours of the accident if `source_population=="3h"`
- only TBI patients from CRASH-2 if `source_population=="tbi"`
- only TBI patients from CRASH-2 who were randomized within 3 hours of the accident if `source_population=="tbi-3h"`

Separate analyses to reproduce paper results

CRASH-2 analysis (results from CRASH-2 paper)

In this part we load the CRASH-2 data and reproduce the results in the publication with the risk ratio (RR). We also provide the results with the ATE to fit the framework of the review.

The outcome is the 28-day all-cause death.

To recover the exact same results as presented in the CRASH-2 paper, we do not remove patients with time since injury of more than 3 hours.

```
## risk_placebo     risk_TXA           RR     lower_ci     upper_ci
##   15.9877398   14.5546418    0.9103627    0.8453210    0.9754044
```

```
##          ATE     lower_ci     upper_ci
## -0.014330979 -0.024249240 -0.004412719
```

We also recover the results for the head-injury related 28d mortality.

```
## risk_placebo     risk_TXA           RR     lower_ci     upper_ci
##    6.1400040    5.9843456    0.9746485    0.8660853    1.0832117
```

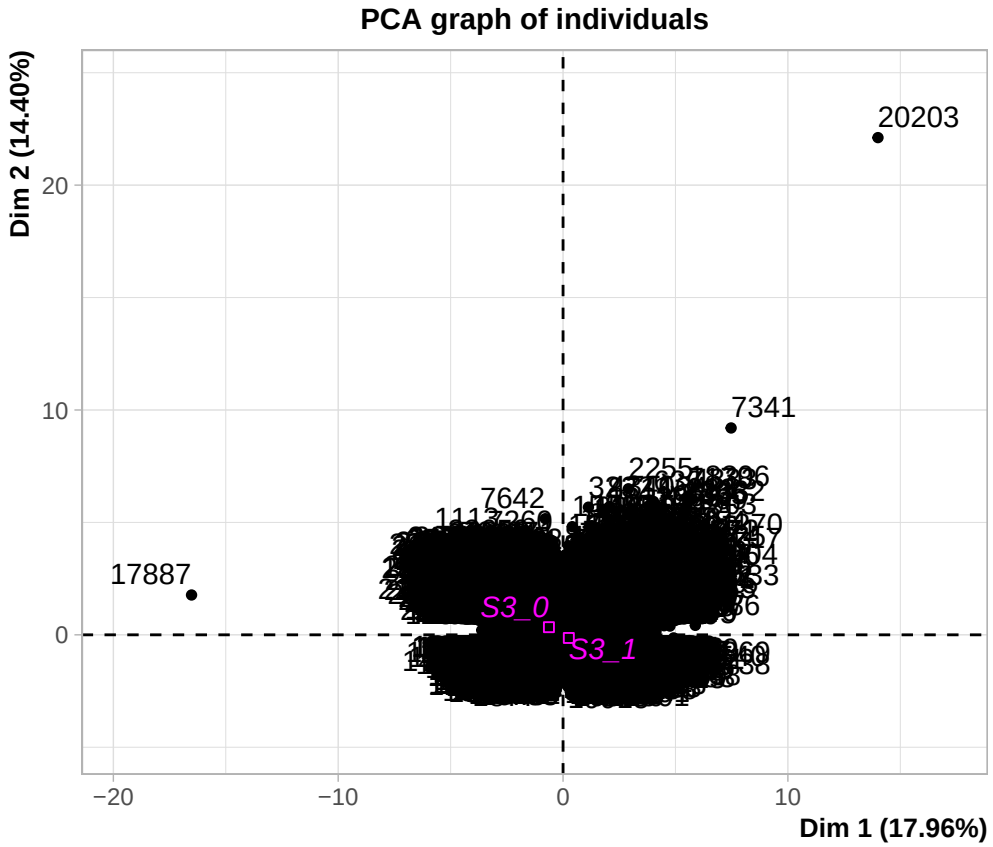
```
##          ATE     lower_ci     upper_ci
## -0.001556584 -0.008137825  0.005024658
```

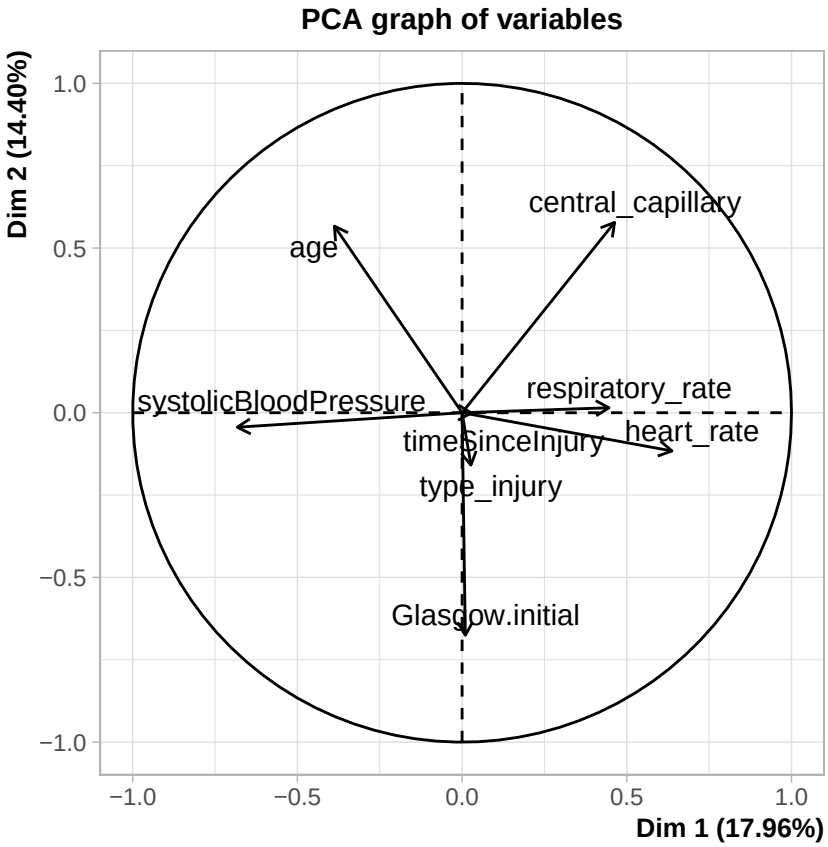
Before computing the estimates on the CRASH-2 data, we first compute the PCA to detect possible outliers.

Analysis on `total`

Before computing the estimates on the RCT data, we first compute the PCA to detect possible outliers.

ACP

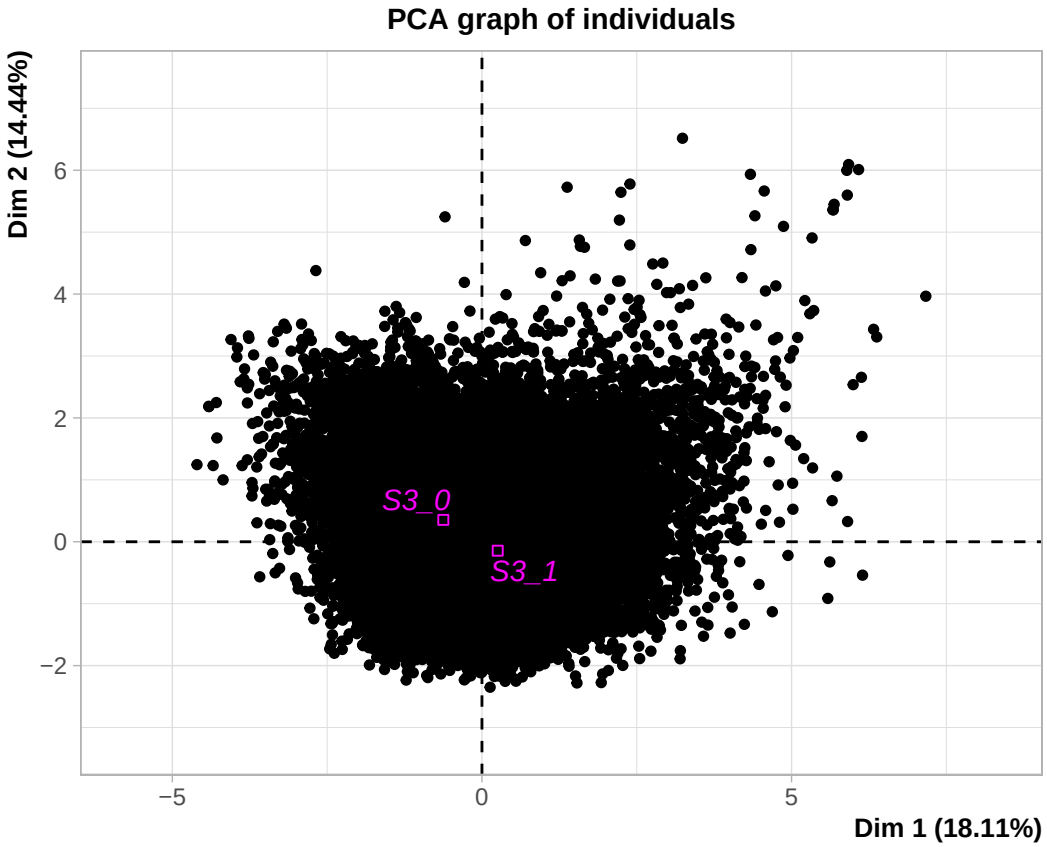

PCA graph of individuals
Dim 2 (14.40%)
Dim 1 (17.96%)
20203
7341
7642
17887
S3_0
S3_1
20
10
0
−20
−10
0
10

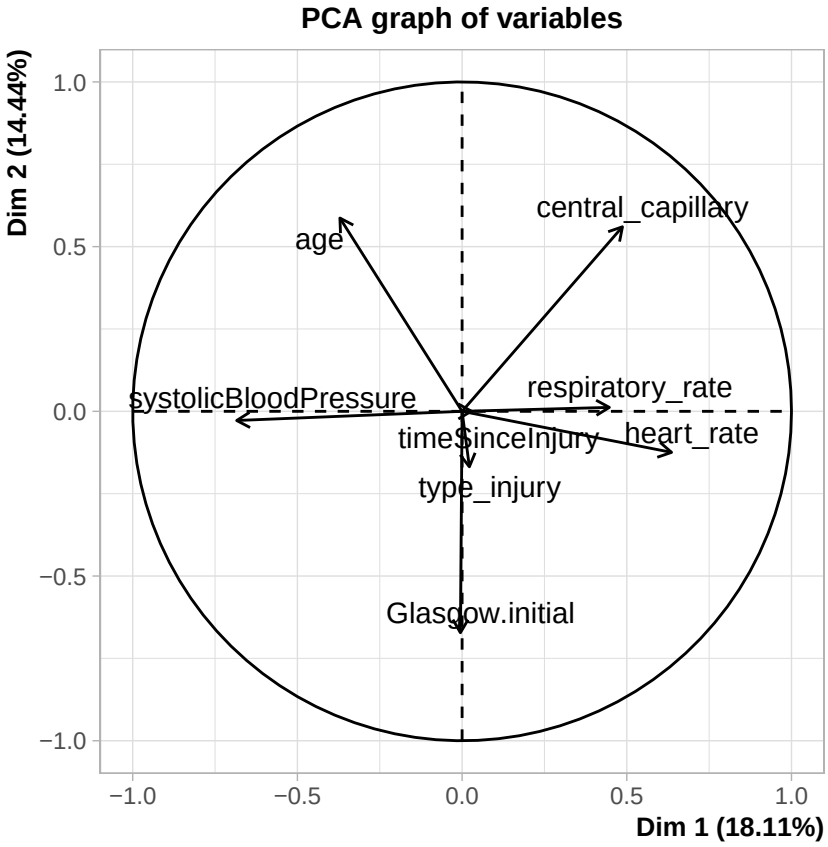


We identify 2 outliers for CRASH-2 and 3-4 outliers for CRASH-3. The corresponding observations are:

```
##       S3 age systolicBloodPressure heart_rate timeSinceInjury Glasgow.initial
## 7341   1  29                  100        110             1.0              15
## 17887  1  70                  999         96             2.0               5
## 20203  1  22                   40          4             0.3               9
##       central_capillary respiratory_rate type_injury
## 7341                 30               18           3
## 17887                 5               25           1
## 20203                60               12           2
```

We will remove these observations


PCA graph of individuals
Dim 2 (14.44%)
6
4
2
0
−2
S3_0
S3_1
−5
0
5
Dim 1 (18.11%)


PCA graph of variables
Dim 2 (14.44%)
1.0
0.5
0.0
−0.5
−1.0
age
central_capillary
systolicBloodPressure
respiratory_rate
timeSinceInjury
heart_rate
type_injury
Glasgow.initial
−1.0
−0.5
0.0
0.5
1.0
Dim 1 (18.11%)

We will keep the results from all RCT patients and from the TBI patients in the RCT

```
## risk_placebo     risk_TXA           RR     lower_ci     upper_ci
##   15.9877398   14.5490585    0.9100135    0.8449590    0.9750680
```

```
##          ATE     lower_ci     upper_ci
## -0.014386813 -0.024305023 -0.004468603
```

```
##          ATE     lower_ci     upper_ci
## -0.011734681 -0.020653496 -0.002815866
```

Final data set overview

Size

The final size of the data.frame is 28452, with

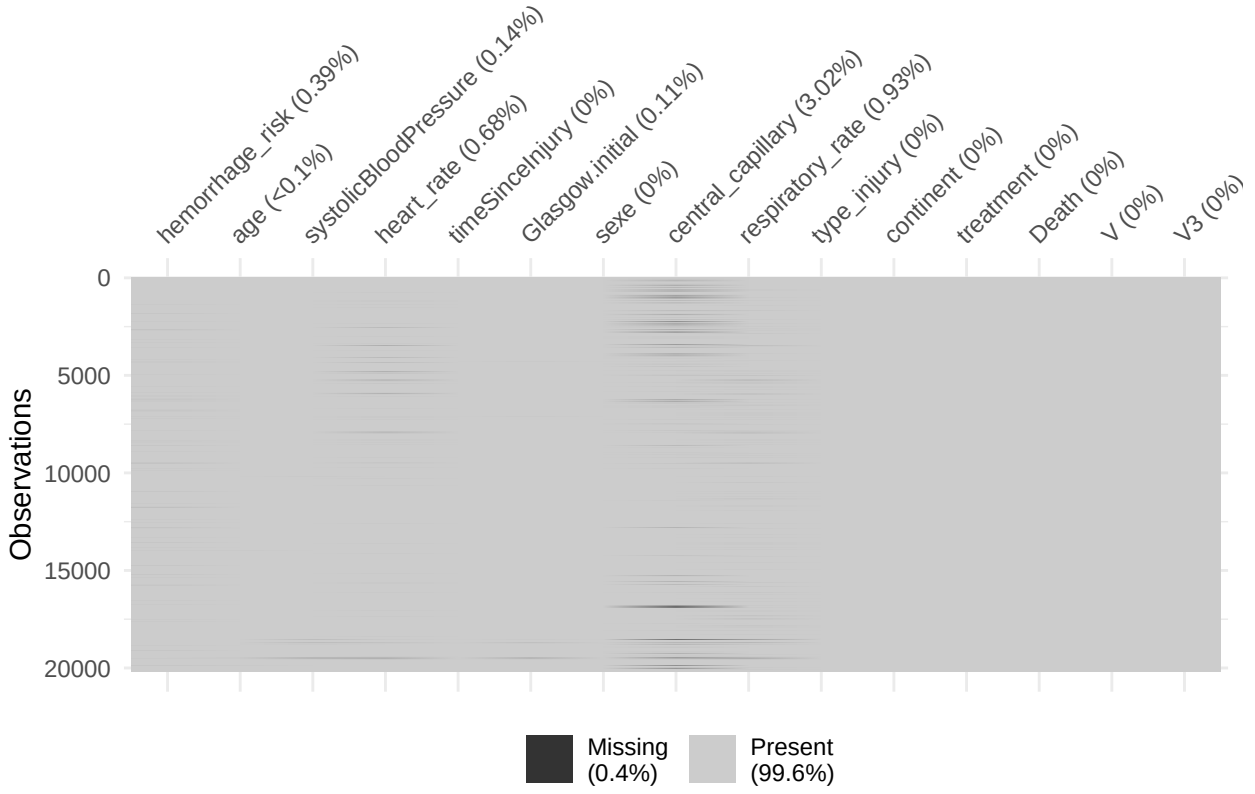
- 20204 observations from the CRASH-2 RCT , and
- 8248 observations from the Traumabase.

```
##             Treated
## Study             0     1
##   CRASH-2     10114 10090
##   Traumabase   7565   683
```

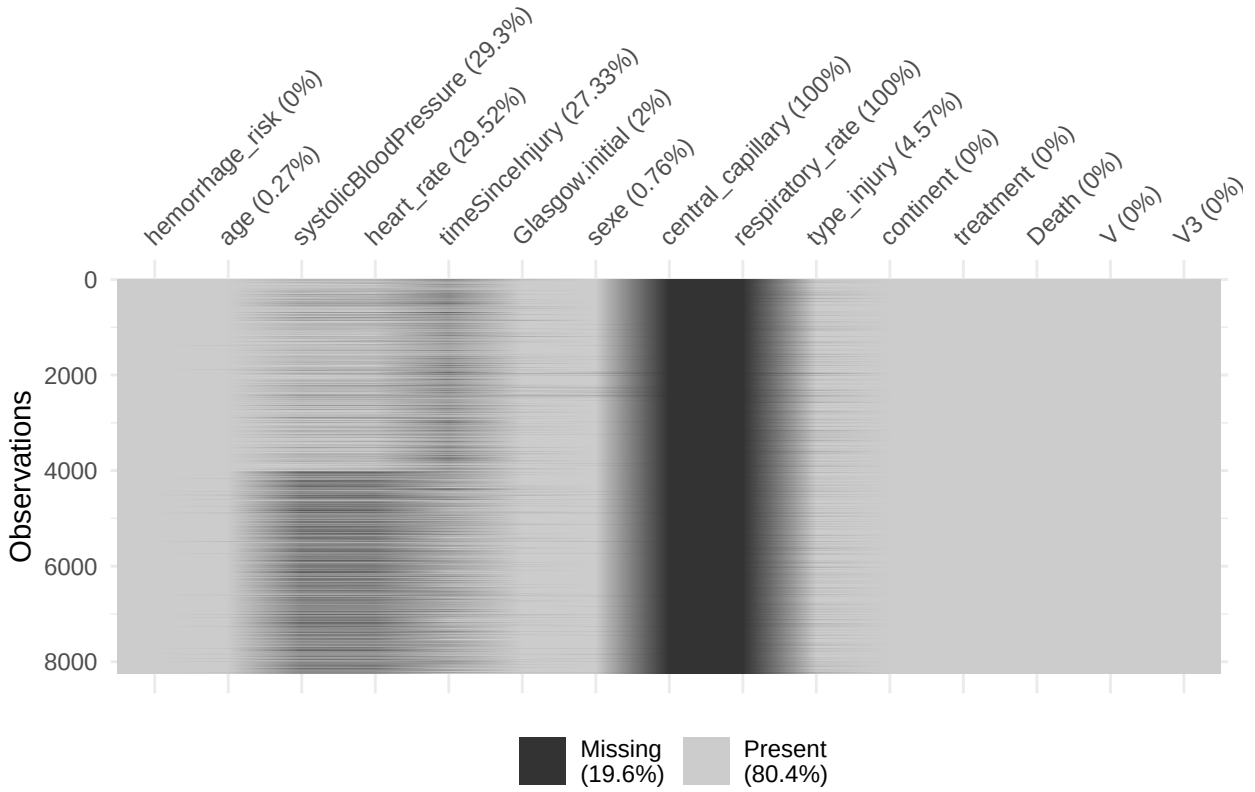
```
##             Died
## Study             0     1
##   CRASH-2     17119  3085
##   Traumabase   6605  1643
```

Missing values

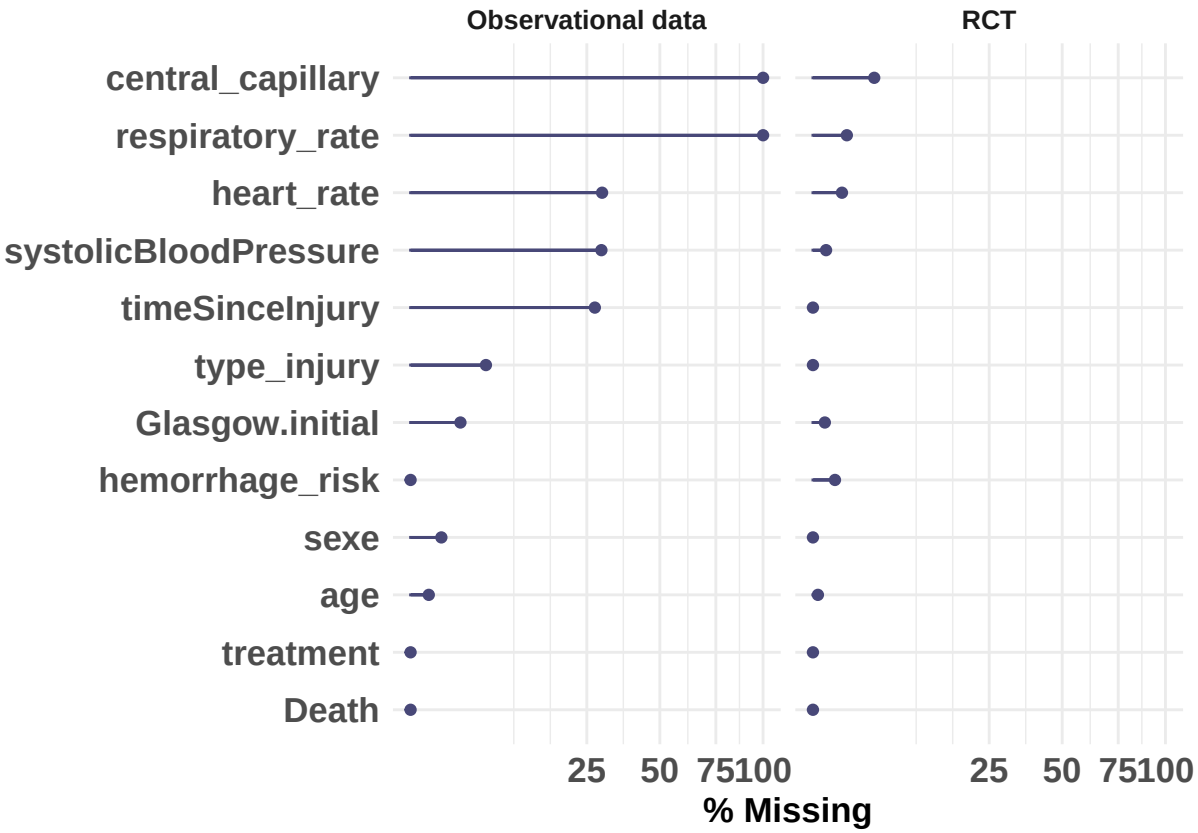
First, note that the RCT contains nearly no missing values.



The Traumabase subset taken contains missing values, it explains why the estimators for transporting the ATE have to be adapted to take into account these missing values.



Alternatively we can plot the barplots of percentage of missing values



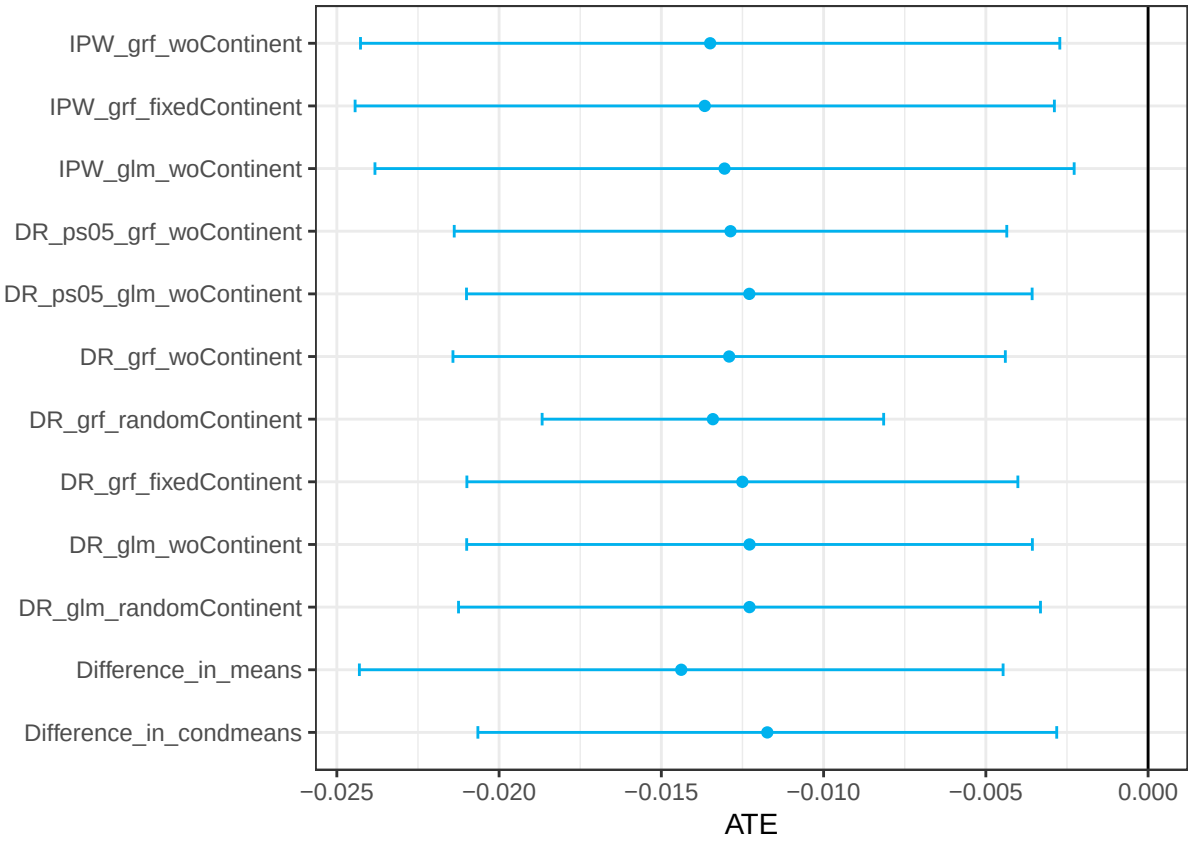
Separate causal analyses using standard causal inference estimators (IPW, etc.)

ATE using only the RCT data

We apply the standard ATE estimators, parametric and non-parametric, on the RCT data, using the baseline variables as regressors.

Estimation

Plot of the final results



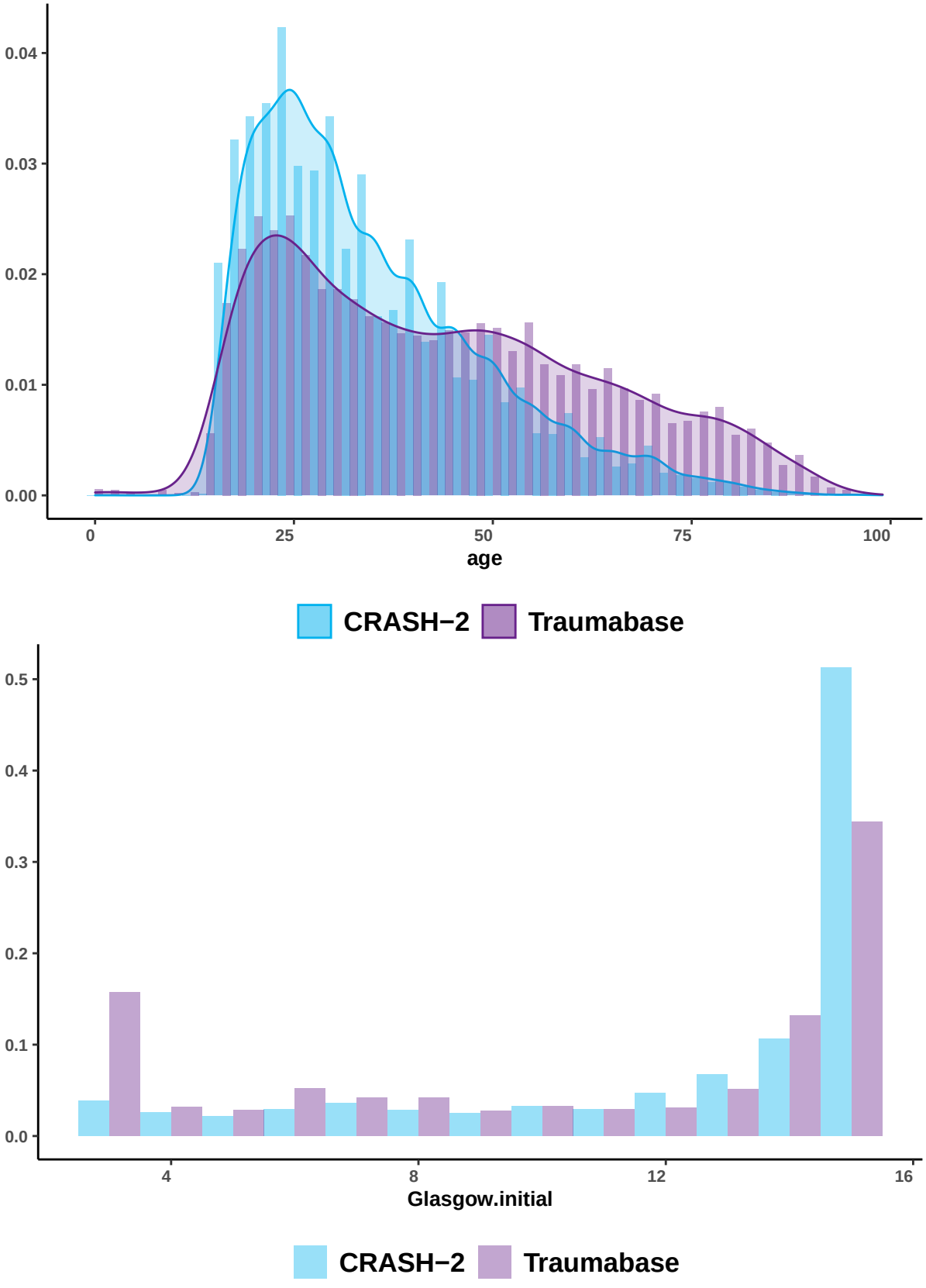
ATE using only the Traumabase data

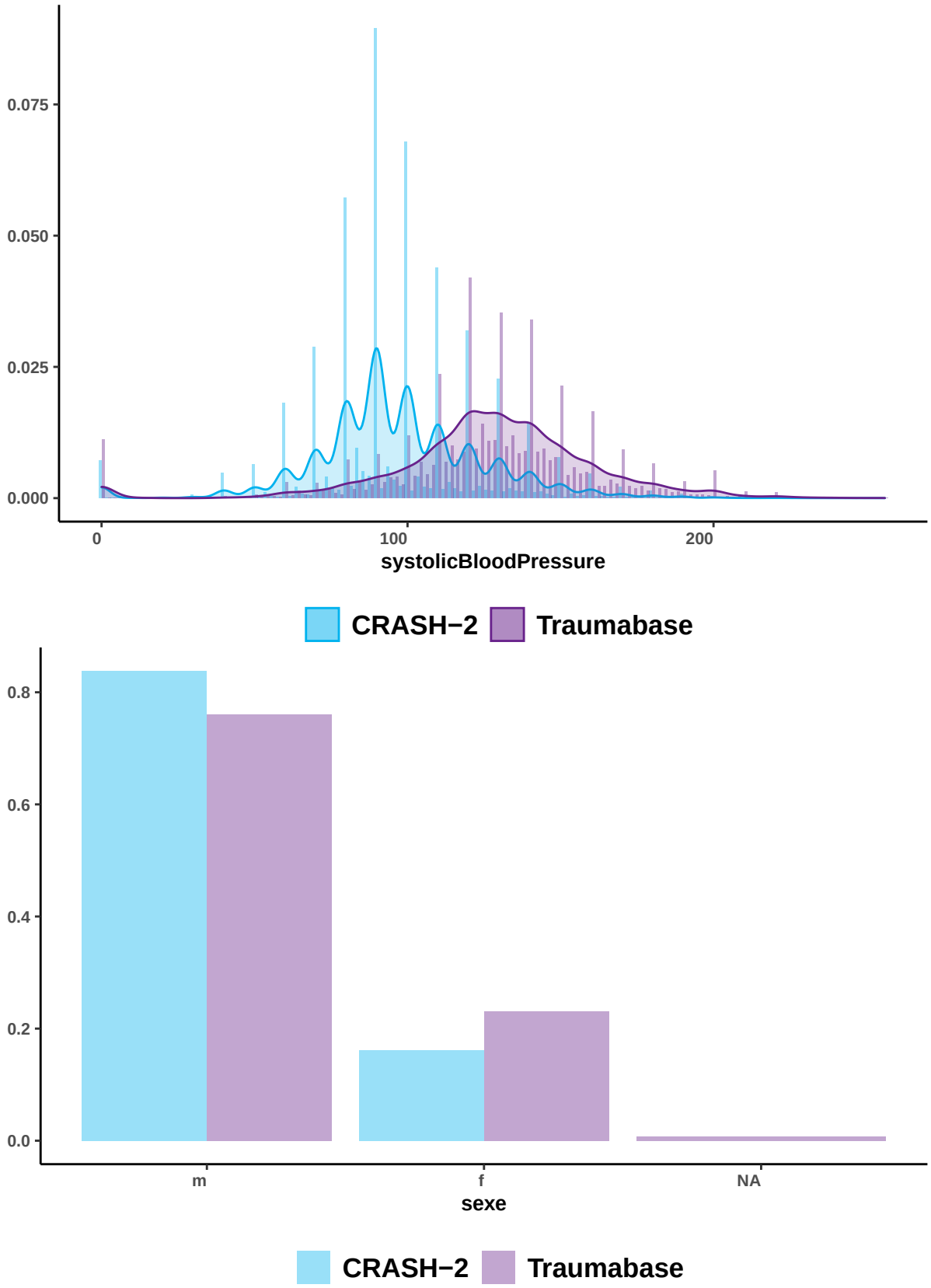
We can first use a naive difference in means, for which we can conclude that TXA increases death (treatment bias, confounding bias, Simpson's paradox).

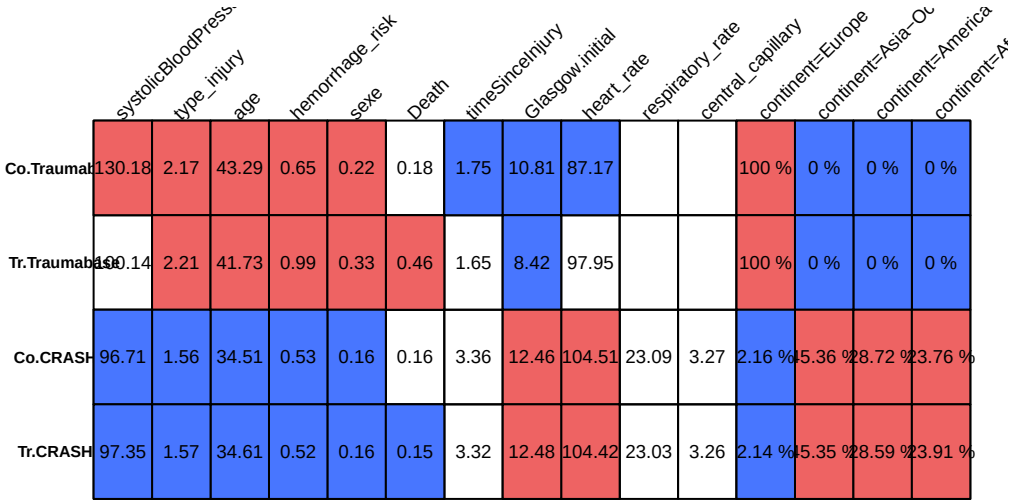
```
## risk_placebo     risk_TXA           RR     lower_ci     upper_ci
##    17.541309    46.266471     2.637572     2.543129     2.732016
```

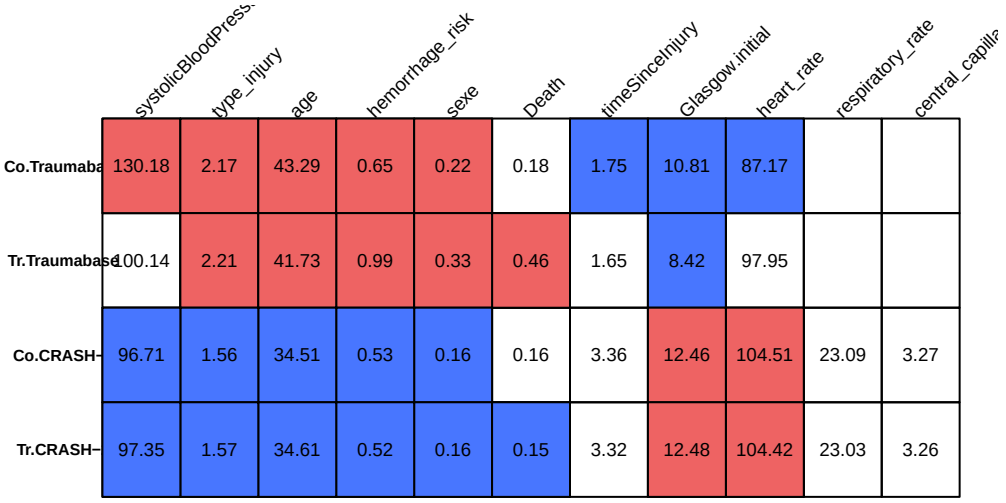
```
##       ATE  lower_ci  upper_ci
## 0.2872516 0.2488344 0.3256689
```

Distributional shift visualization




0.075
0.050
0.025
0.000
0
100
200
systolicBloodPressure
CRASH−2
Traumabase
0.8
0.6
0.4
0.2
0.0
m
f
NA
sexe
CRASH−2
Traumabase





Preliminaries for generalization analysis

We remove `central_capillary` and `respiratory_rate` from the list of outcome regressors since these are not available the Traumabase.

Point estimates

We start by applying all estimators (implemented in the `estimators.R` script) on the `total` data.frame.

```
##      ipsw_hat ipsw.norm_hat  gformula_hat     aipsw_hat     strat_hat
##   -0.05902187   -0.02502415   -0.01860297   -0.05066760   -0.02595748
##        cw_hat
##            NA
```

```
##      ipsw_hat ipsw.norm_hat  gformula_hat     aipsw_hat     strat_hat
##  -0.001203401  -0.007639273  -0.024652020  -0.028795590   0.000855492
##        cw_hat
##            NA
```

Confidence interval estimation (Bootstrap)

The confidence intervals are estimated via non-parametric stratified bootstrap.

```r
stratified_bootstrap <- function(DF,
    nboot = 100, estimator, method,
    outcome_name = "TBI_Death",
    vars_s_model = NULL, vars_y_model = NULL,
    cw_type = "Hajek", do_mi = FALSE,
    micemd_method = NULL, nb_mi = NULL,
    strategy = NULL, complete_cases = FALSE,
    ampute = FALSE, verbose = FALSE) {

    estimands <- c()

    ct_fail <- 0
    if (verbose)
        cat("Iteration ")
    for (i in 1:nboot) {
        if (verbose)
            cat(paste0(i, " "))

        # random resamples from RCT
        n = nrow(DF[DF$V == 1,
            ])
        index_RCT = sample(1:n,
            n, replace = TRUE)

        # random resamples from RWD
        m = nrow(DF[DF$V == 0,
            ])
        index_RWD = sample(1:m,
            m, replace = TRUE)

        # new data set
        RCT_RWD <- rbind(DF[which(DF$V ==
            1), ][index_RCT, ],
            DF[which(DF$V == 0),
                ][index_RWD, ])

        # ampute values to keep similar
        # fraction of NA in RWD part of
        # the data
        if (ampute) {
            prop_miss_RWD <- sapply(DF[DF$V ==
                0, ], function(x) mean(is.na(x)))
            for (j in 1:ncol(DF)) {
                prop_miss_boot <- mean(is.na(RCT_RWD[which(RCT_RWD$V ==
                  0), j]))
                if (prop_miss_RWD[j] >
                  0.1 & prop_miss_RWD[j] >
                  prop_miss_boot) {
                  idx_miss <- which(is.na(RCT_RWD[which(RCT_RWD$V ==
                    0), j]))
                  idx_new_miss <- sample(m -
                    length(idx_miss),
                    floor(m * (prop_miss_RWD[j] -
```

```
                    prop_miss_boot)),
                replace = F)
            }
        }
    }

    # estimation
    estimand <- NULL

    if (do_mi) {
        try(estimand <- unlist(estimator(RCT_RWD,
            outcome_name = outcome_name,
            method = method,
            complete_cases = complete_cases,
            vars_s_model = vars_s_model,
            vars_y_model = vars_y_model,
            nb_mi = nb_mi,
            micemd_method = micemd_method,
            strategy = strategy,
            cw_type = cw_type)))
    } else {
        try(estimand <- unlist(estimator(RCT_RWD,
            outcome_name = outcome_name,
            method = method,
            complete_cases = complete_cases,
            vars_s_model = vars_s_model,
            vars_y_model = vars_y_model)))
    }
    if (!is.null(estimand)) {
        estimands <- rbind(estimands,
            data.frame(t(estimand)))
    } else {
        cat(paste0(i, "-> fail, "))
        ct_fail <- ct_fail +
            1
    }
}
if (as.character(substitute(estimator)) ==
    "compute_ipsw") {
    estimands <- data.frame(estimands)
    colnames(estimands) <- paste0(c("IPSW_",
        "IPSW.norm_"), method)
}
if (as.character(substitute(estimator)) ==
    "compute_all") {
    estimands <- data.frame(estimands)
    if (access_genRCT) {
        colnames(estimands) <- paste0(c("IPSW_",
            "IPSW.norm_", "G-formula_",
            "AIPSW_", "CW_"),
            method)
    } else {
        colnames(estimands) <- paste0(c("IPSW_",
```

```
                "IPSW.norm_", "G-formula_",
                "AIPSW_"), method)
        }
    }
    if (as.character(substitute(estimator)) ==
        "compute_all_mi") {
        estimands <- data.frame(estimands)
        estimands <- dplyr::select(estimands,
            -"nb_mi")
        if (access_genRCT) {
            colnames(estimands) <- paste0(c("IPSW_",
                "IPSW.norm_", "G-formula_",
                "AIPSW_", "CW_"),
                method)
        } else {
            colnames(estimands) <- paste0(c("IPSW_",
                "IPSW.norm_", "G-formula_",
                "AIPSW_"), method)
        }
    }
    print(paste0("Number of failed iterations: ",
        ct_fail))
    return(estimands)
}
```

ATE transported from CRASH-2 study to the Traumabase TBI patients

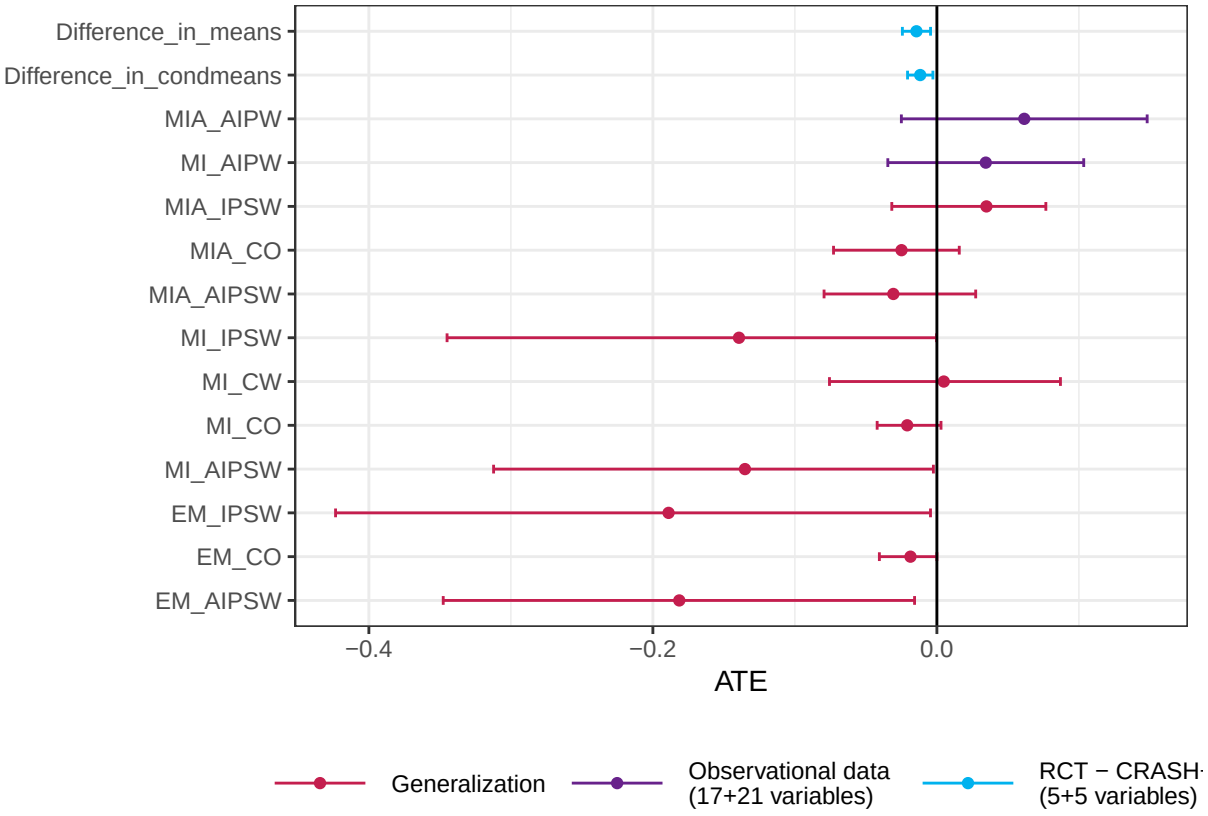
Note that when using the original Traumabase, the standard estimators (IPSW, CO, AIPSW) need to be adapted to handle missing values that are not missing completely at random (MCAR).

We propose three ways of addressing this handling of missing values:

- Logistic regression via Expectation Maximization (EM) that explicitly handles missing values that are missing at random (MAR).
- Generalized random forests that consider that missing values are potentially informative, this is achieved through the *missing incorporated in attributes* (MIA) criterion.
- Multilevel multiple imputation combined with parametric IPSW, CO and AIPSW estimation.

Estimation

Plot of the final results



Appendix

CRASH-3 analysis (results from CRASH-3 paper)

In this part we load the CRASH-3 data and reproduce the results in the publication with the risk ratio (RR). We also provide the results with the ATE to fit the framework of the review.

The outcome is the 28-day death due to brain injury (same output is taken in the Traumabase).

To recover the exact same results as presented in the CRASH-3 paper, we exclude patients with minimal GCS (equal to 3), or bilateral non-reactive pupils (*mydriasis*).

```
## risk_placebo     risk_TXA           RR     lower_ci     upper_ci
##   13.3245383   12.5483693    0.9417489    0.8435506    1.0399472
```

```
##          ATE     lower_ci     upper_ci
## -0.007761690 -0.020464312  0.004940932
```