Overlap analysis: distribution shift and sampling propensity score

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 proposes an analysis of the distributional shift between the CRASH-3 and Traumabase patients. The input is the merged table of both the randomized controlled trial and the observational study (corresponding to the output of `preprocess.Rmd`). The key functions to perform the analysis below come from the script `estimators.R`.

Contents

Load data and define subsets of variables

```
# load combined data
# (corresponds to the output of
# preprocess.Rmd)
DF <- read.csv("./Data/output_preprocess_combined_crash3_TB.csv")
DF <- DF[, 2:ncol(DF)]
trial_eligibility <- c("majorExtracranial",
    "age", "Glasgow.initial")
outcome_impact <- c("systolicBloodPressure",
    "sexe", "pupilReact_num")

# remove outcome and treatment
# from the table for analysis
DF <- DF[, c(trial_eligibility,
    outcome_impact, "V")]
```

We first compute the sampling propensity score, that corresponds to:

$$P(S = 1|X)$$

We recall that the covariates of interest kept in the preprocess part corresponds to:

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

- **majorExtracranial** (binary)
- **age** (continuous)
- **Glasgow.initial** (continuous)
- **systolicBloodPressure** (continuous)
- **sexe** (binary)
- **pupilReact_num** (continuous)

Estimation of propensity scores

```
# estimation of the sampling
# propensity scores with two
# methods
pi_s_hat_glm <- sampling_propensities(DF,
    method = "glm", seed = 100)
```

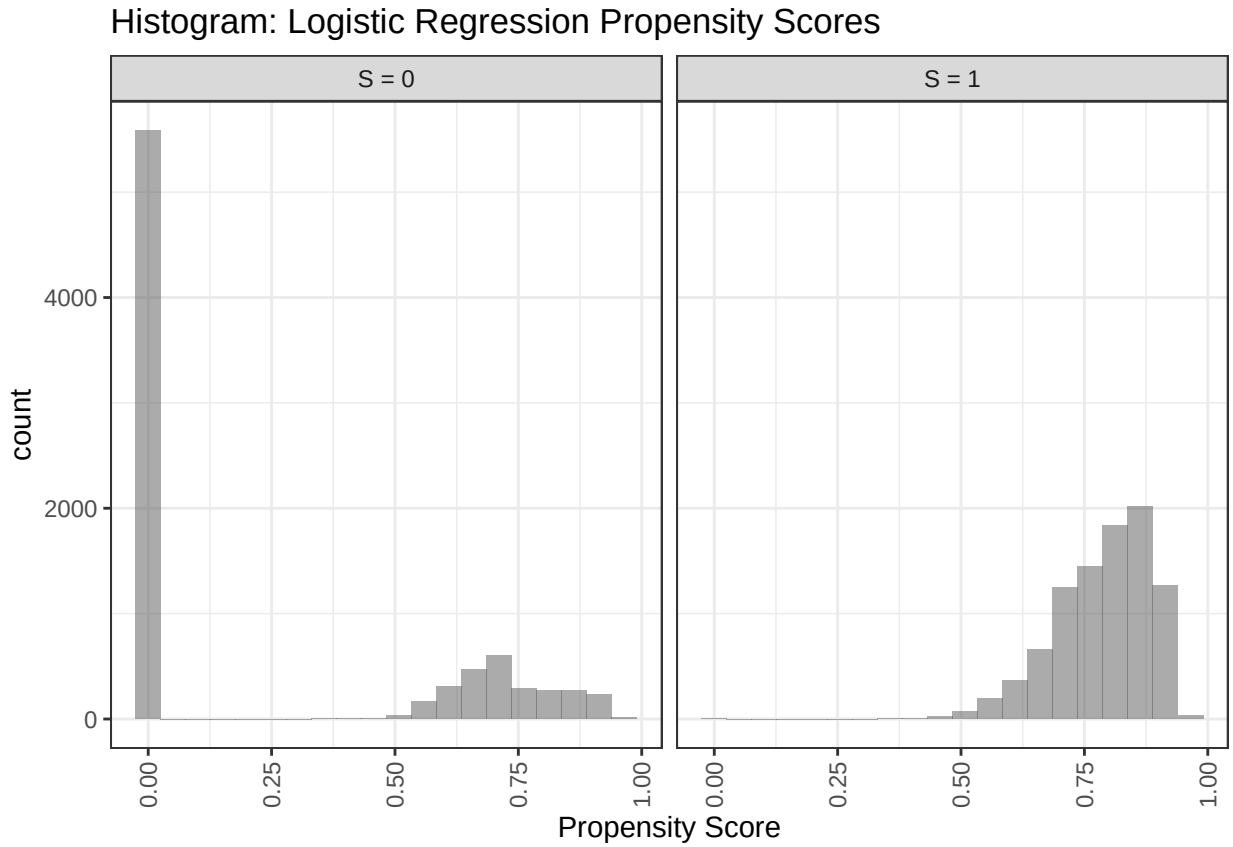
```
## Iteration of SAEM:
## 50
```

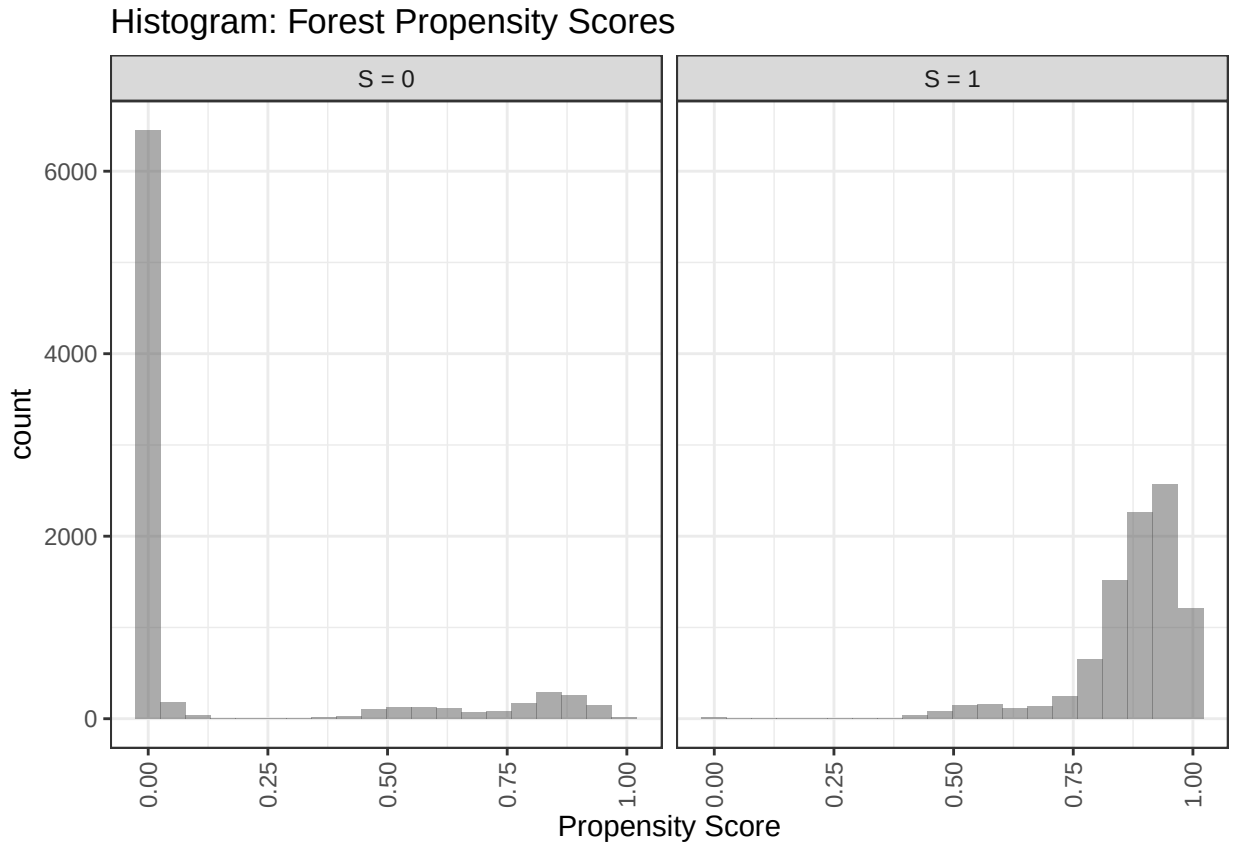
```
pi_s_hat_grf <- sampling_propensities(DF,
    method = "grf", seed = 100)
```

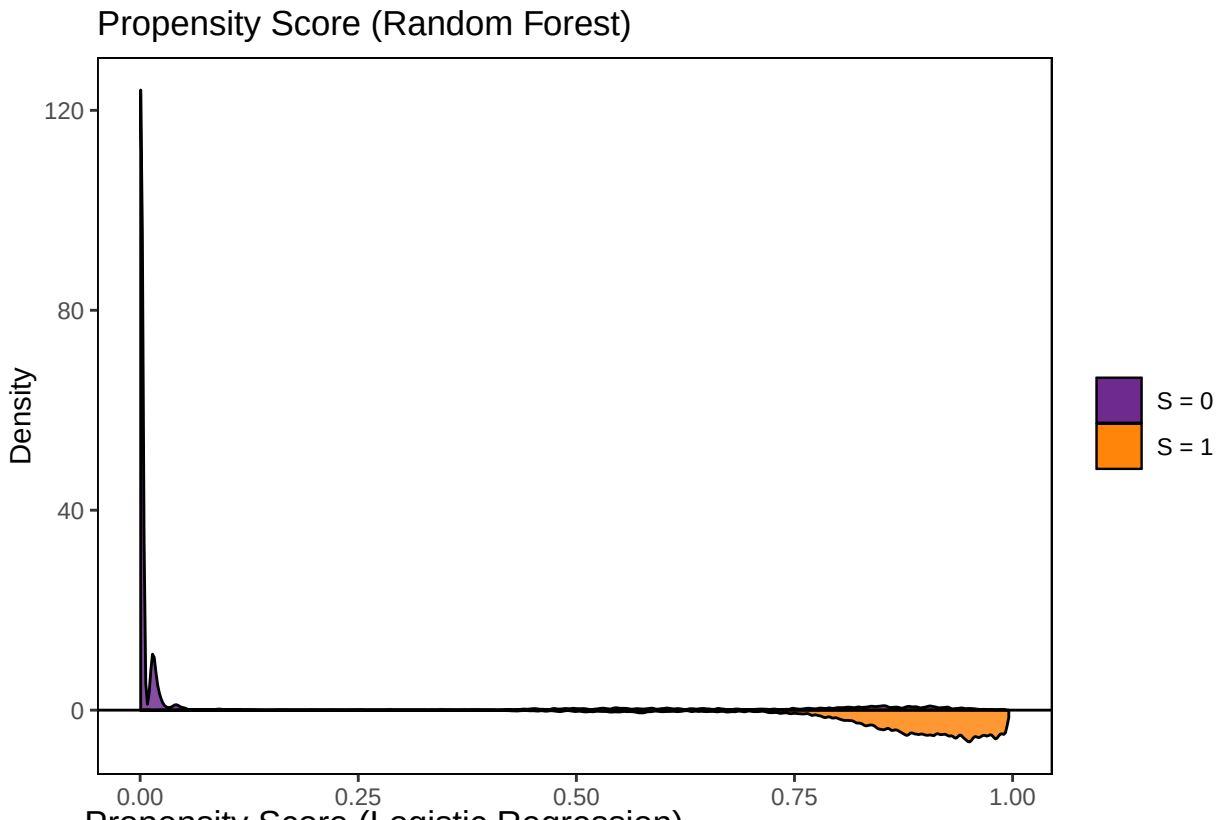
```
# visualization and analysis of
# these sampling propensity
# scores
plot_propensity_scores <- function(DF,
    ps_glm, ps_grf) {
    propensity_scores <- data.frame(logit = ps_glm,
        grf = ps_grf, RCT = paste0("S = ",
            as.factor(DF$V)), extraCranialBleeding = as.factor(DF$majorExtracranial),
        SystolicBloodPressure = DF$systolicBloodPressure,
        GCS = DF$Glasgow.initial,
        pupil = DF$pupilReact_num,
        sexe = DF$sexe, age = DF$age)

    g <- ggplot(propensity_scores,
        aes(x = logit)) + geom_histogram(bins = 20,
        alpha = 0.5) + facet_wrap(~RCT) +
        theme_bw() + theme(axis.text.x = element_text(angle = 90,
        vjust = 0.5)) + labs(title = "Histogram: Logistic Regression Propensity Scores",
        x = "Propensity Score")

    print(g)

    g <- ggplot(propensity_scores,
        aes(x = grf)) + geom_histogram(bins = 20,
        alpha = 0.5) + facet_wrap(~RCT) +
        theme_bw() + theme(axis.text.x = element_text(angle = 90,
        vjust = 0.5)) + labs(title = "Histogram: Forest Propensity Scores",
        x = "Propensity Score")

    print(g)

    g <- bal.plot(x = data.frame(grf = propensity_scores$grf),
        var.name = "grf", which = "unadjusted",
```

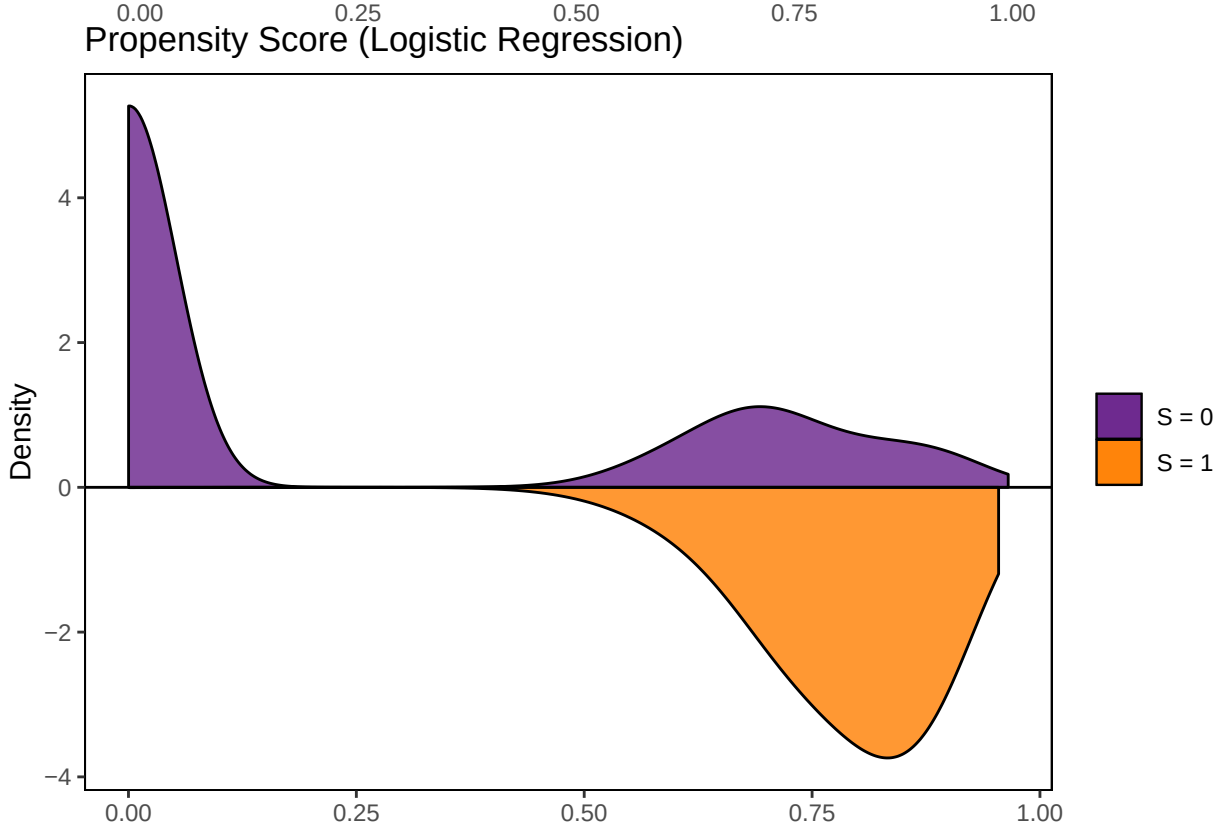
```
    treat = propensity_scores$RCT,
    colors = c("darkorchid4",
        "darkorange1"), type = "density",
    mirror = TRUE) + theme(legend.title = element_blank(),
    strip.text = element_blank()) +
    ggtitle("Propensity Score (Random Forest)") +
    xlab("")
print(g)


g <- bal.plot(x = data.frame(glm = propensity_scores$logit),
    var.name = "glm", which = "unadjusted",
    treat = propensity_scores$RCT,
    colors = c("darkorchid4",
        "darkorange1"), type = "density",
    mirror = TRUE) + theme(legend.title = element_blank(),
    strip.text = element_blank()) +
    ggtitle("Propensity Score (Logistic Regression)") +
    xlab("")
print(g)


g <- ggplot(propensity_scores,
    aes(x = logit, y = grf,
        color = RCT)) + geom_point(alpha = 0.5) +
    theme_bw() + xlab("Propensity Score (Logistic Regression)") +
    ylab("Propensity Score (Random Forest)")

print(g)

g <- ggplot(propensity_scores,
    aes(x = logit, y = grf,
        color = extraCranialBleeding)) +
    geom_point(alpha = 0.5) +
    theme_bw() + xlab("Propensity Score (Logistic Regression)") +
    ylab("Propensity Score (Random Forest)")

print(g)

g <- ggplot(propensity_scores,
    aes(x = logit, y = grf,
        color = SystolicBloodPressure)) +
    geom_point(alpha = 0.5) +
    theme_bw() + xlab("Propensity Score (Logistic Regression)") +
    ylab("Propensity Score (Random Forest)")

print(g)

g <- ggplot(propensity_scores,
    aes(x = logit, y = grf,
        color = GCS)) + geom_point(alpha = 0.5) +
    theme_bw() + xlab("Propensity Score (Logistic Regression)") +
    ylab("Propensity Score (Random Forest)")
```

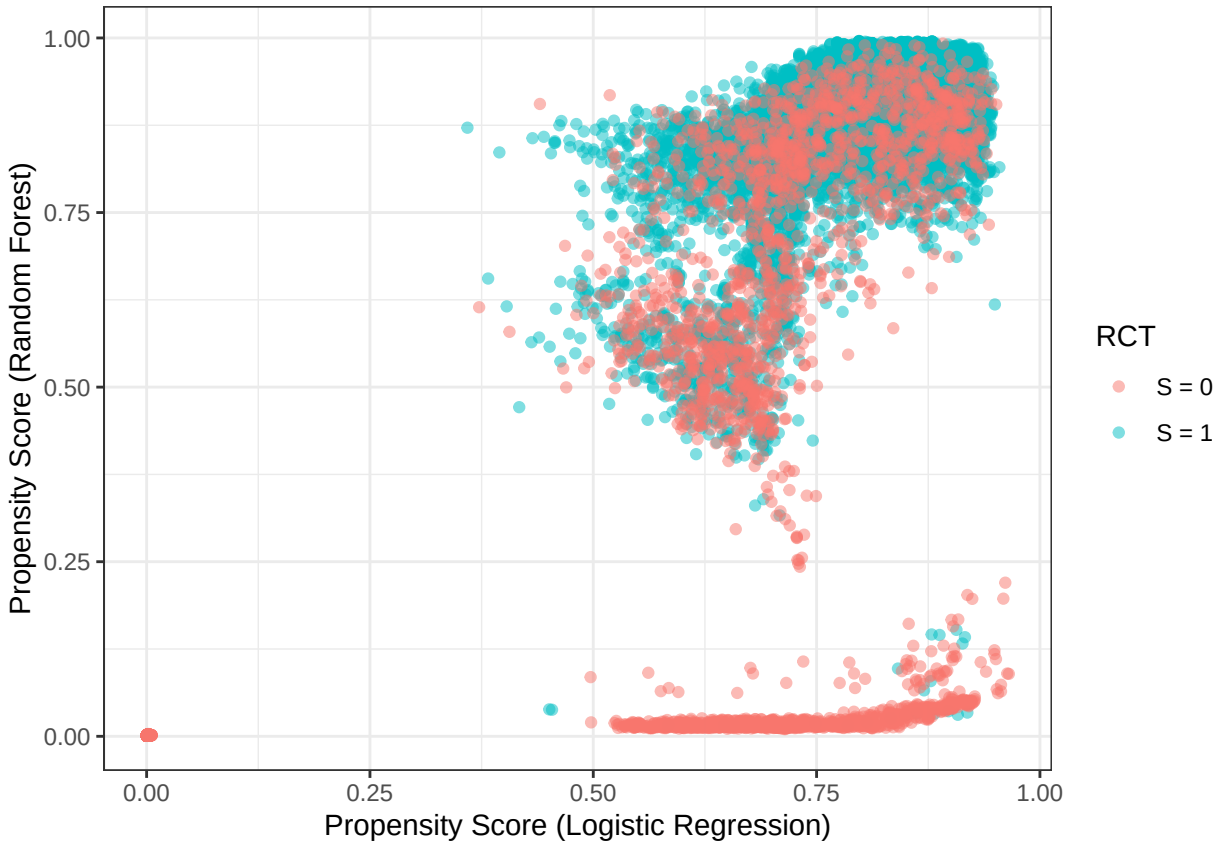
```
    print(g)

    g <- ggplot(propensity_scores,
        aes(x = logit, y = grf,
            color = age)) + geom_point(alpha = 0.5) +
        theme_bw() + xlab("Propensity Score (Logistic Regression)") +
        ylab("Propensity Score (Random Forest)")

    print(g)

    g <- ggplot(propensity_scores,
        aes(x = logit, y = grf,
            color = sexe)) + geom_point(alpha = 0.5) +
        theme_bw() + xlab("Propensity Score (Logistic Regression)") +
        ylab("Propensity Score (Random Forest)")

    print(g)

    g <- ggplot(propensity_scores,
        aes(x = logit, y = grf,
            color = pupil)) + geom_point(alpha = 0.5) +
        theme_bw() + xlab("Propensity Score (Logistic Regression)") +
        ylab("Propensity Score (Random Forest)")

    print(g)
}

plot_propensity_scores(DF, pi_s_hat_glm,
    pi_s_hat_grf)
```

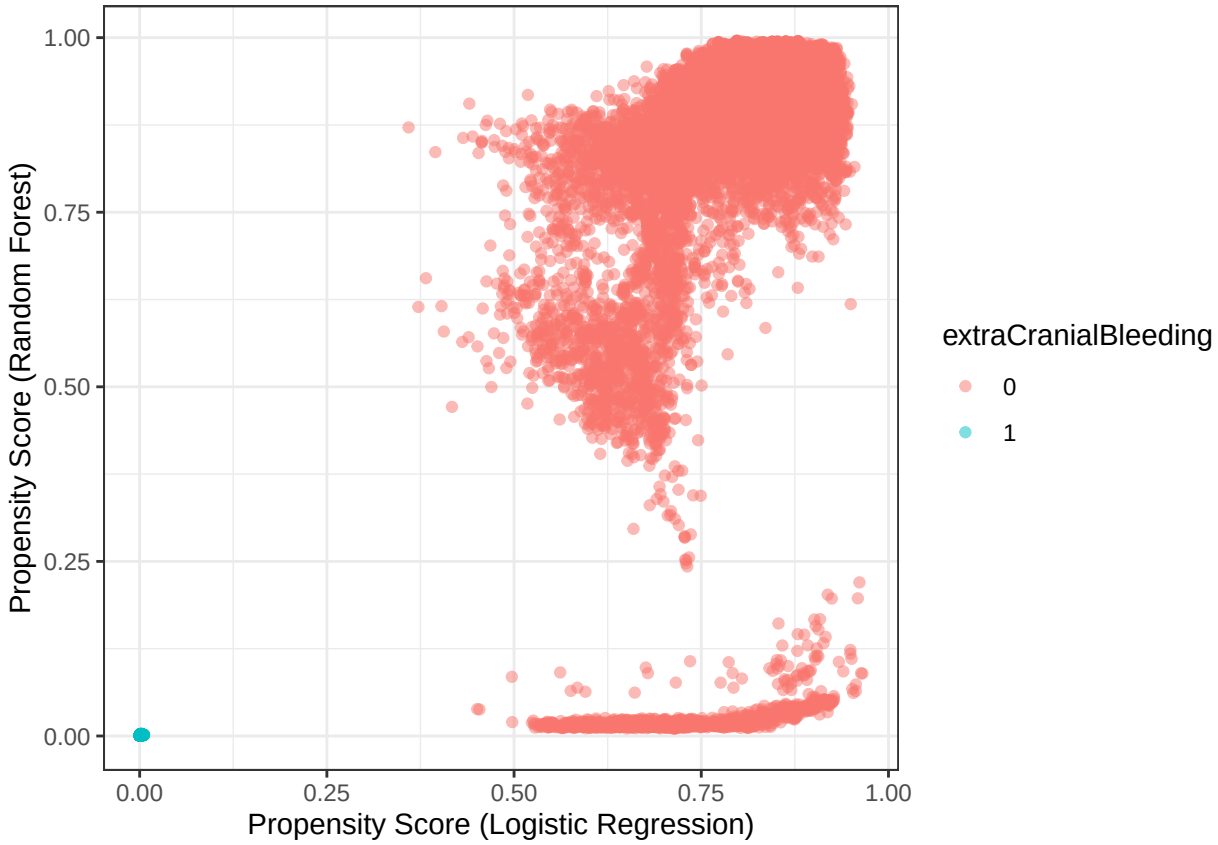

Histogram: Logistic Regression Propensity Scores
S = 0
S = 1
count
4000
2000
0
0.00
0.25
0.50
0.75
1.00
Propensity Score

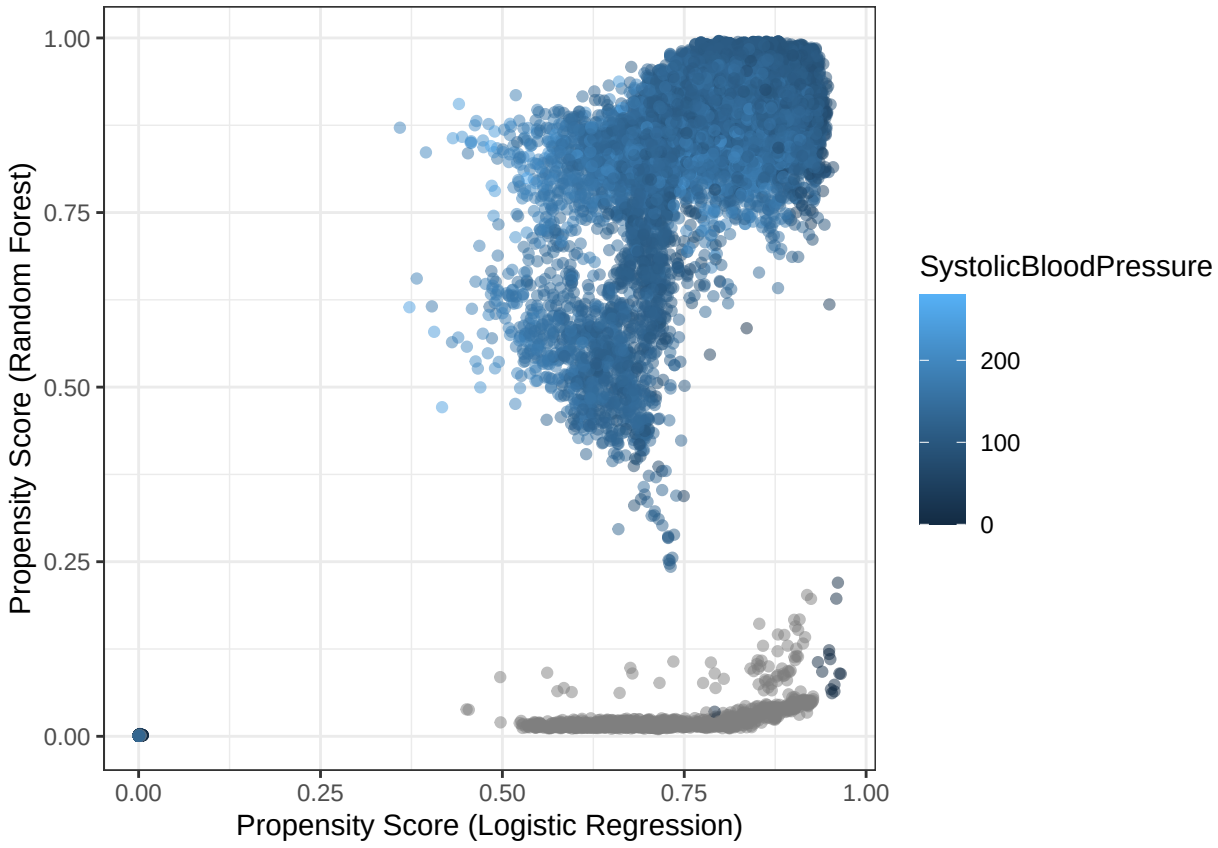

Histogram: Forest Propensity Scores
S = 0
S = 1
count
6000
4000
2000
0
0.00
0.25
0.50
0.75
1.00
Propensity Score

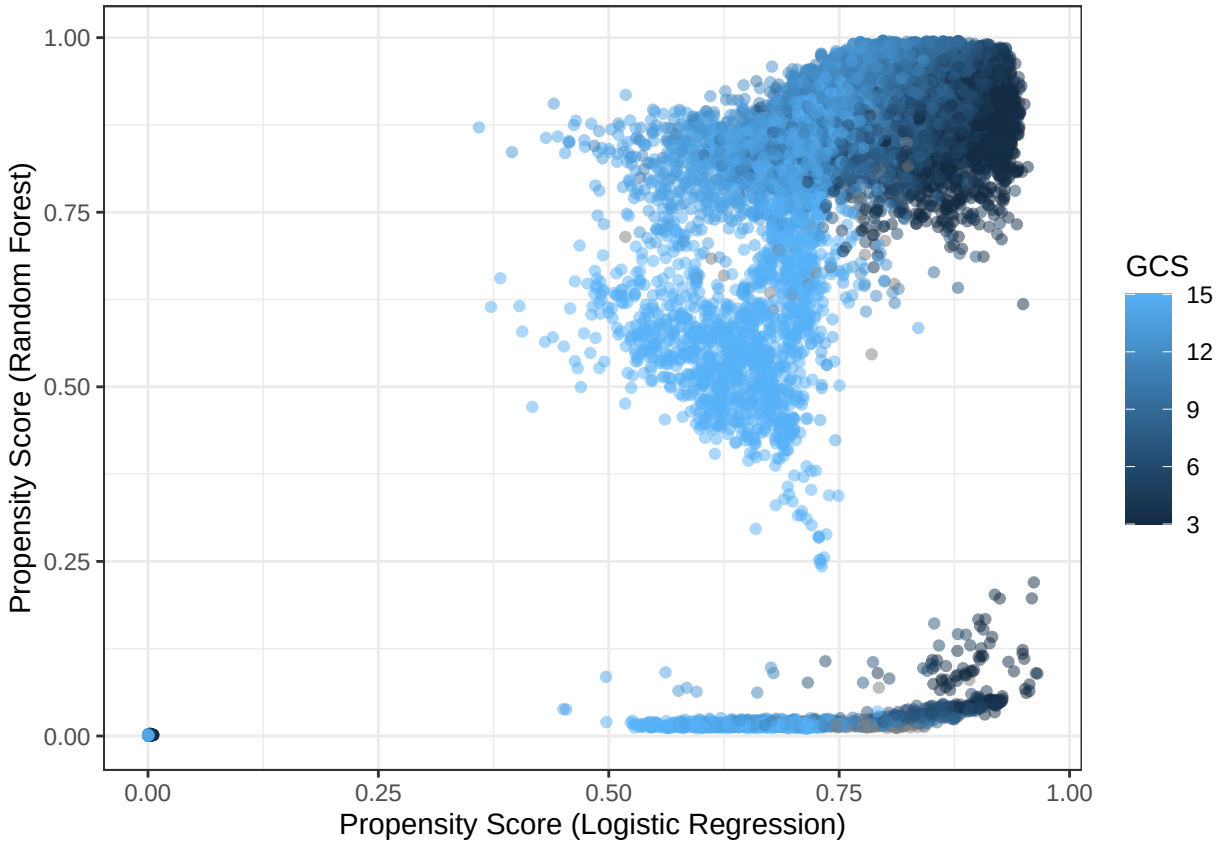

Propensity Score (Random Forest)
Density
120
80
40
0
0.00
0.25
0.50
0.75
1.00
S = 0
S = 1

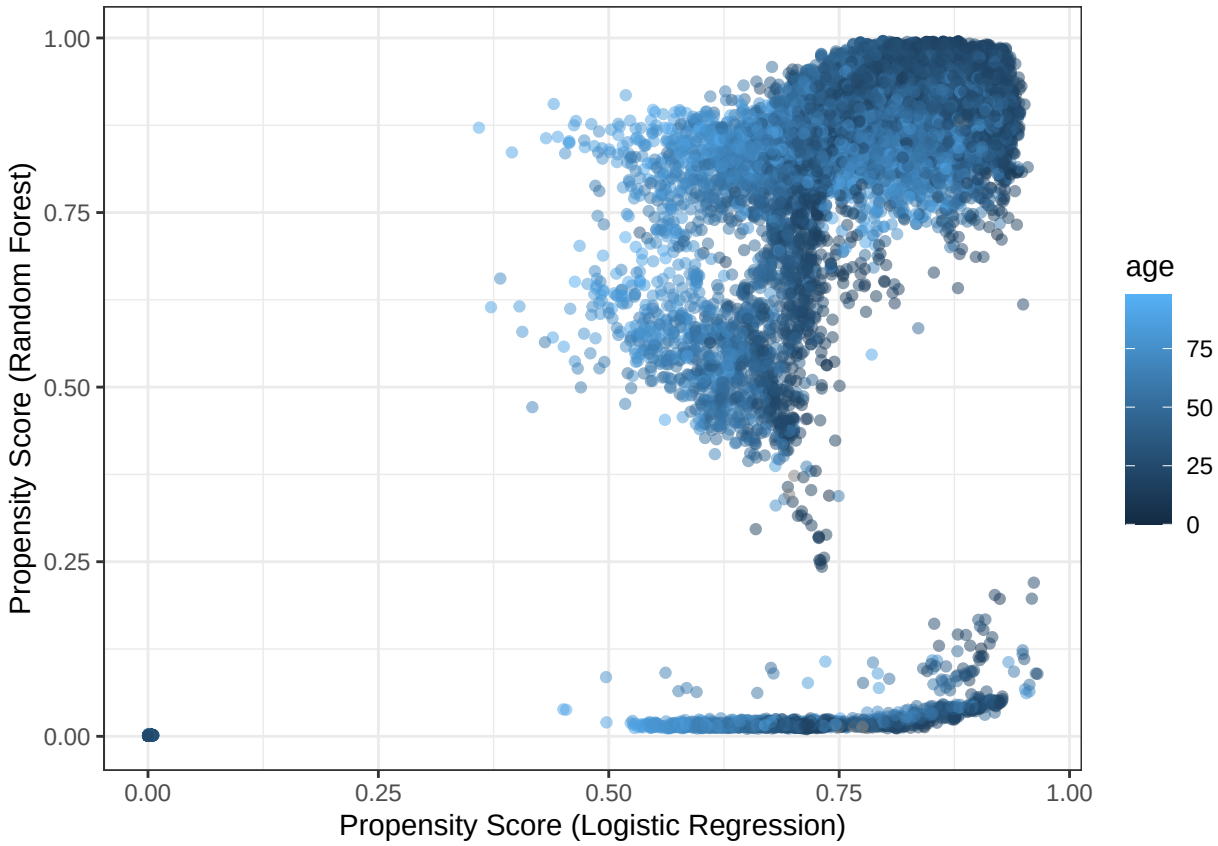

Propensity Score (Logistic Regression)
Density
4
2
0
−2
−4
0.00
0.25
0.50
0.75
1.00
S = 0
S = 1

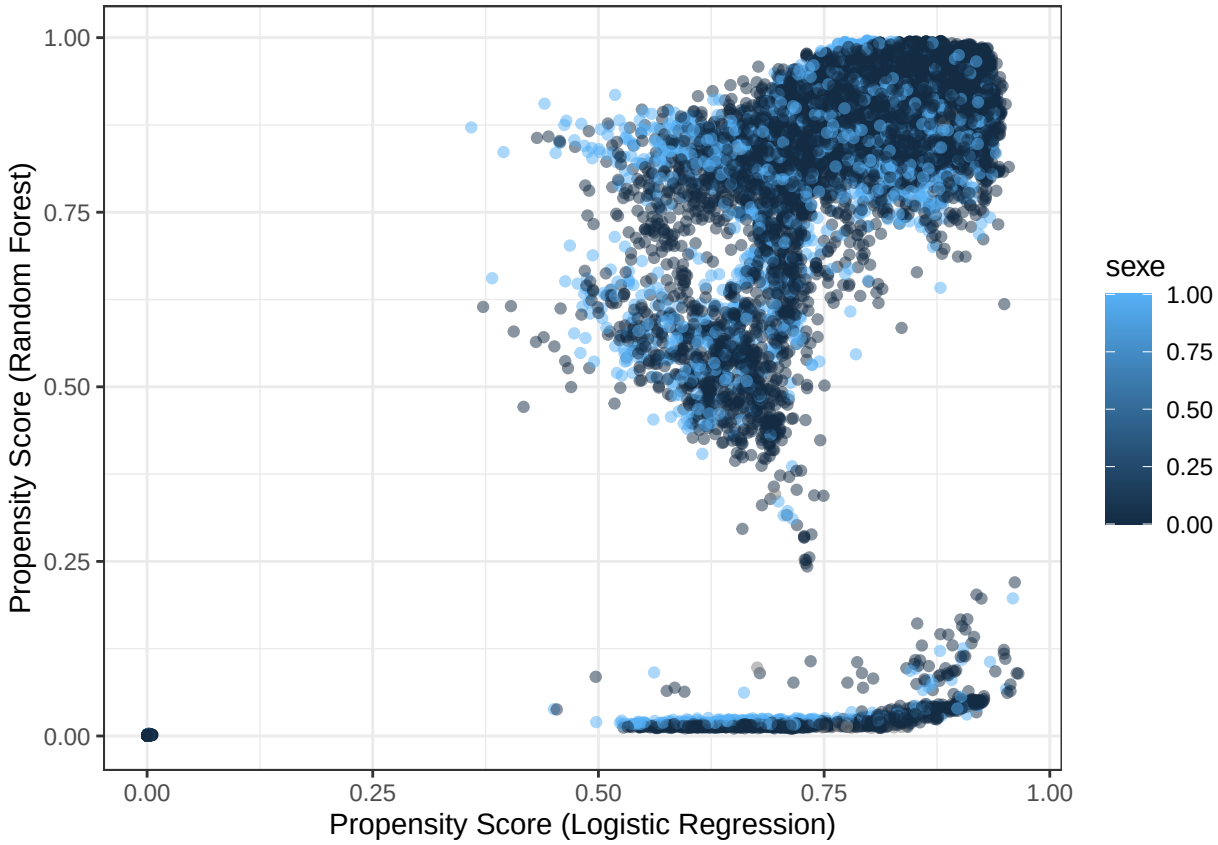

1.00
0.75
0.50
0.25
0.00
Propensity Score (Random Forest)
0.00
0.25
0.50
0.75
1.00
Propensity Score (Logistic Regression)
RCT
S = 0
S = 1

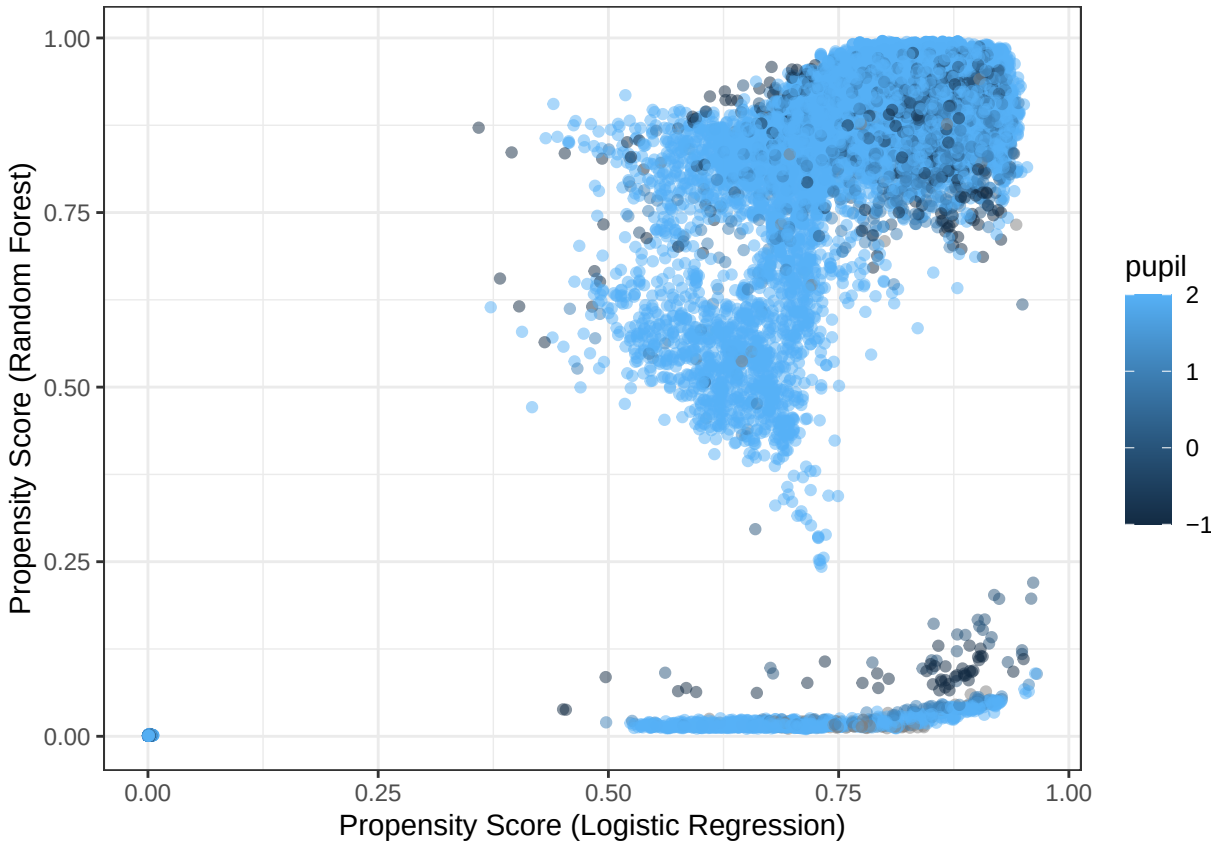

1.00
0.75
0.50
0.25
0.00
Propensity Score (Random Forest)
0.00
0.25
0.50
0.75
1.00
Propensity Score (Logistic Regression)
extraCranialBleeding
0
1


Propensity Score (Random Forest)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Propensity Score (Logistic Regression)
SystolicBloodPressure
200
100
0

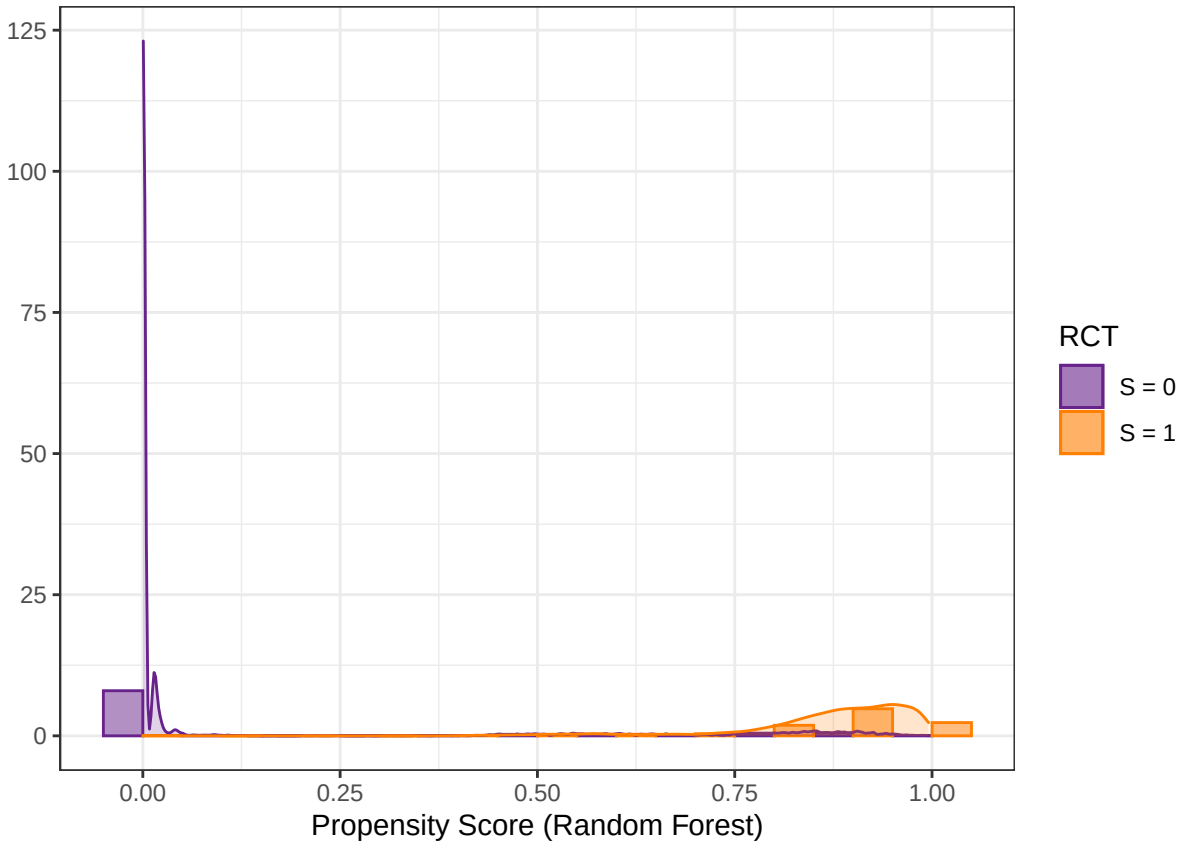

Propensity Score (Random Forest)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Propensity Score (Logistic Regression)
GCS
15
12
9
6
3


Propensity Score (Random Forest)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Propensity Score (Logistic Regression)
age
75
50
25
0


Propensity Score (Random Forest)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Propensity Score (Logistic Regression)
sexe
1.00
0.75
0.50
0.25
0.00



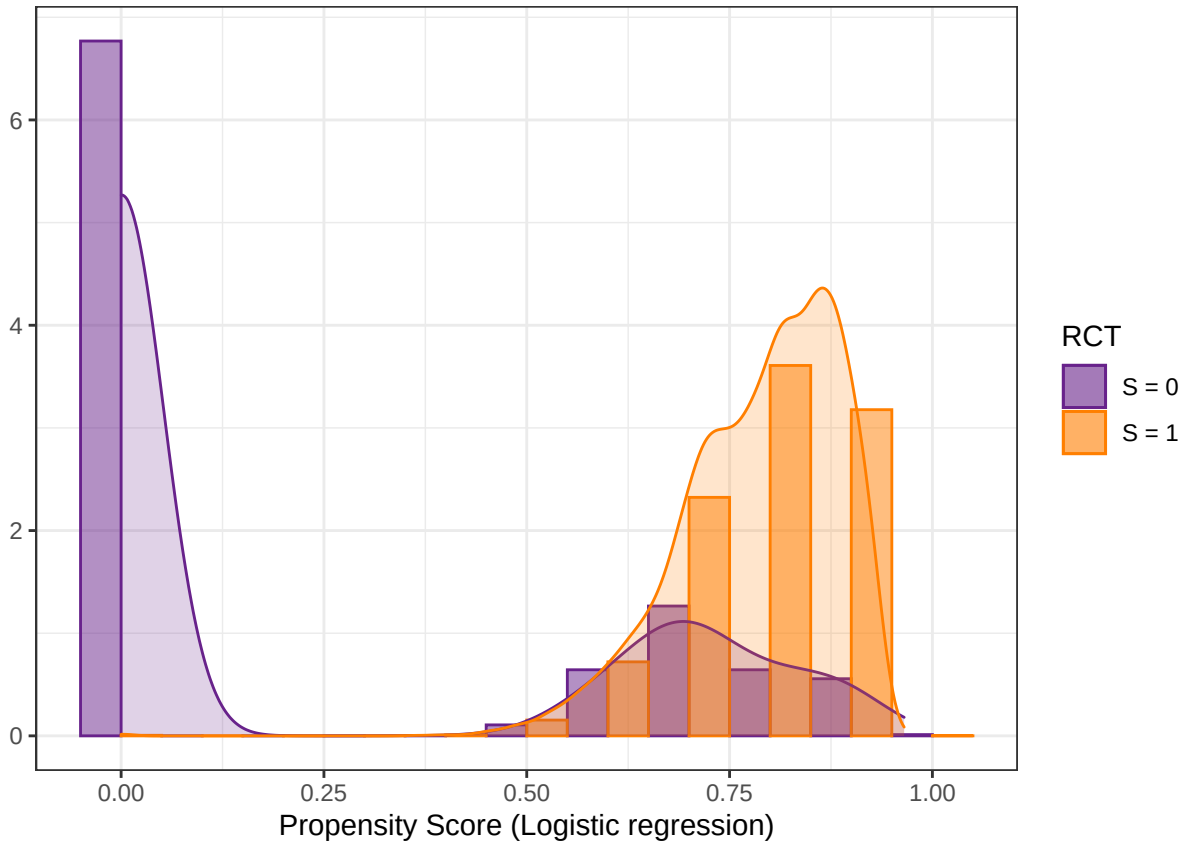

```
# Exact figures from the review
propensity_scores <- data.frame(logit = pi_s_hat_glm,
    grf = pi_s_hat_grf, RCT = paste0("S = ",
        as.factor(DF$V)), extraCranialBleeding = as.factor(DF$majorExtracranial),
    SystolicBloodPressure = DF$systolicBloodPressure,
    GCS = DF$Glasgow.initial, pupil = DF$pupilReact_num,
    sexe = DF$sexe, age = DF$age)
ggplot(propensity_scores, aes(x = grf,
    group = RCT, fill = RCT, colour = RCT)) +
    geom_histogram(aes(y = ..density..),
        position = "dodge", alpha = 0.5,
        binwidth = 0.1) + geom_density(alpha = 0.2) +
    scale_fill_manual(values = c("darkorchid4",
        "darkorange1")) + scale_colour_manual(values = c("darkorchid4",
    "darkorange1")) + theme_bw() +
    ylab("") + xlab("Propensity Score (Random Forest)")
```




```
ggsave("./Figures/propensity_scores_all_grf.png")
```

```
## Saving 6.5 x 4.5 in image
```

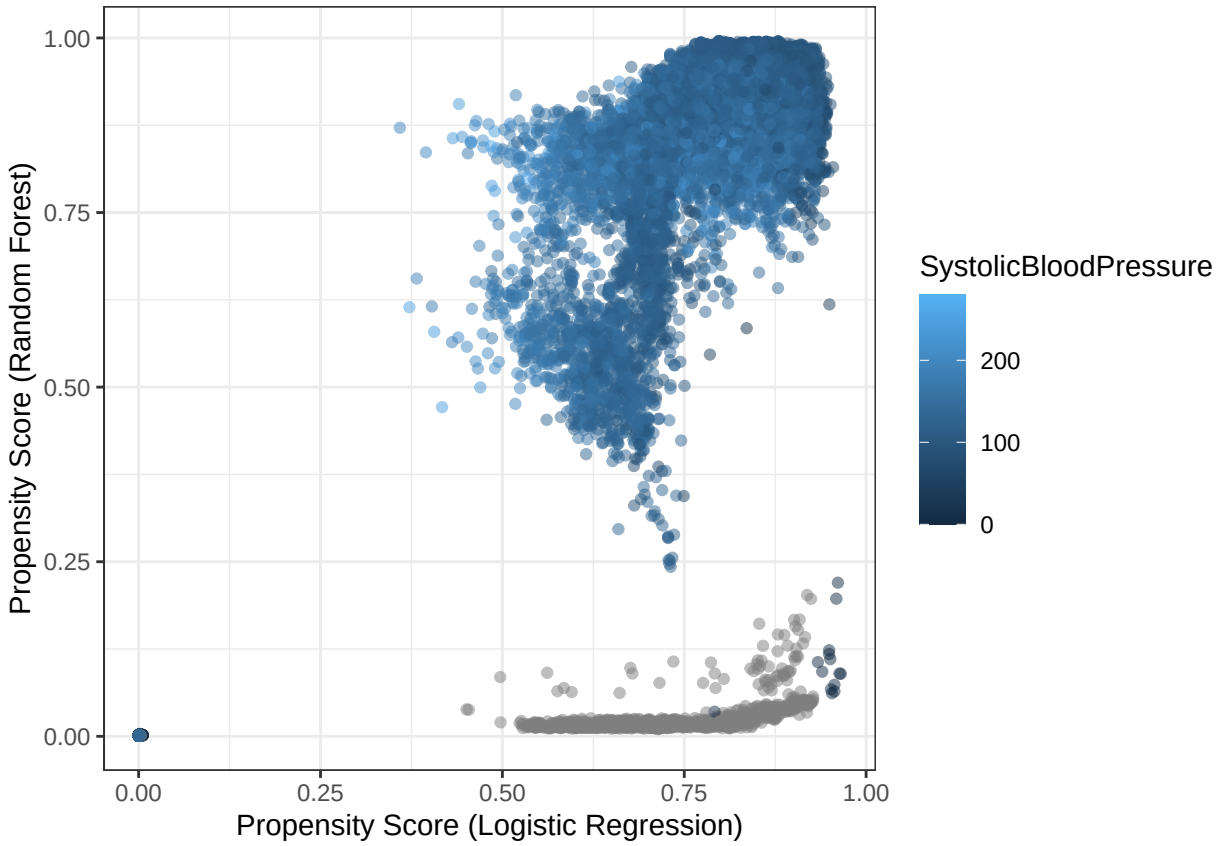
```
ggplot(propensity_scores, aes(x = logit,
    group = RCT, fill = RCT, colour = RCT)) +
    geom_histogram(aes(y = ..density..),
        position = "dodge", alpha = 0.5,
        binwidth = 0.1) + geom_density(alpha = 0.2) +
    scale_fill_manual(values = c("darkorchid4",
        "darkorange1")) + scale_colour_manual(values = c("darkorchid4",
    "darkorange1")) + theme_bw() +
    ylab("") + xlab("Propensity Score (Logistic regression)")
```




```
ggsave("./Figures/propensity_scores_all_glm.png")
```

```
## Saving 6.5 x 4.5 in image
```

```
ggplot(propensity_scores, aes(x = logit,
    y = grf, color = SystolicBloodPressure)) +
    geom_point(alpha = 0.5) + theme_bw() +
    xlab("Propensity Score (Logistic Regression)") +
    ylab("Propensity Score (Random Forest)")
```




```
ggsave("./Figures/scatter_plot_propensity_scores_systolic.png")
```

```
## Saving 6.5 x 4.5 in image
```

Comparison of different approaches to handle missing values

We observe the random forest have more extreme values than the logistic regression. We try to understand why. We draw the scatter plot of propensity scores. We observe that a set of values have almost a zero-probability predicted by the forest, while it seems to be different for the logistic regression. These values are almost all in the observational data set. We also observe that the `majorExtracranial` characteristic is more present in the Traumabase, and that both propensity scores identify it correctly. We also observe that the forest seems to consider missing values for the systolic blood pressure as an indicator of being or not in the Traumabase (which is not necessary a good choice, more importantly because missing values of the systolic blood pressure is not missing completely at random as the absence of the data can reflect the gravity of the patient state).

To further investigate the problem, we can analyze the coefficients of either of the propensity scores algorithm.

```
# logistic regressing
p.fit <- miss.glm(V ~ ., data = DF)
```

```
## Iteration of SAEM:
## 50
```

```
print(p.fit)
```

```
##
## Call:  miss.glm(formula = V ~ ., data = DF)
##
## Coefficients:
```

```
##           (Intercept)      majorExtracranial                    age
##              3.878419              -8.676907             -0.008423
##       Glasgow.initial  systolicBloodPressure                   sexe
##             -0.146138              -0.007294             -0.065250
##        pupilReact_num
##              0.157587
## Standard error estimates:
##           (Intercept)      majorExtracranial                    age
##              0.159883               0.449859              0.001284
##       Glasgow.initial  systolicBloodPressure                   sexe
##              0.007086               0.001085              0.056345
##        pupilReact_num
##              0.037176
## Log-likelihood: -6015
```

```
# forest
na.action <- options()$na.action
options(na.action = "na.pass")
if (is.data.frame(DF)) {
    X.m = model.matrix(~. - 1,
        data = DF[, !names(DF) %in%
            c("V")])
} else {
    # X can also be a matrix
    X.m = model.matrix(~. - 1,
        data = data.frame(DF[,
            !names(DF) %in% c("V")]))
}
options(na.action = na.action)

forest.W = regression_forest(X.m,
    DF$V, tune.parameters = "all")
print(forest.W)
```

```
## GRF forest object of type regression_forest
## Number of trees: 2000
## Number of training samples: 17416
## Variable importance:
##     1     2     3     4     5     6
## 0.616 0.042 0.085 0.220 0.012 0.024
```

As a conclusion, we observe that the GRF forest retains the extracranial bleeding and then the systolic blood pressurs as the most important variables to account for the RCT inclusion. While the absence of extracranial bleeding is a criterion for inclusion in the RCT, the systolic blood pressure may be chosen for a bad reason. Glasgow score and pupils reactivity are more important in the logistic regression coefficients, which accounts for a real distribution difference between the two data as the Traumabase contains more less complicated patients.

To further investigate and understand the importance of the `NA` values in the systolic blood pressure covariates, we can perform the same analysis as the previous, but with an imputed data.

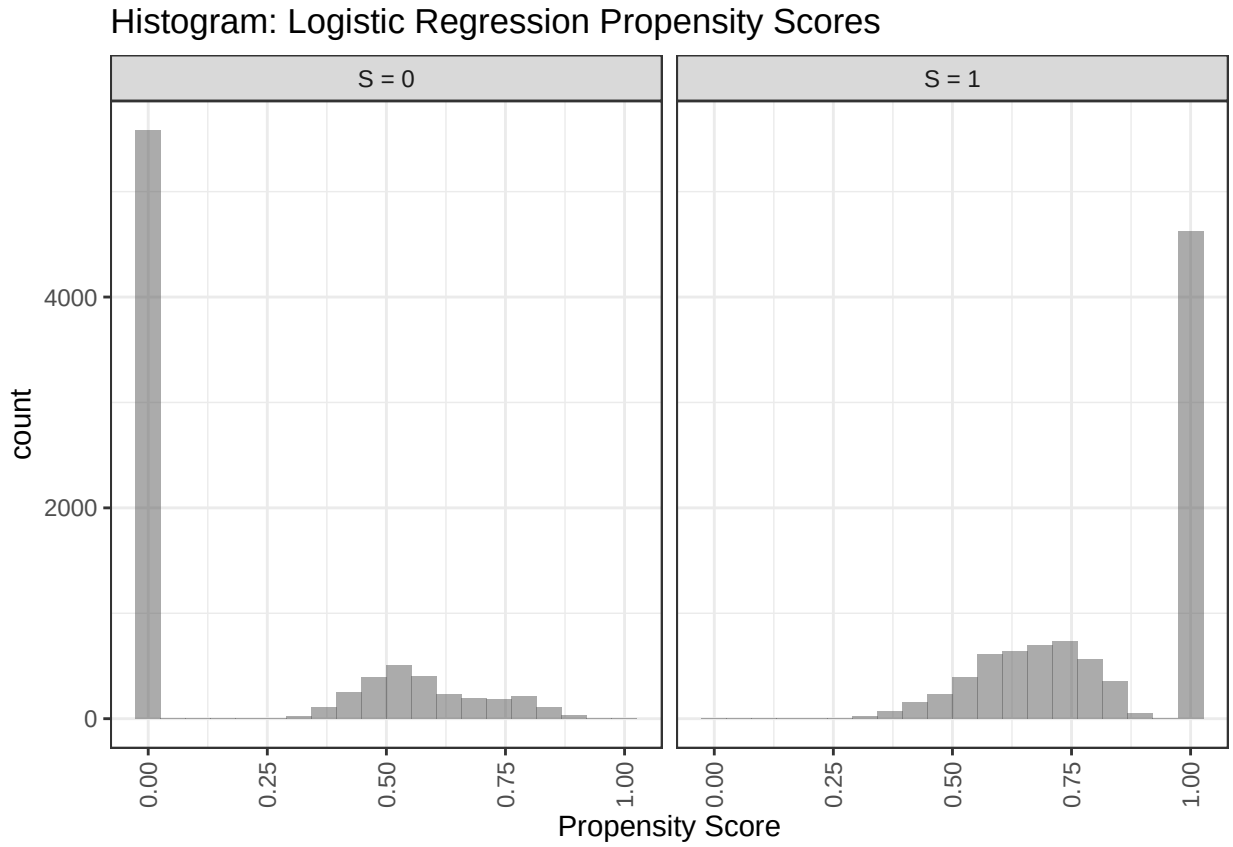
```
DF_imputed <- read.csv("./Data/output_preprocess_combined_crash3_TB_imputed.csv")
DF_imputed <- DF_imputed[, 2:ncol(DF_imputed)]

# estimation of the sampling
# propensity scores with two
```

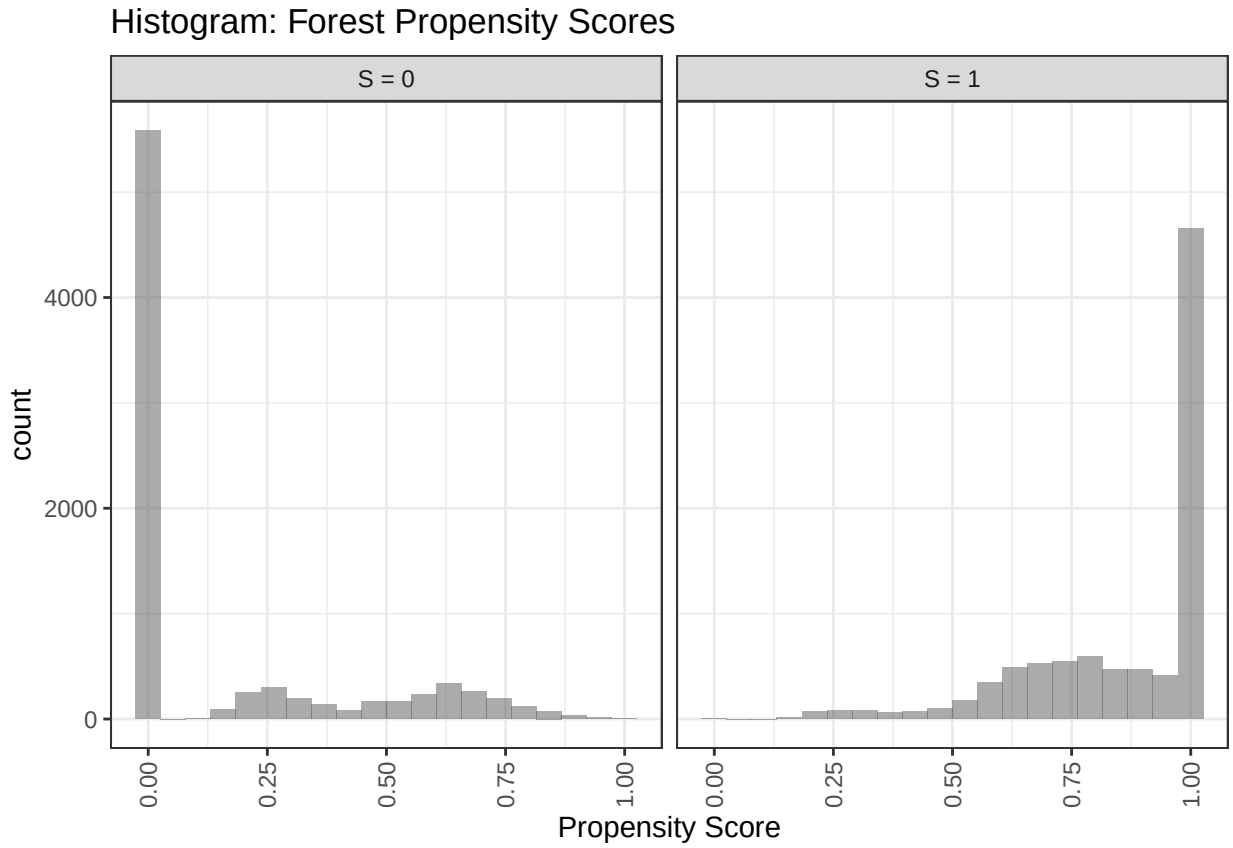
```r
# methods
pi_s_hat_glm_imputed <- sampling_propensities(DF_imputed,
    method = "glm", seed = 100)
```

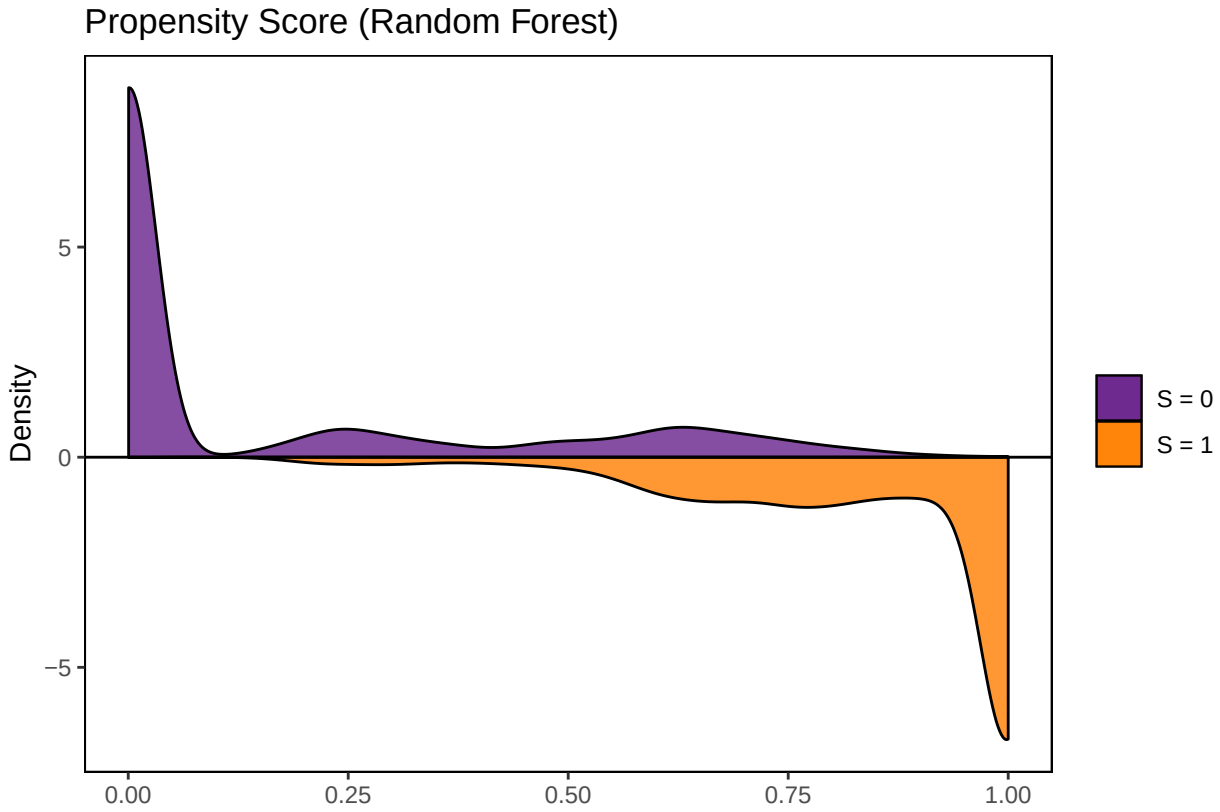
```
## Iteration of SAEM:
## 50
```

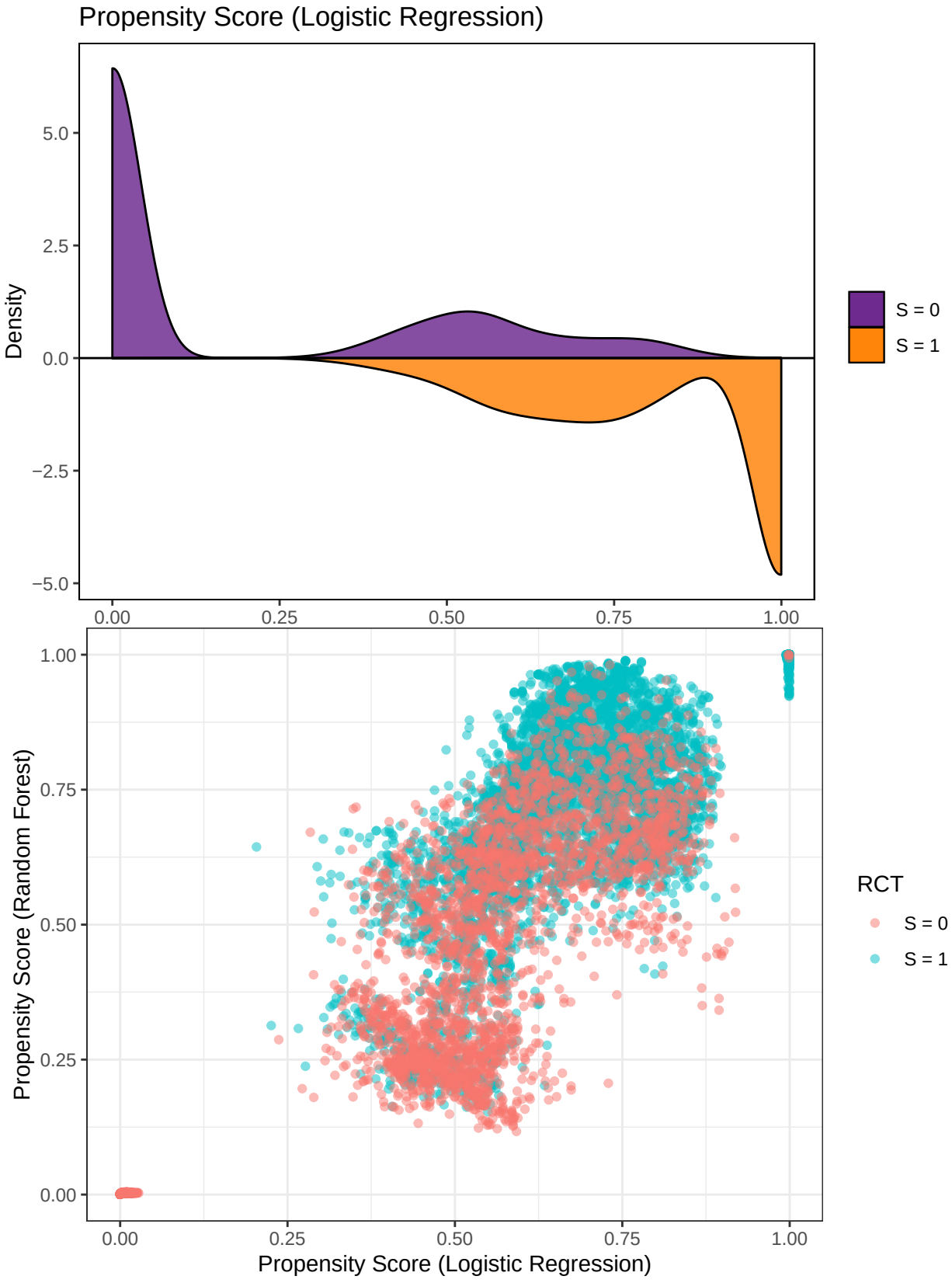
```r
pi_s_hat_grf_imputed <- sampling_propensities(DF_imputed,
    method = "grf", seed = 100)
```

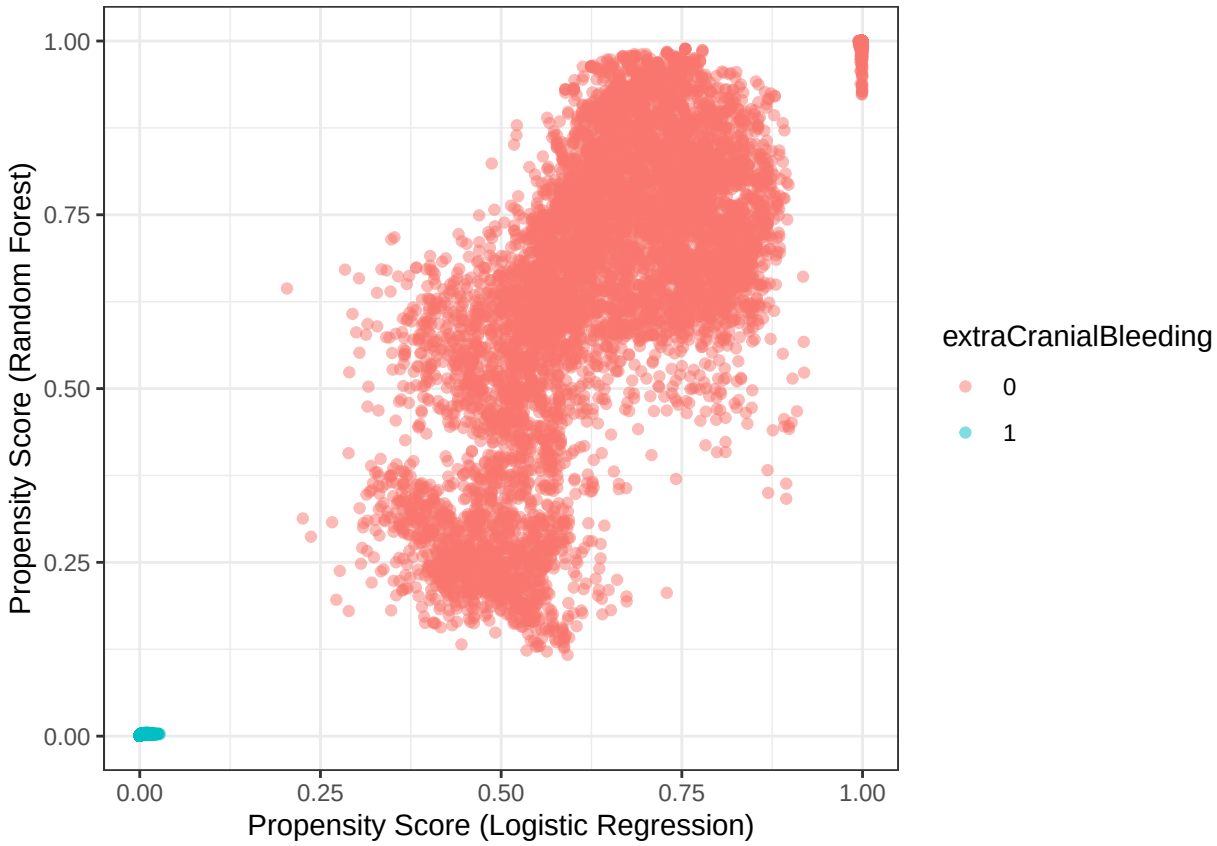
```r
plot_propensity_scores(DF = DF_imputed,
    pi_s_hat_glm_imputed, pi_s_hat_grf_imputed)
```

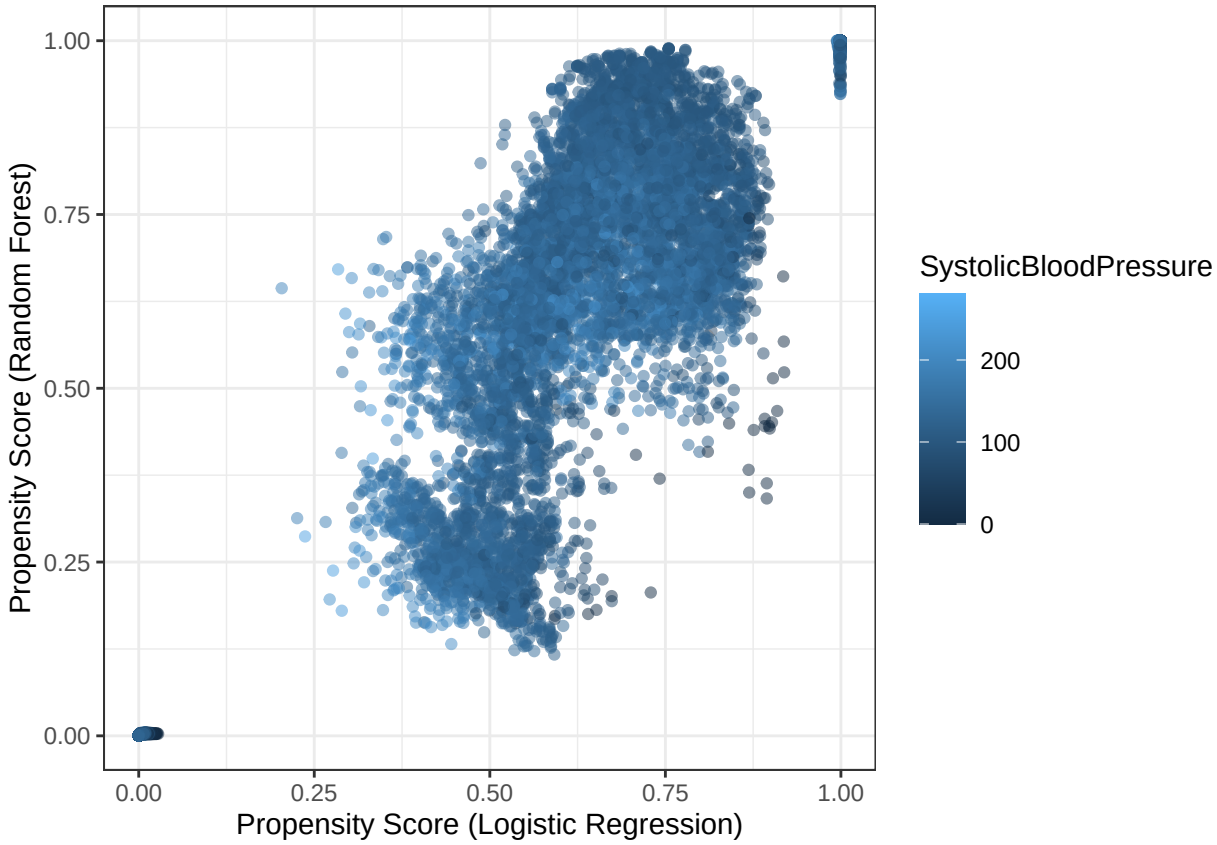


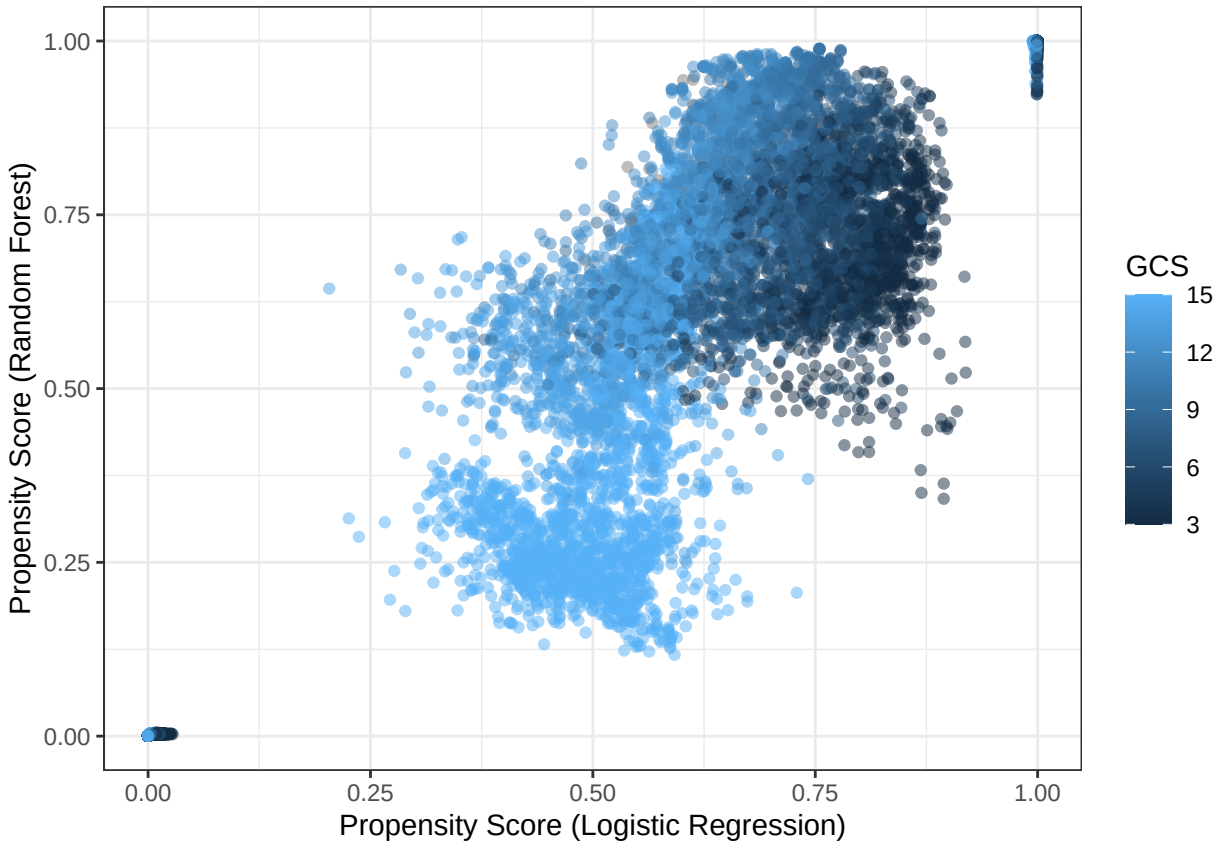

Histogram: Forest Propensity Scores
S = 0
S = 1
count
4000
2000
0
0.00
0.25
0.50
0.75
1.00
Propensity Score

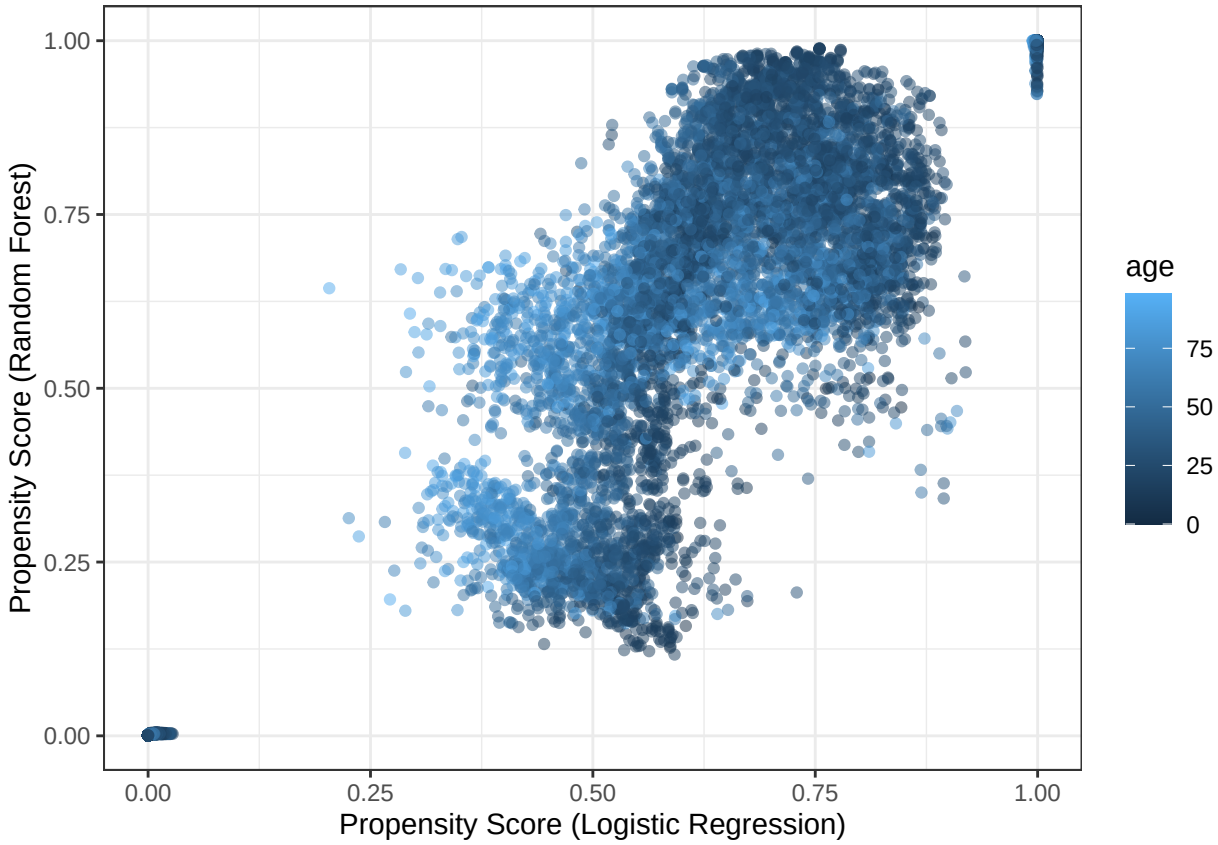

Propensity Score (Random Forest)
Density
5
0
−5
0.00
0.25
0.50
0.75
1.00
S = 0
S = 1

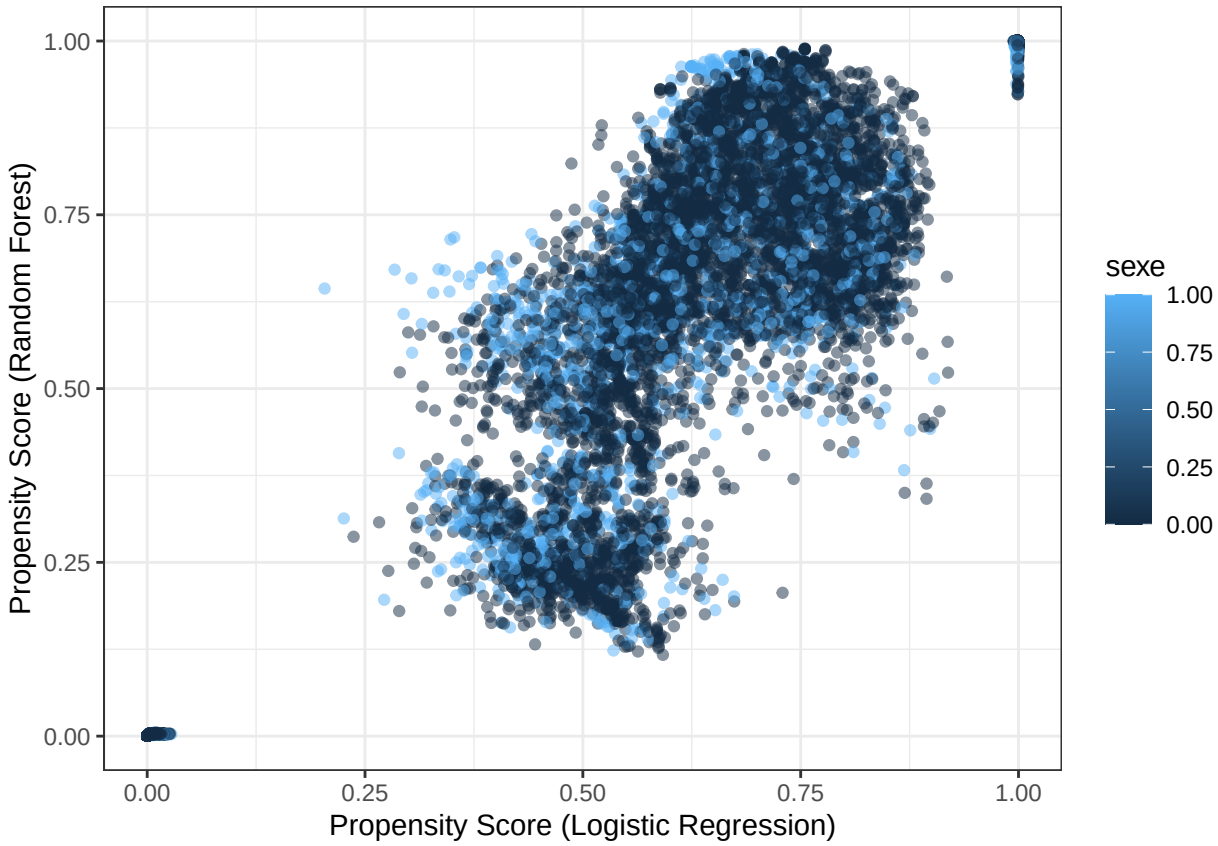

Propensity Score (Logistic Regression)
Density
5.0
2.5
0.0
−2.5
−5.0
0.00
0.25
0.50
0.75
1.00
S = 0
S = 1
Propensity Score (Random Forest)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Propensity Score (Logistic Regression)
RCT
S = 0
S = 1

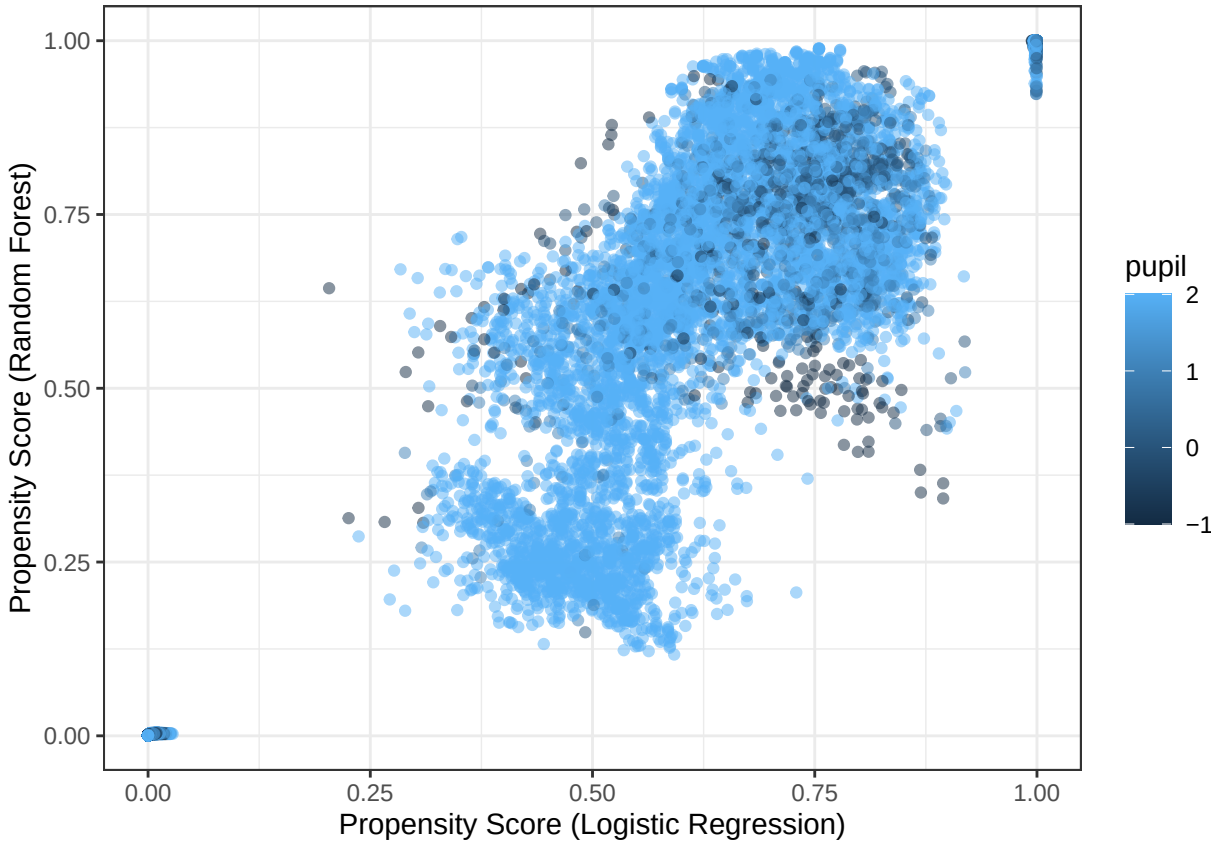

1.00
0.75
0.50
0.25
0.00
Propensity Score (Random Forest)
0.00
0.25
0.50
0.75
1.00
Propensity Score (Logistic Regression)
extraCranialBleeding
0
1


1.00
0.75
0.50
0.25
0.00
Propensity Score (Random Forest)
0.00
0.25
0.50
0.75
1.00
Propensity Score (Logistic Regression)
SystolicBloodPressure
200
100
0


Propensity Score (Random Forest)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Propensity Score (Logistic Regression)
GCS
15
12
9
6
3


Propensity Score (Random Forest)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Propensity Score (Logistic Regression)
age
75
50
25
0


Propensity Score (Random Forest)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Propensity Score (Logistic Regression)
sexe
1.00
0.75
0.50
0.25
0.00

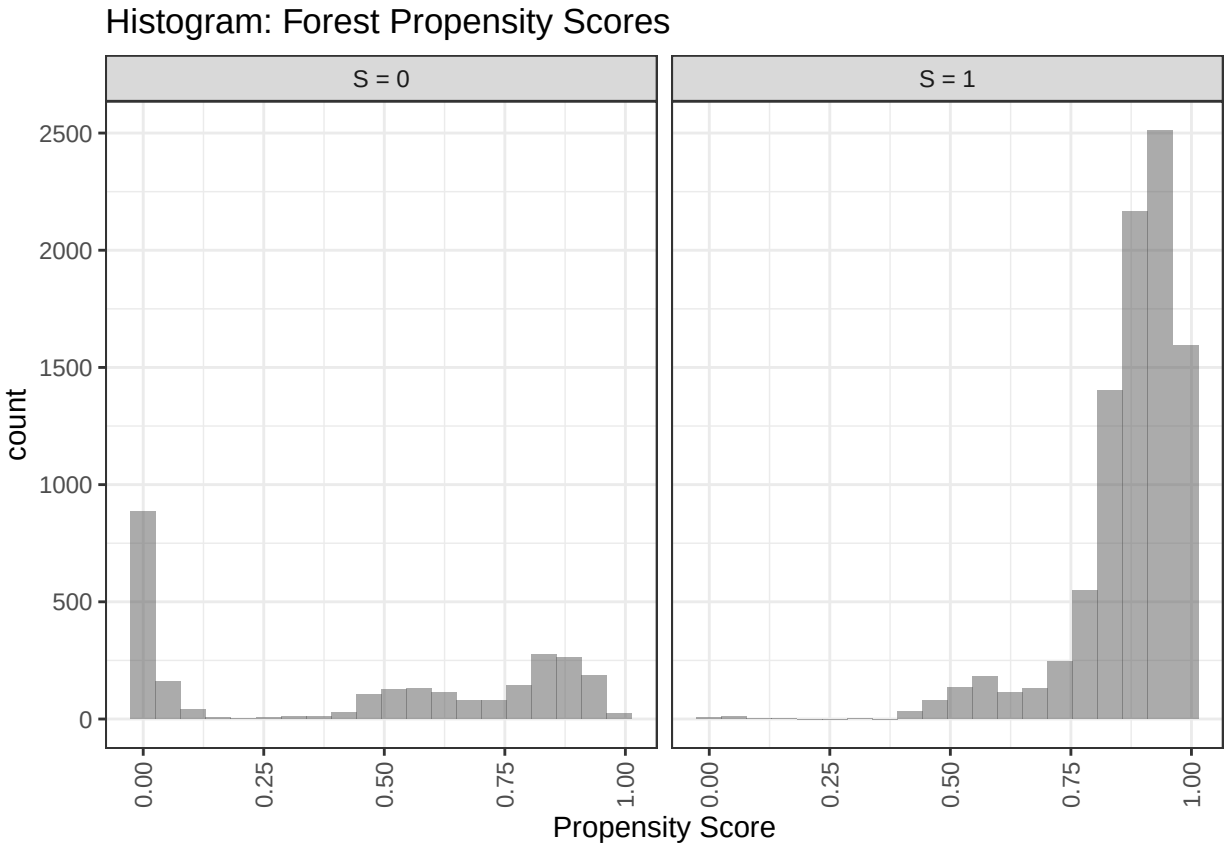

Propensity Score (Random Forest)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Propensity Score (Logistic Regression)
pupil
2
1
0
−1

This time we observe that the analysis are more different between the two methods. This time also, the random forest badly classify the observations with extracranial bleeding. Because the scores are very different with the imputed data we propose not to impute them.

To further understand the effect of each variable, we perform an analysis without the observations that have an extracranial bleeding.

```
DF_without_extracranial <- DF[DF$majorExtracranial ==
    0, c("age", "Glasgow.initial",
    "systolicBloodPressure", "sexe",
    "pupilReact_num", "V")]

# @Imke: it fails on glm... :(
# )
# pi_s_hat_glm_without_extracranial
# <-
# sampling_propensities(DF_without_extracranial,
# method = 'glm', seed = 100)
pi_s_hat_grf_without_extracranial <- sampling_propensities(DF_without_extracranial,
    method = "grf", seed = 100)
```

```
DF_without_extracranial$majorExtracranial <- rep(0,
    nrow(DF_without_extracranial))
propensity_scores <- data.frame(grf = pi_s_hat_grf_without_extracranial,
    RCT = paste0("S = ", as.factor(DF_without_extracranial$V)),
    extraCranialBleeding = as.factor(DF_without_extracranial$majorExtracranial),
    SystolicBloodPressure = DF_without_extracranial$systolicBloodPressure,
    GCS = DF_without_extracranial$Glasgow.initial,
    pupil = DF_without_extracranial$pupilReact_num,
    sexe = DF_without_extracranial$sexe,
    age = DF_without_extracranial$age)

g <- ggplot(propensity_scores,
    aes(x = grf)) + geom_histogram(bins = 20,
    alpha = 0.5) + facet_wrap(~RCT) +
    theme_bw() + theme(axis.text.x = element_text(angle = 90,
    vjust = 0.5)) + labs(title = "Histogram: Forest Propensity Scores",
    x = "Propensity Score")

print(g)
```



General conclusion

For now, we will keep both approaches, but we need to investigate in more detail methodological and theoretical issues to formalize assumptions for combining generalization estimators with missing values handling approaches.