DSC 385 - Project 4 Report

Name and EID

Name: Marina Witherell
EID: mw44788

Background

Right heart catheterization (RHC) is a diagnostic procedure for directly measuring cardiac function in critically ill patients. In an influential study, Connors et al. (1996) studied the effectiveness of RHC with an observational study design. The study collected data on 5735 hospitalized adult patients; 2184 of them are assigned to the experimental treatment, receipt of RHC within 24 hours of admission, and the remaining 3551 assigned to the control condition. The outcome was death at 30 days after admission to the hospital. The goal is to assess the causal effect of RHC (the treatment) on the binary outcome, death at 30 days after admission.

The dataset provided here is a cleaned version of the original dataset. The treatment variable in the dataset is `swang1` and the outcome variable is `death`. To simplify the analysis, we have restricted the data to 20 covariates that have been identified as the top confounders in an ad-hoc exploratory analysis.

Below is a table describing each of the variables in the dataset:

Variable Name	**Interpretation**
age	Age in years
sex	Male / Female
cat1	Primary Diagnosis: COPD, Multiple Organ System Failure (MOSF) w/Sepsis, MOSF w/Malignancy, CHF, Coma, Cirrhosis, Lung Cancer, Colon Cancer
cat2	Secondary Diagnosis: MOSF w/Sepsis, Coma, MOSF w/Malignancy, Lung Cancer, Cirrhosis, Colon Cancer
ca	yes = Cancer localized, no/Metastatic = metastatic
death	Death within 30 days after hospital admission
pafi1	PaO2 / F102 ratio
wtkilo1	Weight
surv2md1	Estimate of probability of surviving 2 months

Variable Name	Interpretation
dementhx	Dementia, stroke or cerebral infarct, Parkinsons's disease
gastr	Gastrointestinal diagnosis
wblc1	White blood count
temp1	Temperature
das2d3pc	DASI-Duke Activity Status Index
chfhx	Congestive heart failure
hema	Hematological diagnosis
chrpulhx	Chronic pulmonary disease, severe pulmonary disease
cardiohx	Cardiovascular symptoms
meta	Metabolic diagnosis

Setup

```
library(tidyverse)
library(MatchIt)
```

```
Warning: package 'MatchIt' was built under R version 4.5.3
```

```
## Load any additional packages that you need here
```

The following code loads the dataset, removes some unnecessary columns, and converts categorical variables to factors.

```
## Read in the data and remove unnecessary columns
dataset <- read_csv("rhc-demo.csv",
                    col_select = c(-1, -ptid, -survtime, -urin1, -adld3p)) |>
    mutate(across(where(is.character), factor)) |> ## Convert to factors
    mutate(across(death:chrpulhx, factor))
```

```
New names:
Rows: 5735 Columns: 20
── Column specification
──────────────────────────────────────────────────────── Delimiter: "," chr
(8): cat1, cat2, ca, sex, swang1, gastr, meta, hema dbl (12): death, cardiohx,
```

```
chfhx, dementhx, chrpulhx, age, surv2md1, das2d3p...
ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
Specify the column types or set `show_col_types = FALSE` to quiet this message.
• `` -> `...1`
```

We can take a quick look at the dataset and all of the variables.

```
dataset
```

```
# A tibble: 5,735 × 20
   cat1  cat2  ca    death cardiohx chfhx dementhx chrpulhx   age sex   surv2md1
   <fct> <fct> <fct> <fct> <fct>    <fct> <fct>    <fct>    <dbl> <fct>    <dbl>
 1 COPD  <NA>  Yes   0     0        0     0        1         70.3 Male     0.641
 2 MOSF  <NA>  No    1     1        1     0        0         78.2 Fema…    0.755
 3 MOSF  MOSF… Yes   0     0        0     0        0         46.1 Fema…    0.317
 4 ARF   <NA>  No    1     0        0     0        0         75.3 Fema…    0.441
 5 MOSF  <NA>  No    1     0        0     0        0         67.9 Male     0.437
 6 COPD  <NA>  No    0     0        1     0        1         86.1 Fema…    0.665
 7 MOSF  <NA>  Meta… 0     0        0     0        0         55.0 Male     0.339
 8 ARF   Coma  No    1     0        0     0        0         43.6 Male     0.632
 9 MOSF  <NA>  Yes   0     0        0     0        0         18.0 Fema…    0.503
10 ARF   <NA>  Yes   0     0        0     0        0         48.4 Fema…    0.669
# ℹ 5,725 more rows
# ℹ 9 more variables: das2d3pc <dbl>, wblc1 <dbl>, temp1 <dbl>, pafi1 <dbl>,
#   swang1 <fct>, wtkilo1 <dbl>, gastr <fct>, meta <fct>, hema <fct>
```

```
dataset |>
    glimpse()
```

```
Rows: 5,735
Columns: 20
$ cat1     <fct> COPD, MOSF, MOSF, ARF, MOSF, COPD, MOSF, ARF, MOSF, ARF, ARF,…
$ cat2     <fct> NA, NA, MOSF w/Sepsis, NA, NA, NA, NA, Coma, NA, NA, NA, NA, …
$ ca       <fct> Yes, No, Yes, No, No, No, Metastatic, No, Yes, Yes, No, No, N…
$ death    <fct> 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 1, 1, 0, 0, 0…
$ cardiohx <fct> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0…
$ chfhx    <fct> 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0…
$ dementhx <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0…
$ chrpulhx <fct> 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0…
$ age      <dbl> 70.25098, 78.17896, 46.09198, 75.33197, 67.90997, 86.07794, 5…
$ sex      <fct> Male, Female, Female, Female, Male, Female, Male, Male, Femal…
$ surv2md1 <dbl> 0.6409912, 0.7549996, 0.3169999, 0.4409790, 0.4369998, 0.6650…
$ das2d3pc <dbl> 23.50000, 14.75195, 18.13672, 22.92969, 21.05078, 17.50000, 1…
$ wblc1    <dbl> 22.09765620, 28.89843750, 0.04999542, 23.29687500, 29.6992188…
$ temp1    <dbl> 38.69531, 38.89844, 36.39844, 35.79688, 34.79688, 39.19531, 3…
$ pafi1    <dbl> 68.00000, 218.31250, 275.50000, 156.65625, 478.00000, 184.187…
$ swang1   <fct> No RHC, RHC, RHC, No RHC, RHC, No RHC, No RHC, No RHC, No RHC…
$ wtkilo1  <dbl> 64.69995, 45.69998, 0.00000, 54.59998, 78.39996, 54.89999, 0.…
$ gastr    <fct> No, No, No, No, No, No, No, No, No, Yes, No, No, No, No, No, …
```

```
$ meta     <fct> No, No, No, No, No, No, No, No, No, No, No, No, No, No, No, N…
$ hema     <fct> No, No, No, No, No, No, No, No, Yes, No, No, No, No, No, No, …
```

Part 1

Let's compare the variables in the dataset between the treatment and control groups. The `swang1` variable indicates whether a person got right heart catheterization (`swang1 == RHC`) or did not (`swang1 == No RHC`).

First take a look at the **numeric** variables in the dataset and compare their means between treatment groups.

Do the two groups look similar to you based on the means of the numeric variables?

```
dataset |>
  group_by(swang1) |>
  select(where(is.numeric)) |>
  aggregate(.~swang1, FUN=mean)
```

```
Adding missing grouping variables: `swang1`
```

	swang1	age	surv2md1	das2d3pc	wblc1	temp1	pafi1	wtkilo1
1	No RHC	61.76093	0.6072064	20.37148	15.26349	37.63288	240.6266	65.04024
2	RHC	60.74984	0.5684576	20.70079	16.26573	37.59473	192.4334	72.36018

Answer

The two groups do seem similar based on the means of the numeric variables. The means of the numeric variables between the groups do not appear to be significantly different, but it would need further testing to definitively make that conclusion.

Part 2

Now let's take a look at some of the other variables. For each of the **factor** variables in the dataset and for each treatment group, compute the proportion of people in the dataset that are equal to the first factor level of the variable.

NOTE: The `levels()` function can be used to identify the levels of a factor. For a variables `x`, the first factor level is `levels(x)[1]`.

Which of the factor variables appears to have the largest absolute difference in proportion between the treatment groups?

```
dataset |>
  group_by(swang1) |>
  summarise(across(where(is.factor),
                   ~mean(. == levels(.)[1], na.rm = TRUE))) |>
  pivot_longer(-swang1, names_to = "variable", values_to = "proportion") %>%
```

```
  pivot_wider(names_from = swang1, values_from = proportion) %>%
  mutate(abs_diff = abs(RHC - `No RHC`)) %>%
  arrange(desc(abs_diff))
```

```
# A tibble: 12 × 4
   variable `No RHC`    RHC abs_diff
   <chr>       <dbl>  <dbl>    <dbl>
 1 chrpulhx   0.782  0.856   0.0737
 2 death      0.370  0.320   0.0507
 3 dementhx   0.884  0.931   0.0472
 4 sex        0.461  0.415   0.0462
 5 gastr      0.853  0.808   0.0453
 6 cardiohx   0.840  0.796   0.0445
 7 cat1       0.445  0.416   0.0290
 8 chfhx      0.832  0.805   0.0268
 9 cat2       0.0392 0.0215  0.0178
10 ca         0.0735 0.0563  0.0172
11 hema       0.933  0.947   0.0146
12 meta       0.952  0.957   0.00585
```

Answer

The factor variables that appear to have the largest absolute difference in proportion between the treatment groups are `death`, `dementhx`, and `chrpulhx`. The variable with the largest difference is `chrpulhx` with an absolute difference of 0.0737 between the two groups.

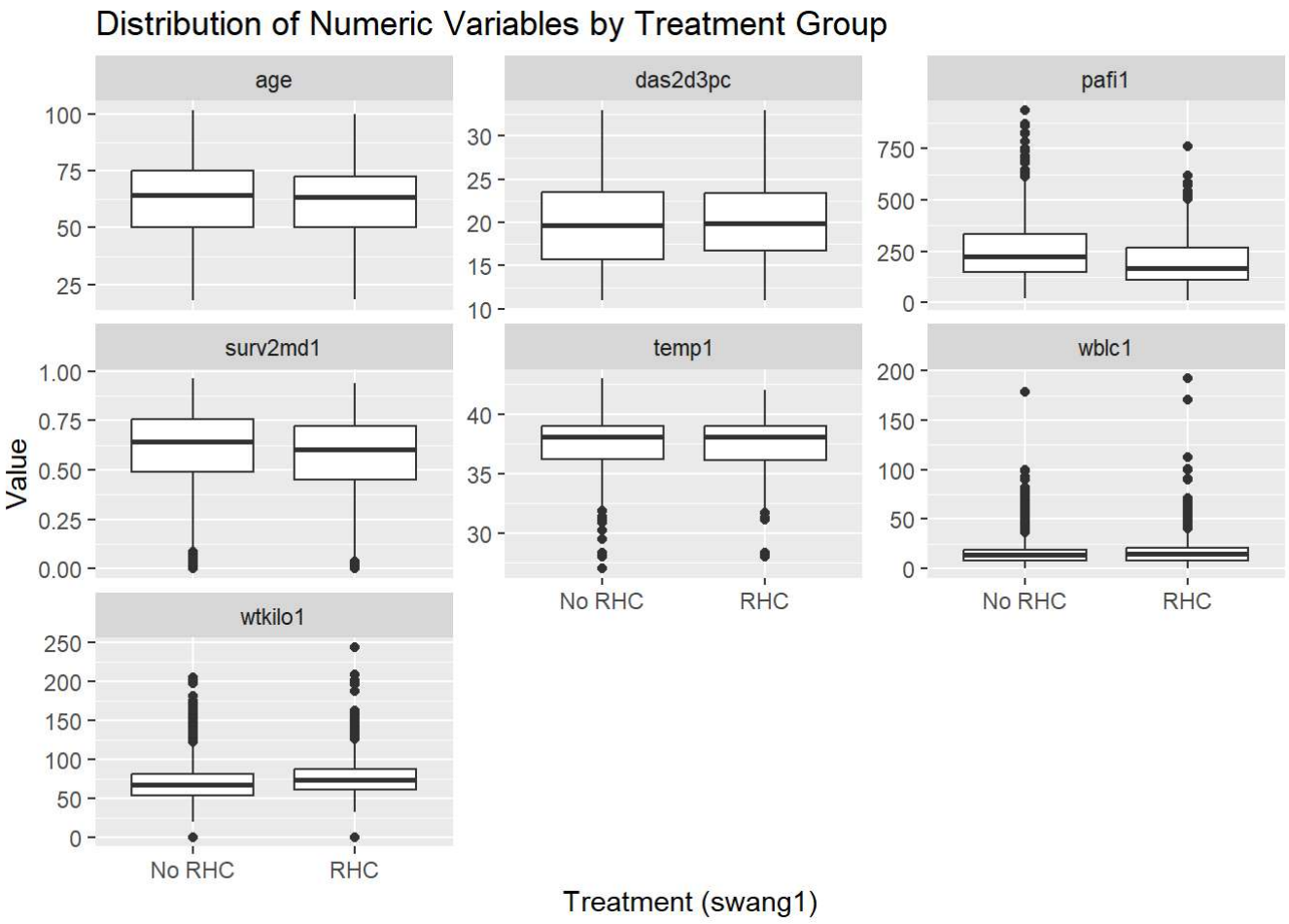
Part 3

Make a panel of boxplot(s) comparing the numeric variables across treatment groups.

HINT: If you use the `facet_wrap()` function, you should set `scales = "free_y"` to allow for different y-axis scales for the different ranges of the variables.

```
dataset |>
  group_by(swang1) |>
  select(where(is.numeric)) |>
  pivot_longer(cols = -swang1) |>
  ggplot(aes(x=factor(swang1), y=value)) +
  geom_boxplot() +
  facet_wrap(~name, scales = "free_y") +
  labs(title = "Distribution of Numeric Variables by Treatment Group",
       x = "Treatment (swang1)", y = "Value")
```

```
Adding missing grouping variables: `swang1`
```



From the boxplots, which numeric variable, appears to have the largest difference between the treatment and control groups?

Answer

From the boxplots, the numeric variable that appears to have the largest difference between groups is the `pafi1` variable.

Part 4

Using the `MatchIt` package, fit a propensity score model using all the predictors **except** the `cat2` variable (which has too many `NA` values) and the `death` variable. The propensity score model should specify a matching ratio of 1:1.

```
df_clean <- dataset |>
  drop_na(cat2)

pfit <- matchit(swang1 ~ . - cat2 - death, data=df_clean, ratio=1)
```

Make a plot of the absolute standardized mean difference in each predictor variable for both before and after matching is done via the propensity score model.

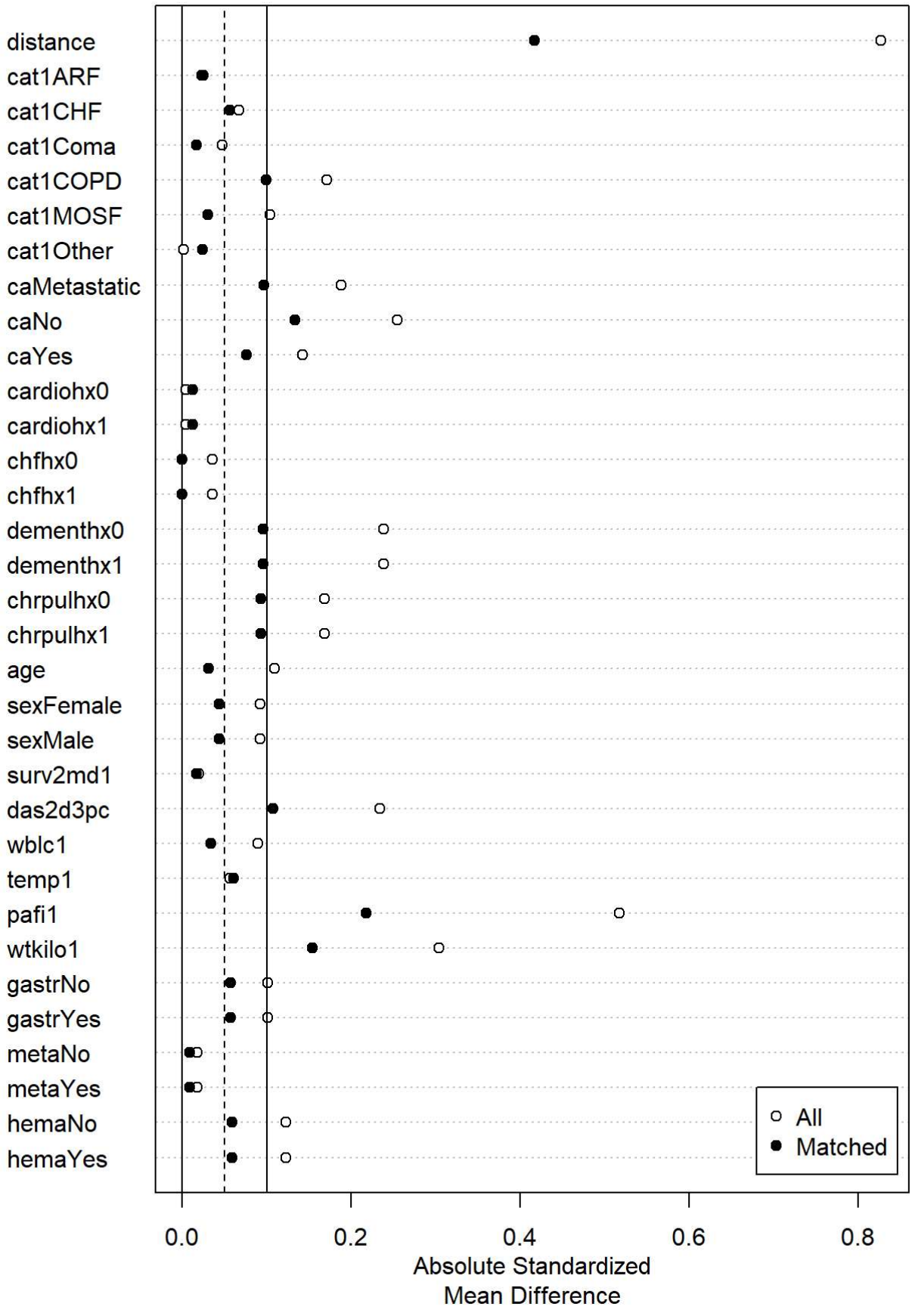
NOTE: You can adjust the height of the figure in the HTML output document by setting `fig-height` to be a larger number.

```
plot(summary(pfit), stat="abs.std.diff")
```

```
Warning in mtext(labels[o], side = 2, line = loffset, at = y, adj = 0, col =
color, : "stat" is not a graphical parameter

Warning in title(main = main, xlab = xlab, ...): "stat" is not a graphical
parameter

Warning in title(ylab = ylab, ...): "stat" is not a graphical parameter
```


distance
cat1ARF
cat1CHF
cat1Coma
cat1COPD
cat1MOSF
cat1Other
caMetastatic
caNo
caYes
cardiohx0
cardiohx1
chfhx0
chfhx1
dementhx0
dementhx1
chrpulhx0
chrpulhx1
age
sexFemale
sexMale
surv2md1
das2d3pc
wblc1
temp1
pafi1
wtkilo1
gastrNo
gastrYes
metaNo
metaYes
hemaNo
hemaYes
All
Matched
0.0
0.2
0.4
0.6
0.8
Absolute Standardized
Mean Difference

How would you describe the balance of the covariates between the treatment and control groups in the matched dataset?

Answer

The covariates between the treatment and control groups in the matched dataset are mostly well-balanced, since most of the absolute standardized mean differences are less than 0.1. Some variables remain slightly imbalanced, but on the whole, the balance improves with match. This can also be seen by the `distance` being smaller when matched vs unmatched.

Part 5

Create a 1-to-1 matched dataset using the propensity score model that you fit in the previous part. Call the new matched dataset `dataset_match`.

```
dataset_match <- match.data(pfit)
```

Run a simple logistic regression using the `dataset_match` object with the outcome `death` (death after 30 days from hospital admission) and the treatment variable `swang1`.

What is the effect of right heart catheterization on 30-day mortality in these participants?

```
fit_matched <- glm(death ~ swang1,
                   data = dataset_match,
                   family = binomial(link = "logit"))
summary(fit_matched)
```

```
Call:
glm(formula = death ~ swang1, family = binomial(link = "logit"),
    data = dataset_match)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.1838     0.1043  11.347   <2e-16 ***
swang1RHC    -0.1059     0.1456  -0.727    0.467
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1138.3  on 1023  degrees of freedom
Residual deviance: 1137.8  on 1022  degrees of freedom
AIC: 1141.8

Number of Fisher Scoring iterations: 4
```

Answer

Based on the simple logistic regression, right heart catheterization decreased 30-day mortality in these participants. However, since the p-value is high (0.467), we fail to reject the null hypothesis that right heart catheterization has no effect on 30-day mortality.

Part 6

Given what you saw in the propensity score analysis above, would you trust the result of this analysis? What potential confounding variables may explain the relationship between RHC and 30-day mortality?

Answer

Although the propensity score analysis showed balance of the covariates, the result of the logistic regression analysis was not statistically significant. Some potential confounding variables that may explain the relationship between RHC and 30-day mortality are the socioeconomic status of the patient, how effective their hospital is (better equipment, doctors, etc.), the other diagnostic procedures done on the patient, and many more.

A common issue with studies such as this one is that it is almost impossible to get a result that accounts for all confounding variables that is applicable to the population, since all factors of a large sample size in a large number of locations can not be controlled in practice.