Balance and variance inflation checks for completeness–propensity weights

Roger B. Newson
r.newson@qmul.ac.uk
http://www.rogernewsonresources.org.uk

Cancer Prevention Group, Wolfson Institute of Population Health, Queen Mary University London

Presented at the 2024 UK Stata Conference, London,
12–13 September, 2024
Downloadable from the conference website at
https://econpapers.repec.org/paper/boclsug24/

The Rubin causal model using propensity weights (revisited)

- The **Rubin causal model**[1] is a 2–stage process for estimating treatment effects, adjusting for confounders.
- In Stage 1 (“Design”), we find a **propensity model** in the data on treatment and confounders, predicting treatment from confounders.
- This model is used to compute **inverse treatment–propensity weights**, which can be used to directly standardize the sample to a fantasy target population, with a real–world distribution of confounders, in which treatment is independent of confounders.
- In Stage 2 (“Analysis”), we bring in the outcome data, and estimate the mean treated–control difference in that fantasy target population, using the inverse treatment–propensity weights to standardize.

The Rubin causal model using propensity weights (revisited)

- The **Rubin causal model**[1] is a 2–stage process for estimating treatment effects, adjusting for confounders.
- In Stage 1 ("Design"), we find a **propensity model** in the data on treatment and confounders, predicting treatment from confounders.
- This model is used to compute **inverse treatment–propensity weights**, which can be used to directly standardize the sample to a fantasy target population, with a real–world distribution of confounders, in which treatment is independent of confounders.
- In Stage 2 ("Analysis"), we bring in the outcome data, and estimate the mean treated–control difference in that fantasy target population, using the inverse treatment–propensity weights to standardize.

The Rubin causal model using propensity weights (revisited)

- The **Rubin causal model**[1] is a 2–stage process for estimating treatment effects, adjusting for confounders.
- In Stage 1 ("Design"), we find a **propensity model** in the data on treatment and confounders, predicting treatment from confounders.
- This model is used to compute **inverse treatment–propensity weights**, which can be used to directly standardize the sample to a fantasy target population, with a real–world distribution of confounders, in which treatment is independent of confounders.
- In Stage 2 ("Analysis"), we bring in the outcome data, and estimate the mean treated–control difference in that fantasy target population, using the inverse treatment–propensity weights to standardize.

The Rubin causal model using propensity weights (revisited)

- The **Rubin causal model**[1] is a 2–stage process for estimating treatment effects, adjusting for confounders.
- In Stage 1 ("Design"), we find a **propensity model** in the data on treatment and confounders, predicting treatment from confounders.
- This model is used to compute **inverse treatment–propensity weights**, which can be used to directly standardize the sample to a fantasy target population, with a real–world distribution of confounders, in which treatment is independent of confounders.
- In Stage 2 ("Analysis"), we bring in the outcome data, and estimate the mean treated–control difference in that fantasy target population, using the inverse treatment–propensity weights to standardize.

The Rubin causal model using propensity weights (revisited)

- The **Rubin causal model**[1] is a 2–stage process for estimating treatment effects, adjusting for confounders.
- In Stage 1 (“Design”), we find a **propensity model** in the data on treatment and confounders, predicting treatment from confounders.
- This model is used to compute **inverse treatment–propensity weights**, which can be used to directly standardize the sample to a fantasy target population, with a real–world distribution of confounders, in which treatment is independent of confounders.
- In Stage 2 (“Analysis”), we bring in the outcome data, and estimate the mean treated–control difference in that fantasy target population, using the inverse treatment–propensity weights to standardize.

Example: Maternal smoking and birth weight in the `cattaneo2` data

- An example from the `cattaneo2` data appears in Newson and Falcaro (2023)[2], and in the example do–file for this presentation.
- Observations are 4642 pregnancies.
- The outcome is birthweight in grams. The "treatment" (or **exposure**) `mbsmoke` is self–reported maternal smoking,
- And there are 17 confounding covariates (mostly health– or wealth–related), entered into a logit propensity model to predict maternal smoking and to derive **average treatment effect (ATE) weights**.
- We checked these weights for balance and variance inflation, using the SSC packages `somersd`[3] and `haif`, respectively.
- In the analysis phase, we used the ATE weights in a regression model to estimate mean smoking effect on birthweight (which was negative).

Example: Maternal smoking and birth weight in the `cattaneo2` data

- An example from the `cattaneo2` data appears in Newson and Falcaro (2023)[2], and in the example do–file for this presentation.
- Observations are 4642 pregnancies.
- The outcome is birthweight in grams. The “treatment” (or **exposure**) `mbsmoke` is self–reported maternal smoking,
- And there are 17 confounding covariates (mostly health– or wealth–related), entered into a logit propensity model to predict maternal smoking and to derive **average treatment effect (ATE) weights**.
- We checked these weights for balance and variance inflation, using the SSC packages `somersd`[3] and `haif`, respectively.
- In the analysis phase, we used the ATE weights in a regression model to estimate mean smoking effect on birthweight (which was negative).

Example: Maternal smoking and birth weight in the `cattaneo2` data

- An example from the `cattaneo2` data appears in Newson and Falcaro (2023)[2], and in the example do–file for this presentation.
- Observations are 4642 pregnancies.
- The outcome is birthweight in grams. The “treatment” (or **exposure**) `mbsmoke` is self–reported maternal smoking,
- And there are 17 confounding covariates (mostly health– or wealth–related), entered into a logit propensity model to predict maternal smoking and to derive **average treatment effect (ATE) weights**.
- We checked these weights for balance and variance inflation, using the SSC packages `somersd`[3] and `haif`, respectively.
- In the analysis phase, we used the ATE weights in a regression model to estimate mean smoking effect on birthweight (which was negative).

Example: Maternal smoking and birth weight in the `cattaneo2` data

- An example from the `cattaneo2` data appears in Newson and Falcaro (2023)[2], and in the example do–file for this presentation.
- Observations are 4642 pregnancies.
- The outcome is birthweight in grams. The “treatment” (or **exposure**) `mbsmoke` is self–reported maternal smoking,
- And there are 17 confounding covariates (mostly health– or wealth–related), entered into a logit propensity model to predict maternal smoking and to derive **average treatment effect (ATE) weights**.
- We checked these weights for balance and variance inflation, using the SSC packages `somersd`[3] and `haif`, respectively.
- In the analysis phase, we used the ATE weights in a regression model to estimate mean smoking effect on birthweight (which was negative).

Example: Maternal smoking and birth weight in the `cattaneo2` data

- An example from the `cattaneo2` data appears in Newson and Falcaro (2023)[2], and in the example do–file for this presentation.
- Observations are 4642 pregnancies.
- The outcome is birthweight in grams. The “treatment” (or **exposure**) `mbsmoke` is self–reported maternal smoking,
- And there are 17 confounding covariates (mostly health– or wealth–related), entered into a logit propensity model to predict maternal smoking and to derive **average treatment effect (ATE) weights**.
- We checked these weights for balance and variance inflation, using the SSC packages `somersd`[3] and `haif`, respectively.
- In the analysis phase, we used the ATE weights in a regression model to estimate mean smoking effect on birthweight (which was negative).

Example: Maternal smoking and birth weight in the `cattaneo2` data

- An example from the `cattaneo2` data appears in Newson and Falcaro (2023)[2], and in the example do–file for this presentation.
- Observations are 4642 pregnancies.
- The outcome is birthweight in grams. The “treatment” (or **exposure**) `mbsmoke` is self–reported maternal smoking,
- And there are 17 confounding covariates (mostly health– or wealth–related), entered into a logit propensity model to predict maternal smoking and to derive **average treatment effect (ATE) weights**.
- We checked these weights for balance and variance inflation, using the SSC packages `somersd`[3] and `haif`, respectively.
- In the analysis phase, we used the ATE weights in a regression model to estimate mean smoking effect on birthweight (which was negative).

Example: Maternal smoking and birth weight in the `cattaneo2` data

- An example from the `cattaneo2` data appears in Newson and Falcaro (2023)[2], and in the example do–file for this presentation.
- Observations are 4642 pregnancies.
- The outcome is birthweight in grams. The “treatment” (or **exposure**) `mbsmoke` is self–reported maternal smoking,
- And there are 17 confounding covariates (mostly health– or wealth–related), entered into a logit propensity model to predict maternal smoking and to derive **average treatment effect (ATE) weights**.
- We checked these weights for balance and variance inflation, using the SSC packages `somersd`[3] and `haif`, respectively.
- In the analysis phase, we used the ATE weights in a regression model to estimate mean smoking effect on birthweight (which was negative).

Inverse completeness–propensity weights

- Inverse completeness–propensity weights are sometimes known simply as inverse–probability weights[4].
- They are sometimes used to correct for the presence of incomplete observations, with missing values for one or more important variables.
- *For instance*, in a randomized controlled trial, some randomized subjects may have missing values for the primary outcome designated in the protocol.
- A possible remedy might be to weight the surviving subjects inversely proportionally to their predicted probability of completeness, or **completeness–propensity**, given a list of baseline variables that are always complete.
- This procedure might (we hope) directly standardize the results observed in the complete subjects to the total set of randomized subjects.

Inverse completeness–propensity weights

- Inverse completeness–propensity weights are sometimes known simply as inverse–probability weights[4].
- They are sometimes used to correct for the presence of incomplete observations, with missing values for one or more important variables.
- *For instance*, in a randomized controlled trial, some randomized subjects may have missing values for the primary outcome designated in the protocol.
- A possible remedy might be to weight the surviving subjects inversely proportionally to their predicted probability of completeness, or **completeness–propensity**, given a list of baseline variables that are always complete.
- This procedure might (we hope) directly standardize the results observed in the complete subjects to the total set of randomized subjects.

Inverse completeness–propensity weights

- Inverse completeness–propensity weights are sometimes known simply as inverse–probability weights[4].
- They are sometimes used to correct for the presence of incomplete observations, with missing values for one or more important variables.
- *For instance*, in a randomized controlled trial, some randomized subjects may have missing values for the primary outcome designated in the protocol.
- A possible remedy might be to weight the surviving subjects inversely proportionally to their predicted probability of completeness, or **completeness–propensity**, given a list of baseline variables that are always complete.
- This procedure might (we hope) directly standardize the results observed in the complete subjects to the total set of randomized subjects.

Inverse completeness–propensity weights

- Inverse completeness–propensity weights are sometimes known simply as inverse–probability weights[4].
- They are sometimes used to correct for the presence of incomplete observations, with missing values for one or more important variables.
- *For instance*, in a randomized controlled trial, some randomized subjects may have missing values for the primary outcome designated in the protocol.
- A possible remedy might be to weight the surviving subjects inversely proportionally to their predicted probability of completeness, or **completeness–propensity**, given a list of baseline variables that are always complete.
- This procedure might (we hope) directly standardize the results observed in the complete subjects to the total set of randomized subjects.

Inverse completeness–propensity weights

- Inverse completeness–propensity weights are sometimes known simply as inverse–probability weights[4].
- They are sometimes used to correct for the presence of incomplete observations, with missing values for one or more important variables.
- *For instance*, in a randomized controlled trial, some randomized subjects may have missing values for the primary outcome designated in the protocol.
- A possible remedy might be to weight the surviving subjects inversely proportionally to their predicted probability of completeness, or **completeness–propensity**, given a list of baseline variables that are always complete.
- This procedure might (we hope) directly standardize the results observed in the complete subjects to the total set of randomized subjects.

Inverse completeness–propensity weights

- Inverse completeness–propensity weights are sometimes known simply as inverse–probability weights[4].
- They are sometimes used to correct for the presence of incomplete observations, with missing values for one or more important variables.
- *For instance*, in a randomized controlled trial, some randomized subjects may have missing values for the primary outcome designated in the protocol.
- A possible remedy might be to weight the surviving subjects inversely proportionally to their predicted probability of completeness, or **completeness–propensity**, given a list of baseline variables that are always complete.
- This procedure might (we hope) directly standardize the results observed in the complete subjects to the total set of randomized subjects.

Example: the Positional Therapy for Positional Obstructive Sleep Apnoea (POSA) trial

- This multi–center trial[5] was organized jointly between the Royal Brompton Hospital, Imperial College London, and Oxford Respiratory Trials Unit.
- It was intended to test the usefulness, in patients with a sleep apnoea problem, of a proprietary device (Night Shift™), strapped around the neck at night, which alerts patients by vibrational feedback if they attempt to sleep in a supine position.
- The primary outcome was an **apnoea–hyperpnoea index (AHI)**, expressed as a mean number of apnoeic breathing (snoring) events per hour, and measured at baseline and after 3 months.
- Patients were randomized to a working device (with vibrational feedback, 59 subjects) or a sham device (set to monitoring only, 61 subjects).
- *Unfortunately*, following an unforeseen pandemic, the primary outcome was only available at baseline and 3 months for 45 intervention and 47 control subjects!

Example: the Positional Therapy for Positional Obstructive Sleep Apnoea (POSA) trial

- This multi–center trial[5] was organized jointly between the Royal Brompton Hospital, Imperial College London, and Oxford Respiratory Trials Unit.
- It was intended to test the usefulness, in patients with a sleep apnoea problem, of a proprietary device (Night Shift™), strapped around the neck at night, which alerts patients by vibrational feedback if they attempt to sleep in a supine position.
- The primary outcome was an **apnoea–hyperpnoea index (AHI)**, expressed as a mean number of apnoeic breathing (snoring) events per hour, and measured at baseline and after 3 months.
- Patients were randomized to a working device (with vibrational feedback, 59 subjects) or a sham device (set to monitoring only, 61 subjects).
- *Unfortunately*, following an unforeseen pandemic, the primary outcome was only available at baseline and 3 months for 45 intervention and 47 control subjects!

Example: the Positional Therapy for Positional Obstructive Sleep Apnoea (POSA) trial

- This multi–center trial[5] was organized jointly between the Royal Brompton Hospital, Imperial College London, and Oxford Respiratory Trials Unit.
- It was intended to test the usefulness, in patients with a sleep apnoea problem, of a proprietary device (Night Shift™), strapped around the neck at night, which alerts patients by vibrational feedback if they attempt to sleep in a supine position.
- The primary outcome was an **apnoea–hyperpnoea index (AHI)**, expressed as a mean number of apnoeic breathing (snoring) events per hour, and measured at baseline and after 3 months.
- Patients were randomized to a working device (with vibrational feedback, 59 subjects) or a sham device (set to monitoring only, 61 subjects).
- *Unfortunately*, following an unforeseen pandemic, the primary outcome was only available at baseline and 3 months for 45 intervention and 47 control subjects!

Example: the Positional Therapy for Positional Obstructive Sleep Apnoea (POSA) trial

- This multi–center trial[5] was organized jointly between the Royal Brompton Hospital, Imperial College London, and Oxford Respiratory Trials Unit.
- It was intended to test the usefulness, in patients with a sleep apnoea problem, of a proprietary device (Night Shift™), strapped around the neck at night, which alerts patients by vibrational feedback if they attempt to sleep in a supine position.
- The primary outcome was an **apnoea–hyperpnoea index (AHI)**, expressed as a mean number of apnoeic breathing (snoring) events per hour, and measured at baseline and after 3 months.
- Patients were randomized to a working device (with vibrational feedback, 59 subjects) or a sham device (set to monitoring only, 61 subjects).
- *Unfortunately*, following an unforeseen pandemic, the primary outcome was only available at baseline and 3 months for 45 intervention and 47 control subjects!

Example: the Positional Therapy for Positional Obstructive Sleep Apnoea (POSA) trial

- This multi–center trial[5] was organized jointly between the Royal Brompton Hospital, Imperial College London, and Oxford Respiratory Trials Unit.
- It was intended to test the usefulness, in patients with a sleep apnoea problem, of a proprietary device (Night Shift™), strapped around the neck at night, which alerts patients by vibrational feedback if they attempt to sleep in a supine position.
- The primary outcome was an **apnoea–hyperpnoea index (AHI)**, expressed as a mean number of apnoeic breathing (snoring) events per hour, and measured at baseline and after 3 months.
- Patients were randomized to a working device (with vibrational feedback, 59 subjects) or a sham device (set to monitoring only, 61 subjects).
- *Unfortunately*, following an unforeseen pandemic, the primary outcome was only available at baseline and 3 months for 45 intervention and 47 control subjects!

Example: the Positional Therapy for Positional Obstructive Sleep Apnoea (POSA) trial

- This multi–center trial[5] was organized jointly between the Royal Brompton Hospital, Imperial College London, and Oxford Respiratory Trials Unit.
- It was intended to test the usefulness, in patients with a sleep apnoea problem, of a proprietary device (Night Shift™), strapped around the neck at night, which alerts patients by vibrational feedback if they attempt to sleep in a supine position.
- The primary outcome was an **apnoea–hyperpnoea index (AHI)**, expressed as a mean number of apnoeic breathing (snoring) events per hour, and measured at baseline and after 3 months.
- Patients were randomized to a working device (with vibrational feedback, 59 subjects) or a sham device (set to monitoring only, 61 subjects).
- *Unfortunately*, following an unforeseen pandemic, the primary outcome was only available at baseline and 3 months for 45 intervention and 47 control subjects!

The model for the primary estimand

This was a multi–factorial model, regressing 3–month AHI with respect to baseline AHI and randomization age group. The analysis was "treble–blind", so treatment groups were labelled "Group 1" (actually working device) and "Group 2" (actually sham device). The parameters were:

- A Group 2 effect (expressed in AHI units of events per hour). compared to a reference level of Group 1. This was the primary estimand.
- A randomization age group effect for 65+ years (in AHI units), compared to a reference level of <65 years.
- A linear effect of baseline AHI (in units per unit).
- A constant term (in AHI units), representing the mean AHI for a Group 1 subject aged <65 years with a mean baseline AHI of 15.245.

So, the primary estimand was a difference between 3–month AHI event rate in sham–device patients and 3–month AHI event rate in working–device patients, other things being equal.

The model for the primary estimand

This was a multi–factorial model, regressing 3–month AHI with respect to baseline AHI and randomization age group. The analysis was "treble–blind", so treatment groups were labelled "Group 1" (actually working device) and "Group 2" (actually sham device). The parameters were:

- A Group 2 effect (expressed in AHI units of events per hour). compared to a reference level of Group 1. This was the primary estimand.
- A randomization age group effect for 65+ years (in AHI units), compared to a reference level of <65 years.
- A linear effect of baseline AHI (in units per unit).
- A constant term (in AHI units), representing the mean AHI for a Group 1 subject aged <65 years with a mean baseline AHI of 15.245.

So, the primary estimand was a difference between 3–month AHI event rate in sham–device patients and 3–month AHI event rate in working–device patients, other things being equal.

The model for the primary estimand

This was a multi–factorial model, regressing 3–month AHI with respect to baseline AHI and randomization age group. The analysis was "treble–blind", so treatment groups were labelled "Group 1" (actually working device) and "Group 2" (actually sham device). The parameters were:

- A Group 2 effect (expressed in AHI units of events per hour). compared to a reference level of Group 1. This was the primary estimand.
- A randomization age group effect for 65+ years (in AHI units), compared to a reference level of <65 years.
- A linear effect of baseline AHI (in units per unit).
- A constant term (in AHI units), representing the mean AHI for a Group 1 subject aged <65 years with a mean baseline AHI of 15.245.

So, the primary estimand was a difference between 3–month AHI event rate in sham–device patients and 3–month AHI event rate in working–device patients, other things being equal.

The model for the primary estimand

This was a multi–factorial model, regressing 3–month AHI with respect to baseline AHI and randomization age group. The analysis was "treble–blind", so treatment groups were labelled "Group 1" (actually working device) and "Group 2" (actually sham device). The parameters were:

- A Group 2 effect (expressed in AHI units of events per hour). compared to a reference level of Group 1. This was the primary estimand.
- A randomization age group effect for 65+ years (in AHI units), compared to a reference level of <65 years.
- A linear effect of baseline AHI (in units per unit).
- A constant term (in AHI units), representing the mean AHI for a Group 1 subject aged <65 years with a mean baseline AHI of 15.245.

So, the primary estimand was a difference between 3–month AHI event rate in sham–device patients and 3–month AHI event rate in working–device patients, other things being equal.

The model for the primary estimand

This was a multi–factorial model, regressing 3–month AHI with respect to baseline AHI and randomization age group. The analysis was "treble–blind", so treatment groups were labelled "Group 1" (actually working device) and "Group 2" (actually sham device). The parameters were:

- A Group 2 effect (expressed in AHI units of events per hour). compared to a reference level of Group 1. This was the primary estimand.
- A randomization age group effect for 65+ years (in AHI units), compared to a reference level of <65 years.
- A linear effect of baseline AHI (in units per unit).
- A constant term (in AHI units), representing the mean AHI for a Group 1 subject aged <65 years with a mean baseline AHI of 15.245.

So, the primary estimand was a difference between 3–month AHI event rate in sham–device patients and 3–month AHI event rate in working–device patients, other things being equal.

The per–protocol analysis for the primary estimand

The regression model for 3–month AHI with respect to `group` (treatment group), `randagp` (randomization age group), and `ybase` (baseline AHI) gave the following results:

```
      Source |       SS           df       MS      Number of obs   =        92
-------------+----------------------------------   F(3, 88)        =      9.98
       Model |  1959.49338         3  653.164459   Prob > F        =    0.0000
    Residual |  5759.09489        88  65.4442601   R-squared       =    0.2539
-------------+----------------------------------   Adj R-squared   =    0.2284
       Total |  7718.58826        91  84.8196512   Root MSE        =    8.0898

------------------------------------------------------------------------------
E~AHI_E4_C48 | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       group |
    Group 2  |   4.413558   1.687882     2.61   0.011     1.059247    7.767869
             |
     randagp |
        65+  |    .635988   2.013702     0.32   0.753    -3.365821    4.637797
       ybase |   .4601508   .0945316     4.87   0.000     .2722892    .6480125
       _cons |   8.254424   1.286945     6.41   0.000     5.696891    10.81196
------------------------------------------------------------------------------
```

We see that the Group 2 effect was 4.414 events (95% CI, 1.059 to 7.768 events; *P*=0.011). *However*, this was based on 92 patients, a little over 3/4 of the 120 that we originally randomized! And how representative were these 92 patients?

Post hoc sensitivity analysis: completeness–propensity model

- We decided to do this before breaking the blind.
- We found in the data a logit completeness–propensity model, regressing AHI completeness both at baseline and at 3 months with respect to 10 baseline covariates: female gender, ex–smoker status, current smoker status, missing smoking status, Group 2 membership, randomization age in years (centered at 60), randomization age missingness, GP visits in previous month, sick days in previous month, sick days missingness. (Missingness indicators are allowed in propensity models[6].)
- We computed completeness–propensity scores and ATE weights for complete and incomplete patients, using the same formula for AHI completeness as for maternal smoking status in the `cattaneo2` data earlier.
- We then did balance and variance–inflation checks for the completeness–propensity model.
- Finally, we re–ran the regression model on the complete patients, using the ATE weights as `pweights`.

Post hoc sensitivity analysis: completeness–propensity model

- We decided to do this before breaking the blind.
- We found in the data a logit completeness–propensity model, regressing AHI completeness both at baseline and at 3 months with respect to 10 baseline covariates: female gender, ex–smoker status, current smoker status, missing smoking status, Group 2 membership, randomization age in years (centered at 60), randomization age missingness, GP visits in previous month, sick days in previous month, sick days missingness. (Missingness indicators are allowed in propensity models[6].)
- We computed completeness–propensity scores and ATE weights for complete and incomplete patients, using the same formula for AHI completeness as for maternal smoking status in the `cattaneo2` data earlier.
- We then did balance and variance–inflation checks for the completeness–propensity model.
- Finally, we re–ran the regression model on the complete patients, using the ATE weights as `pweights`.

Post hoc sensitivity analysis: completeness–propensity model

- We decided to do this before breaking the blind.
- We found in the data a logit completeness–propensity model, regressing AHI completeness both at baseline and at 3 months with respect to 10 baseline covariates: female gender, ex–smoker status, current smoker status, missing smoking status, Group 2 membership, randomization age in years (centered at 60), randomization age missingness, GP visits in previous month, sick days in previous month, sick days missingness. (Missingness indicators are allowed in propensity models[6].)
- We computed completeness–propensity scores and ATE weights for complete and incomplete patients, using the same formula for AHI completeness as for maternal smoking status in the `cattaneo2` data earlier.
- We then did balance and variance–inflation checks for the completeness–propensity model.
- Finally, we re–ran the regression model on the complete patients, using the ATE weights as `pweights`.

Post hoc sensitivity analysis: completeness–propensity model

- We decided to do this before breaking the blind.
- We found in the data a logit completeness–propensity model, regressing AHI completeness both at baseline and at 3 months with respect to 10 baseline covariates: female gender, ex–smoker status, current smoker status, missing smoking status, Group 2 membership, randomization age in years (centered at 60), randomization age missingness, GP visits in previous month, sick days in previous month, sick days missingness. (Missingness indicators are allowed in propensity models[6].)
- We computed completeness–propensity scores and ATE weights for complete and incomplete patients, using the same formula for AHI completeness as for maternal smoking status in the `cattaneo2` data earlier.
- We then did balance and variance–inflation checks for the completeness–propensity model.
- Finally, we re–ran the regression model on the complete patients, using the ATE weights as `pweights`.

Post hoc sensitivity analysis: completeness–propensity model

- We decided to do this before breaking the blind.
- We found in the data a logit completeness–propensity model, regressing AHI completeness both at baseline and at 3 months with respect to 10 baseline covariates: female gender, ex–smoker status, current smoker status, missing smoking status, Group 2 membership, randomization age in years (centered at 60), randomization age missingness, GP visits in previous month, sick days in previous month, sick days missingness. (Missingness indicators are allowed in propensity models[6].)
- We computed completeness–propensity scores and ATE weights for complete and incomplete patients, using the same formula for AHI completeness as for maternal smoking status in the `cattaneo2` data earlier.
- We then did balance and variance–inflation checks for the completeness–propensity model.
- Finally, we re–ran the regression model on the complete patients, using the ATE weights as `pweights`.

Post hoc sensitivity analysis: completeness–propensity model

- We decided to do this before breaking the blind.
- We found in the data a logit completeness–propensity model, regressing AHI completeness both at baseline and at 3 months with respect to 10 baseline covariates: female gender, ex–smoker status, current smoker status, missing smoking status, Group 2 membership, randomization age in years (centered at 60), randomization age missingness, GP visits in previous month, sick days in previous month, sick days missingness. (Missingness indicators are allowed in propensity models[6].)
- We computed completeness–propensity scores and ATE weights for complete and incomplete patients, using the same formula for AHI completeness as for maternal smoking status in the `cattaneo2` data earlier.
- We then did balance and variance–inflation checks for the completeness–propensity model.
- Finally, we re–ran the regression model on the complete patients, using the ATE weights as `pweights`.

Balance checks for completeness–propensity scores

- These are necessary, as weights that do not balance are not balancing weights.
- *However*, they are different from balance checks for treatment–propensity scores, as we are comparing a subset (the completes) with the full sample, not two exclusive treatment groups with each other.
- To compare the completes with the full sample, we should use the SSC package `scsomersd`, which compares 2 **scenarios** (versions of the same dataset), called “Scenario 0” and “Scenario 1”.
- “Scenario 0” might be the completes, weighted either equally or by inverse completeness–propensity weights.
- “Scenario 1” might be the whole sample (complete or incomplete), weighted equally.
- And we might compare either a propensity score or a component covariate between scenarios.

Balance checks for completeness–propensity scores

- These are necessary, as weights that do not balance are not balancing weights.
- *However*, they are different from balance checks for treatment–propensity scores, as we are comparing a subset (the completes) with the full sample, not two exclusive treatment groups with each other.
- To compare the completes with the full sample, we should use the SSC package `scsomersd`, which compares 2 **scenarios** (versions of the same dataset), called “Scenario 0” and “Scenario 1”.
- “Scenario 0” might be the completes, weighted either equally or by inverse completeness–propensity weights.
- “Scenario 1” might be the whole sample (complete or incomplete), weighted equally.
- And we might compare either a propensity score or a component covariate between scenarios.

Balance checks for completeness–propensity scores

- These are necessary, as weights that do not balance are not balancing weights.
- *However*, they are different from balance checks for treatment–propensity scores, as we are comparing a subset (the completes) with the full sample, not two exclusive treatment groups with each other.
- To compare the completes with the full sample, we should use the SSC package `scsomersd`, which compares 2 **scenarios** (versions of the same dataset), called “Scenario 0” and “Scenario 1”.
- “Scenario 0” might be the completes, weighted either equally or by inverse completeness–propensity weights.
- “Scenario 1” might be the whole sample (complete or incomplete), weighted equally.
- And we might compare either a propensity score or a component covariate between scenarios.

Balance checks for completeness–propensity scores

- These are necessary, as weights that do not balance are not balancing weights.
- *However*, they are different from balance checks for treatment–propensity scores, as we are comparing a subset (the completes) with the full sample, not two exclusive treatment groups with each other.
- To compare the completes with the full sample, we should use the SSC package `scsomersd`, which compares 2 **scenarios** (versions of the same dataset), called "Scenario 0" and "Scenario 1".
- "Scenario 0" might be the completes, weighted either equally or by inverse completeness–propensity weights.
- "Scenario 1" might be the whole sample (complete or incomplete), weighted equally.
- And we might compare either a propensity score or a component covariate between scenarios.

Balance checks for completeness–propensity scores

- These are necessary, as weights that do not balance are not balancing weights.
- *However*, they are different from balance checks for treatment–propensity scores, as we are comparing a subset (the completes) with the full sample, not two exclusive treatment groups with each other.
- To compare the completes with the full sample, we should use the SSC package `scsomersd`, which compares 2 **scenarios** (versions of the same dataset), called "Scenario 0" and "Scenario 1".
- "Scenario 0" might be the completes, weighted either equally or by inverse completeness–propensity weights.
- "Scenario 1" might be the whole sample (complete or incomplete), weighted equally.
- And we might compare either a propensity score or a component covariate between scenarios.

Balance checks for completeness–propensity scores

- These are necessary, as weights that do not balance are not balancing weights.
- *However*, they are different from balance checks for treatment–propensity scores, as we are comparing a subset (the completes) with the full sample, not two exclusive treatment groups with each other.
- To compare the completes with the full sample, we should use the SSC package `scsomersd`, which compares 2 **scenarios** (versions of the same dataset), called "Scenario 0" and "Scenario 1".
- "Scenario 0" might be the completes, weighted either equally or by inverse completeness–propensity weights.
- "Scenario 1" might be the whole sample (complete or incomplete), weighted equally.
- And we might compare either a propensity score or a component covariate between scenarios.

Balance checks for completeness–propensity scores

- These are necessary, as weights that do not balance are not balancing weights.
- *However*, they are different from balance checks for treatment–propensity scores, as we are comparing a subset (the completes) with the full sample, not two exclusive treatment groups with each other.
- To compare the completes with the full sample, we should use the SSC package `scsomersd`, which compares 2 **scenarios** (versions of the same dataset), called "Scenario 0" and "Scenario 1".
- "Scenario 0" might be the completes, weighted either equally or by inverse completeness–propensity weights.
- "Scenario 1" might be the whole sample (complete or incomplete), weighted equally.
- And we might compare either a propensity score or a component covariate between scenarios.

Estimating the unweighted Somers' *D* of completeness–propensity with respect to completeness using **scsomersd**

We assume that variables `ahipres`, `cpropscor`, and `cpropwt` contain AHI completeness, completeness propensity, and completeness propensity weight, respectively. We compute the unweighted Somers' *D* of completeness–propensity with respect to completeness, with "Scenario 0" specified as the unweighted completes by `[pweight=ahipres]`, and "Scenario 1" specified as the unweighted full sample by `sweight(1)`:

```
. scsomersd cpropscor [pweight=ahipres], sweight(1) tdist;
Von Mises Somers' D with variable: _scen0
Transformation: Untransformed
Valid observations: 212
Number of clusters: 120
Degrees of freedom: 119

Symmetric 95% CI
                                 (Std. err. adjusted for 120 clusters in _obs)
------------------------------------------------------------------------------
             |              Jackknife
      _scen0 | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       _yvar |    .102808   .0335109     3.07   0.003      .036453     .169163
------------------------------------------------------------------------------
```

The unweighted Somers *D* of propensity with respect to completeness is 0.103. This means that, if we sample one random patient from the completes and one from the full sample, then it is 10.3% more likely for the more completeness–prone to be the first than to be the second.

Estimating the propensity–weighted Somers' *D* of completeness–propensity with respect to completeness using **scsomersd**

This time, we compute the weighted Somers' *D* of completeness–propensity with respect to completeness, with "Scenario 0" specified as the propensity–weighted completes by `[pweight=ahipres*cpropwt]`, and "Scenario 1" specified as the unweighted full sample by `sweight(1)`:

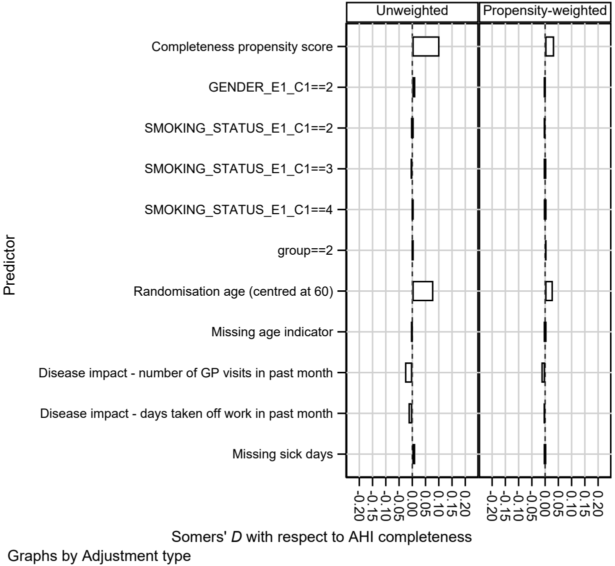
```
. scsomersd cpropscor [pweight=ahipres*cpropwt], sweight(1) tdist;
Von Mises Somers' D with variable: _scen0
Transformation: Untransformed
Valid observations: 212
Number of clusters: 120
Degrees of freedom: 119

Symmetric 95% CI
                                (Std. err. adjusted for 120 clusters in _obs)
------------------------------------------------------------------------------
             |              Jackknife
      _scen0 | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       _yvar |   .0345057   .0362212     0.95   0.343     -.037216    .1062273
------------------------------------------------------------------------------
```

The weighted Somers' *D* of propensity with respect to completeness is 0.035. So, if we sample a random patient from the completes, with probability inversely proportional to propensity, and one equiprobably from the full sample, then it is 3.5% more likely for the more completeness–prone patient to be the first than to be the second.

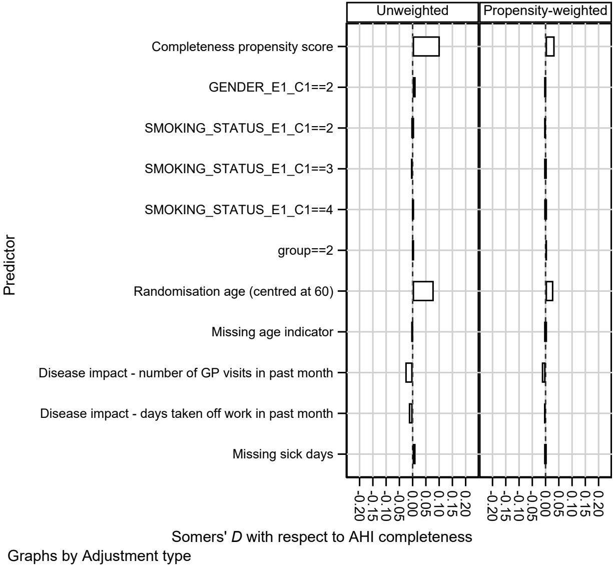
Unweighted and propensity–weighted Somers' *D* of completeness predictors with respect to completeness

- These predictors include the propensity score and its component covariates.
- The unweighted values show that older and/or less diseased patients are more likely to be AHI–complete.
- And the propensity–weighted values show that these associations are *mostly* removed by propensity weighting.



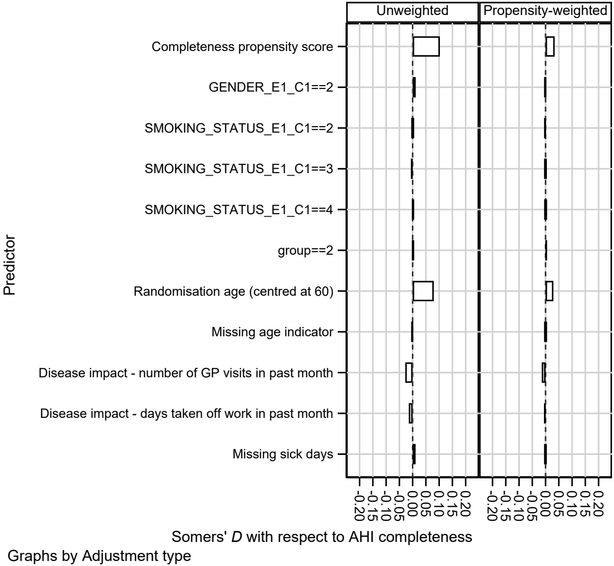
Unweighted and propensity–weighted Somers’ *D* of completeness predictors with respect to completeness

- These predictors include the propensity score and its component covariates.
- The unweighted values show that older and/or less diseased patients are more likely to be AHI–complete.
- And the propensity–weighted values show that these associations are *mostly* removed by propensity weighting.



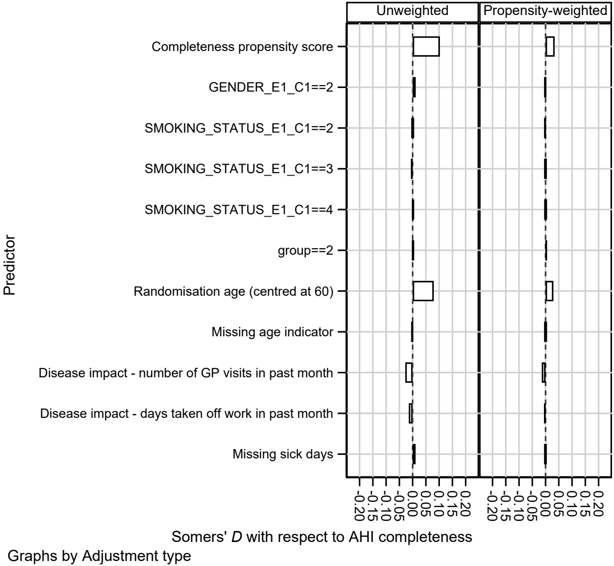
Unweighted and propensity–weighted Somers' *D* of completeness predictors with respect to completeness

- These predictors include the propensity score and its component covariates.
- The unweighted values show that older and/or less diseased patients are more likely to be AHI–complete.
- And the propensity–weighted values show that these associations are *mostly* removed by propensity weighting.



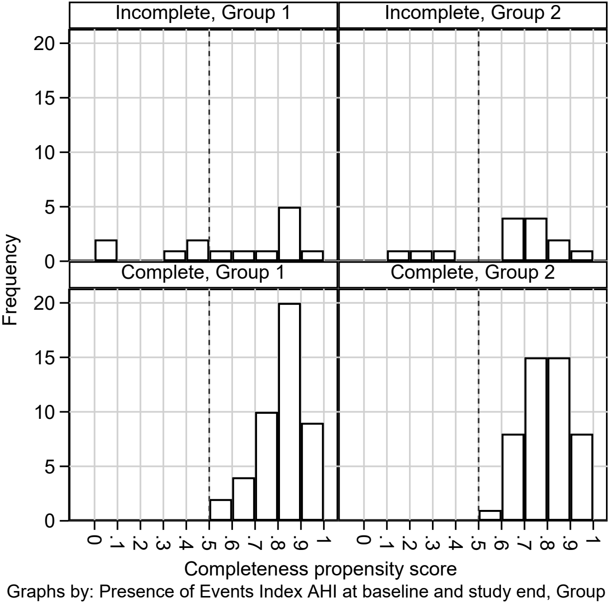
Unweighted and propensity–weighted Somers’ *D* of completeness predictors with respect to completeness

- These predictors include the propensity score and its component covariates.
- The unweighted values show that older and/or less diseased patients are more likely to be AHI–complete.
- And the propensity–weighted values show that these associations are *mostly* removed by propensity weighting.



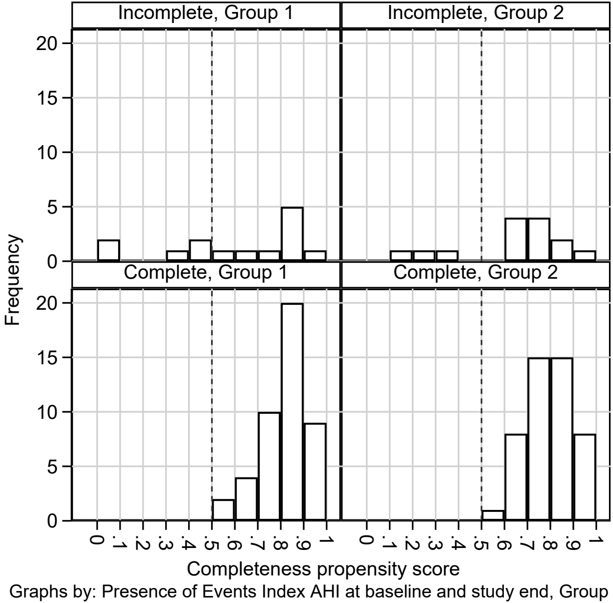
Post–post–hoc caution: A minor non–overlap problem!

- Of the 120 patients randomized, 8 have completeness–propensity less than 0.5 (5 in Group 1, 3 in Group 2).
- And *all 8* of these are incomplete!
- We should therefore *probably* view our inverse–propensity weights as standardizing only to the remaining 112 patients. (Which is the best we can do.)



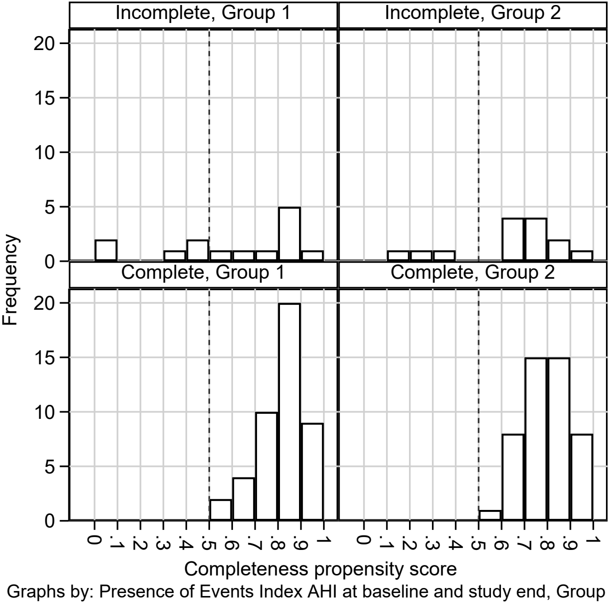
Post–post–hoc caution: A minor non–overlap problem!

- Of the 120 patients randomized, 8 have completeness–propensity less than 0.5 (5 in Group 1, 3 in Group 2).
- And *all 8* of these are incomplete!
- We should therefore *probably* view our inverse–propensity weights as standardizing only to the remaining 112 patients. (Which is the best we can do.)



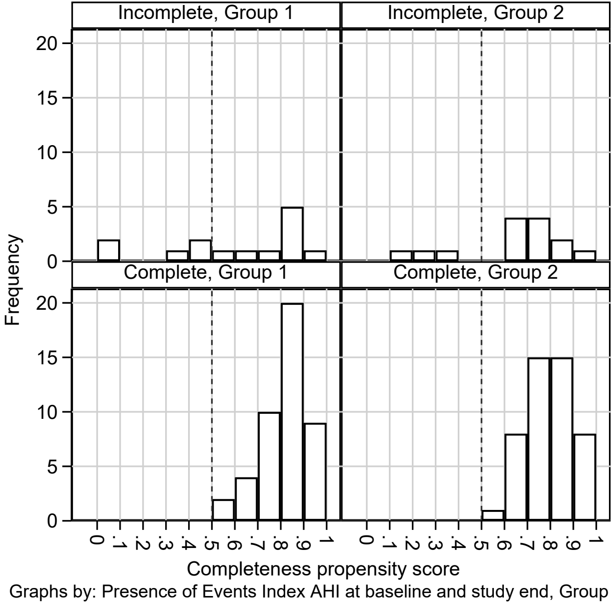
Post–post–hoc caution: A minor non–overlap problem!

- Of the 120 patients randomized, 8 have completeness–propensity less than 0.5 (5 in Group 1, 3 in Group 2).
- And *all 8* of these are incomplete!
- We should therefore *probably* view our inverse–propensity weights as standardizing only to the remaining 112 patients. (Which is the best we can do.)



Post–post–hoc caution: A minor non–overlap problem!

- Of the 120 patients randomized, 8 have completeness–propensity less than 0.5 (5 in Group 1, 3 in Group 2).
- And *all 8* of these are incomplete!
- We should therefore *probably* view our inverse–propensity weights as standardizing only to the remaining 112 patients. (Which is the best we can do.)



Variance–inflation checks using the SSC package `haif`

These are a bit different from variance inflation checks for treatment–propensity weights, as this time they are restricted to the 92 AHI–complete patients:

```
. haif ib1.group ib1.randagp ybase if ahipres, pweight(cpropwt);
Number of observations: 92
Homoskedastic adjustment inflation factors
for variances and standard errors:
               Variance         SE
    1b.group          .          .
     2.group   1.019749   1.009826
  1b.randagp          .          .
   2.randagp   1.007883   1.003934
       ybase   1.010497   1.005235
       _cons   1.021375   1.010631
```

The rows represent model parameters for the per–protocol model for 3–month AHI, with respect to `group` (treatment group), `randagp` (randomization age group), and `ybase` (baseline AHI), now weighted by inverse completeness–propensity weights. We see that very little variance or standard–error inflation is expected, even if the completeness predictors have absolutely no effect on the outcome.

The completeness–propensity weighted analysis for the primary estimand

The regression model for 3–month AHI with respect to `group` (treatment group), `randagp` (randomization age group), and `ybase` (baseline AHI) gave the following output:

```
(sum of wgt is 115.7552886306068)

Linear regression                               Number of obs     =         92
                                                F(3, 88)          =       7.12
                                                Prob > F          =     0.0002
                                                R-squared         =     0.2462
                                                Root MSE          =     8.1676

------------------------------------------------------------------------------
             |               Robust
E~AHI_E4_C48 | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       group |
    Group 2  |   4.153327   1.722201     2.41   0.018     .7308137     7.57584
             |
     randagp |
        65+  |   .5245646   1.888545     0.28   0.782    -3.228522    4.277651
       ybase |   .4697795   .1173178     4.00   0.000     .2366351     .702924
       _cons |   8.469091   1.274587     6.64   0.000     5.936118    11.00206
------------------------------------------------------------------------------
```

We see that the Group 2 effect was 4.153 events (95% CI, 0.731 to 7.576 events; *P*=0.018). This is reassuringly similar to the per–protocol estimate of 4.414 events (95% CI, 1.059 to 7.768 events; *P*=0.011).

POSA *post hoc* sensitivity analysis: Summary

- In the design phase, we found a model to predict AHI completeness from baseline patient features in the 120 randomized patients.
- We used this model to define inverse completeness–propensity weights for the 92 complete patients, hoping to standardize the distribution of baseline features to the 120 randomized patients.
- Balance checks showed that these weights removed most (but not quite all) of the imbalance, as 8 incomplete patients had very low completeness propensity, and could therefore not be represented by up–weighting comparable complete patients.
- *However*, the weights were not expected to inflate the variance of the primary estimand very much.
- *And*, proceeding to the analysis phase, we found the effect estimate to be reassuringly similar to the per–protocol estimate.
- *And*, even more reassuringly, when we broke the blind, we found that the treatment group with the better outcome ("Group 1") was the intervention!

POSA *post hoc* sensitivity analysis: Summary

- In the design phase, we found a model to predict AHI completeness from baseline patient features in the 120 randomized patients.
- We used this model to define inverse completeness–propensity weights for the 92 complete patients, hoping to standardize the distribution of baseline features to the 120 randomized patients.
- Balance checks showed that these weights removed most (but not quite all) of the imbalance, as 8 incomplete patients had very low completeness propensity, and could therefore not be represented by up–weighting comparable complete patients.
- *However*, the weights were not expected to inflate the variance of the primary estimand very much.
- *And*, proceeding to the analysis phase, we found the effect estimate to be reassuringly similar to the per–protocol estimate.
- *And*, even more reassuringly, when we broke the blind, we found that the treatment group with the better outcome (“Group 1”) was the intervention!

POSA *post hoc* sensitivity analysis: Summary

- In the design phase, we found a model to predict AHI completeness from baseline patient features in the 120 randomized patients.
- We used this model to define inverse completeness–propensity weights for the 92 complete patients, hoping to standardize the distribution of baseline features to the 120 randomized patients.
- Balance checks showed that these weights removed most (but not quite all) of the imbalance, as 8 incomplete patients had very low completeness propensity, and could therefore not be represented by up–weighting comparable complete patients.
- *However*, the weights were not expected to inflate the variance of the primary estimand very much.
- *And*, proceeding to the analysis phase, we found the effect estimate to be reassuringly similar to the per–protocol estimate.
- *And*, even more reassuringly, when we broke the blind, we found that the treatment group with the better outcome ("Group 1") was the intervention!

POSA *post hoc* sensitivity analysis: Summary

- In the design phase, we found a model to predict AHI completeness from baseline patient features in the 120 randomized patients.
- We used this model to define inverse completeness–propensity weights for the 92 complete patients, hoping to standardize the distribution of baseline features to the 120 randomized patients.
- Balance checks showed that these weights removed most (but not quite all) of the imbalance, as 8 incomplete patients had very low completeness propensity, and could therefore not be represented by up–weighting comparable complete patients.
- *However*, the weights were not expected to inflate the variance of the primary estimand very much.
- *And*, proceeding to the analysis phase, we found the effect estimate to be reassuringly similar to the per–protocol estimate.
- *And*, even more reassuringly, when we broke the blind, we found that the treatment group with the better outcome ("Group 1") was the intervention!

POSA *post hoc* sensitivity analysis: Summary

- In the design phase, we found a model to predict AHI completeness from baseline patient features in the 120 randomized patients.
- We used this model to define inverse completeness–propensity weights for the 92 complete patients, hoping to standardize the distribution of baseline features to the 120 randomized patients.
- Balance checks showed that these weights removed most (but not quite all) of the imbalance, as 8 incomplete patients had very low completeness propensity, and could therefore not be represented by up–weighting comparable complete patients.
- *However*, the weights were not expected to inflate the variance of the primary estimand very much.
- *And*, proceeding to the analysis phase, we found the effect estimate to be reassuringly similar to the per–protocol estimate.
- *And*, even more reassuringly, when we broke the blind, we found that the treatment group with the better outcome ("Group 1") was the intervention!

POSA *post hoc* sensitivity analysis: Summary

- In the design phase, we found a model to predict AHI completeness from baseline patient features in the 120 randomized patients.
- We used this model to define inverse completeness–propensity weights for the 92 complete patients, hoping to standardize the distribution of baseline features to the 120 randomized patients.
- Balance checks showed that these weights removed most (but not quite all) of the imbalance, as 8 incomplete patients had very low completeness propensity, and could therefore not be represented by up–weighting comparable complete patients.
- *However*, the weights were not expected to inflate the variance of the primary estimand very much.
- *And*, proceeding to the analysis phase, we found the effect estimate to be reassuringly similar to the per–protocol estimate.
- *And*, even more reassuringly, when we broke the blind, we found that the treatment group with the better outcome (“Group 1”) was the intervention!

POSA *post hoc* sensitivity analysis: Summary

- In the design phase, we found a model to predict AHI completeness from baseline patient features in the 120 randomized patients.
- We used this model to define inverse completeness–propensity weights for the 92 complete patients, hoping to standardize the distribution of baseline features to the 120 randomized patients.
- Balance checks showed that these weights removed most (but not quite all) of the imbalance, as 8 incomplete patients had very low completeness propensity, and could therefore not be represented by up–weighting comparable complete patients.
- *However*, the weights were not expected to inflate the variance of the primary estimand very much.
- *And*, proceeding to the analysis phase, we found the effect estimate to be reassuringly similar to the per–protocol estimate.
- *And*, even more reassuringly, when we broke the blind, we found that the treatment group with the better outcome ("Group 1") was the intervention!

References

[1] Rubin, D. B. 2008. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics* **2(3)**: 808–840.

[2] Newson, R. B. and Falcaro, M. 2023. Robit regression in Stata. *The Stata Journal* **23(3)**: 658–682.

[3] Newson, R. B. 2016. The role of Somers' *D* in propensity modelling. Presented at the 22nd UK Stata User Meeting, 8—9 September, 2016. Downloadable from the conference website at *https://ideas.repec.org/p/boc/usug16/01.html*

[4] Seaman, S. R. and White, I. R. 2011. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research* **22**: 278–295.

[5] ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000 Feb 29 –. Identifier NCT04153240, The POSA Trial – Positional Therapy for Positional OSA (POSA); 2023 March 30 [cited 19 January 2024]; [about 4 screens]. Available from: *https://clinicaltrials.gov/study/NCT04153240*.

[6] Rosenbaum, P. R. and Rubin, D. B. 1984. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association* **79 (387)**: 516–524.

The presentation, and the example do–files, can be downloaded from the conference website. The packages can be downloaded from SSC.