Propensity score weighting for covariate adjustment in randomized clinical trials

Fan Li

Department of Biostatistics
Center for Methods in Implementation and Prevention Science
Yale School of Public Health

Online Causal Inference Seminar
November 24, 2020

Joint work with Shuxi Zeng (Duke), Rui Wang (Harvard) and Fan Li (Duke)
Randomized controlled trials

- Randomized controlled trials (RCTs) – gold standard for evaluating the efficacy and safety of new treatments
  - balances both measured and unmeasured confounders in expectation
  - ensures internal validity
- Difference-in-means (unadjusted) estimator unbiased for treatment effect
- Important patient characteristics are collected at baseline
- **Chance imbalance** can often occur in a single trial
  - face validity
  - statistical efficiency and power
Example: “Table 1” in BestAIR RCT
(Bakker et al. 2016)

<table>
<thead>
<tr>
<th></th>
<th>All patients</th>
<th>CPAP group</th>
<th>Control group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$N = 169$</td>
<td>$N_1 = 83$</td>
<td>$N_0 = 86$</td>
</tr>
<tr>
<td><strong>Baseline categorical covariates and number of units in each group.</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender (male)</td>
<td>107</td>
<td>54</td>
<td>53</td>
</tr>
<tr>
<td>Race &amp; ethnicity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>152</td>
<td>75</td>
<td>77</td>
</tr>
<tr>
<td>Black</td>
<td>11</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Other</td>
<td>5</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td><strong>Recruiting center</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Site 1</td>
<td>54</td>
<td>26</td>
<td>28</td>
</tr>
<tr>
<td>Site 2</td>
<td>10</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Site 3</td>
<td>105</td>
<td>52</td>
<td>53</td>
</tr>
<tr>
<td><strong>Baseline continuous covariates, mean and standard deviation (in parenthesis).</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (years)</td>
<td>64.4 (7.4)</td>
<td>64.4 (8.0)</td>
<td>64.3 (6.8)</td>
</tr>
<tr>
<td>BMI (kg/m$^2$)</td>
<td>31.7 (6.0)</td>
<td>31.0 (5.3)</td>
<td>32.4 (6.5)</td>
</tr>
<tr>
<td>Baseline SBP (mmHg)</td>
<td>124.3 (13.2)</td>
<td>121.6 (11.1)</td>
<td>127.0 (14.6)</td>
</tr>
<tr>
<td>Baseline SDP (beats/minute)</td>
<td>63.1 (10.7)</td>
<td>63.0 (10.4)</td>
<td>63.2 (10.9)</td>
</tr>
<tr>
<td>Baseline AHI (events/hr)</td>
<td>28.8 (15.4)</td>
<td>26.5 (13.0)</td>
<td>31.1 (17.2)</td>
</tr>
<tr>
<td>Baseline ESS</td>
<td>8.3 (4.5)</td>
<td>8.0 (4.5)</td>
<td>8.5 (4.6)</td>
</tr>
</tbody>
</table>
Outcome regression for covariate adjustment

- Regression adjustment via ANCOVA
  - outcome is modeled as a function of treatment, (centered) covariates and their interactions

- Improve power over the unadjusted estimator

- Under some conditions, asymptotic equivalent to the semiparametric efficient estimator (Tsiatis et al. 2008)

- Unbiased point estimator even under misspecification

- Valid model-based variance under misspecification with balanced allocation (Wang et al. 2019)
Misspecification of the outcome model decreases precision in unbalanced experiments with treatment effect heterogeneity (Freedom, 2008)

Potential for inviting a ‘fishing expedition’ (Williamson et al. 2014; Zeng et al. 2020)

Still most commonly used in biomedical studies
Inverse probability weighting for covariate adjustment

- **IPW** may serve as an objective alternative to ANCOVA in RCTs
  - **Known** treatment assignment mechanism is modeled as a function of baseline covariates
  - Create the inverse propensity score weights, and difference in the weighted mean outcomes between groups
- Propensity score model always correctly specified, and hence point estimator unbiased
Inverse probability weighting for covariate adjustment
(Williamson et al., 2014)

- Advantages
  - separates the design and analysis without involving outcome in the design stage
  - avoids the fishing expedition by promoting objectivity in pre-specifying the analytical adjustment
  - avoids convergence issues with regression under rare outcomes

- Limitation
  - may be inefficient compared to ANCOVA with limited sample sizes and unbalanced treatment allocations
Objective and notation

- Propose to weight beyond IPW for covariate adjustment
  - maintain the objectivity of weighting, but could improve finite-sample performance of IPW

- Notation
  - $N$: total sample size
  - $Z$: randomized treatment indicator $\in \{0, 1\}$
  - $Y(1), Y(0)$: potential outcomes under treatment and control
  - $X = (X_1, \ldots, X_p)^T$: recorded baseline variables

- Interested in the additive causal estimand, ATE

  \[ \tau = \mathbb{E}[Y(1) - Y(0)] = \mu_1 - \mu_0 \]
Assumptions

- SUTVA: the observed outcome $Y = ZY(1) + (1 - Z)Y(0)$

- Randomization: define $0 < r < 1$ as the randomization probability

$$P(Z = 1|X, Y(1), Y(0)) = P(Z = 1) = r$$

- most typically $r = 1/2$, but other values possible with perceived benefit of the treatment

- Under these assumptions, $\tau$ is easily identified by

$$\mathbb{E}(Y|Z = 1) - \mathbb{E}(Y|Z = 0),$$

which motivates the unadjusted difference-in-means estimator
Weighted average treatment effect
(Li, Morgan, Zaslavsky, 2018)

- ATE is a special case of a class of weighted average treatment effect (WATE)

- Recall the conditional average treatment effect (CATE) is

\[ \tau(x) \equiv \mathbb{E}(Y(1)|X = x) - \mathbb{E}(Y(0)|X = x). \]

- Assume density of the observed covariates, \( f(x) \), exists

- Consider a target population, denoted by a density \( g(x) \), possibly different from \( f(x) \)

- The ratio \( h(x) = g(x)/f(x) \) is called a *tilting function*, which re-weights the observed sample to represent the target population
Weighted average treatment effect - Cont’d

- Class of estimands: the ATE over the target population \( g \)

\[
\tau^h \equiv \mathbb{E}_g[Y(1) - Y(0)] = \frac{\int \tau(x)f(x)h(x)\mu(dx)}{\int f(x)h(x)\mu(dx)} = \frac{\mathbb{E}\{h(x)\tau(x)\}}{\mathbb{E}\{h(x)\}}.
\]

- When \( h(x) \propto 1, f(x) = g(x) \), the target population is the observed population; \( \tau_h \) is the ATE

- Varying \( h(x) \) can vary target population. In practice, we pre-specify the tilting function \( h(x) \) based on \( e(x) \)

- Under randomization, \( e(x) = P(Z = 1) = r \), as long as \( h(x) \) is a function of \( e(x) \), different \( h \) corresponds to the same \( g \), and WATE becomes the ATE
Balancing weights

Li, Morgan, Zaslavsky, 2018

- This last special feature under RCT provides the basis for alternative weighting strategies to achieve better finite-sample performance in covariate adjustment

- For a given $h(x)$, to estimate $\tau^h$, we can use balancing weights to construct weighting estimators

$$
\begin{cases}
w_1(x) \propto \frac{f(x)h(x)}{f_1(x)} = \frac{f(x)h(x)}{f(x)e(x)} = \frac{h(x)}{e(x)}, \\
w_0(x) \propto \frac{f(x)h(x)}{f_0(x)} = \frac{f(x)h(x)}{f(x)(1-e(x))} = \frac{h(x)}{1-e(x)}.
\end{cases}
$$

- The class of weights $(w_0, w_1)$ is called balancing weights: balance the distributions of the weighted covariates between comparison groups
Sample weighting estimators

- Hájek-type estimator of WATE

\[
\hat{\tau}^h = \hat{\mu}_1^h - \hat{\mu}_0^h = \frac{\sum_{i=1}^{N} w_1(x_i)Z_iY_i}{\sum_{i=1}^{N} w_1(x_i)Z_i} - \frac{\sum_{i=1}^{N} w_0(x_i)(1 - Z_i)Y_i}{\sum_{i=1}^{N} w_0(x_i)(1 - Z_i)}.
\]

- inverse probability weights: \((w_1, w_0) = (1/e(x), 1/\{1 - e(x)\})\)

- overlap weights: \((w_1, w_0) = (1 - e(x), e(x))\)

- ATT weights and matching weights are also members of the class of balancing weights

- Choices of \(h\) modifies the target estimands in observational studies, but corresponds to same ATE in RCTs
Overlap weights

- In observational studies, the overlap weights (OW) correspond to a target population with the most overlap in the baseline characteristics
  - theoretically to give the smallest asymptotic variance of $\hat{\tau}^h$
  - empirically reduce the variance of $\tau^h$ in finite samples.

- In RCTs, the true propensity score is constant, OW and IPW target the same estimand $\tau$, but their finite-sample operating characteristics can be markedly different

- Consider applying the OW estimator to RCTs

$$\hat{\tau}^{\text{OW}} = \hat{\mu}_1 - \hat{\mu}_0 = \frac{\sum_{i=1}^{N} (1 - \hat{e}_i)Z_iY_i}{\sum_{i=1}^{N}(1 - \hat{e}_i)Z_i} - \frac{\sum_{i=1}^{N} \hat{e}_i(1 - Z_i)Y_i}{\sum_{i=1}^{N} \hat{e}_i(1 - Z_i)},$$

where $\hat{e}_i$ is the estimated propensity scores from a working model.
The working propensity score model is often the logistic model with

\[ e_i = e(X_i; \theta) = \frac{\exp(\theta_0 + X_i^T \theta_1)}{1 + \exp(\theta_0 + X_i^T \theta_1)}, \]

with \( \theta = (\theta_0, \theta_1^T)^T \) and \( \hat{\theta} \) is the MLE

- \( X \) include stratification variables and other key prognostic factors pre-specified in the design stage (Zeng et al. 2020)

- The purpose is not to learn the assignment mechanism (hence flexible models may not be as useful), but to implicitly perform covariate adjustment
Exact balance property

(Li, Morgan, Zaslavsky, 2018)

- OW estimated from logistic model lead to exact mean balance of any predictor included in the model (Theorem 3 in Li et al. (2018)):

\[
\frac{\sum_{i=1}^{N} (1 - \hat{e}_i) Z_i X_{ji}}{\sum_{i=1}^{N} (1 - \hat{e}_i) Z_i} - \frac{\sum_{i=1}^{N} \hat{e}_i (1 - Z_i) X_{ji}}{\sum_{i=1}^{N} \hat{e}_i (1 - Z_i)} = 0, \quad \text{for } j = 1, \ldots, p.
\]

- Implication for RCTs:
  - weighted differences in the usual "Table 1" are identically zero (face validity)
Implications with exact balance

(Zeng et al., 2020)

- Exact mean balance property translates into better efficiency in estimating $\tau$
- Consider an additive outcome surface $Y_i = \alpha + Z_i \tau + X_i^T \beta_0 + \epsilon_i$ with $E(\epsilon_i|Z_i, X_i) = 0$
- Sources of estimation error in RCTs come from weighted chance imbalance, defined as
  \[
  \Delta_X(w_0, w_1) = \frac{\sum_{i=1}^N w_1(X_i)Z_iX_i}{\sum_{i=1}^N w_1(X_i)Z_i} - \frac{\sum_{i=1}^N w_0(X_i)(1 - Z_i)X_i}{\sum_{i=1}^N w_0(X_i)(1 - Z_i)},
  \]
  and weighted difference in noise, defined as
  \[
  \Delta_\epsilon(w_0, w_1) = \frac{\sum_{i=1}^N w_1(X_i)Z_i\epsilon_i}{\sum_{i=1}^N w_1(X_i)Z_i} - \frac{\sum_{i=1}^N w_0(X_i)(1 - Z_i)\epsilon_i}{\sum_{i=1}^N w_0(X_i)(1 - Z_i)}.\]
Implications with exact balance

(Zeng et al., 2020)

- Error of the unadjusted estimator

\[ \hat{\tau}^{\text{UNADJ}} - \tau = \Delta_X(1, 1)^T \beta_0 + \Delta_\varepsilon(1, 1) \]

explodes when \( X \) exhibits chance imbalance and \( \beta_0 \) large

- Error of the IPW estimator

\[ \hat{\tau}^{\text{IPW}} - \tau = \Delta_X(1/(1 - \hat{e}), 1/\hat{e})^T \beta_0 + \Delta_\varepsilon(1/(1 - \hat{e}), 1/\hat{e}) \]

is reduced because usually \( \|\Delta_X(1/(1 - \hat{e}), 1/\hat{e})\| < \|\Delta_X(1, 1)\| \)

- Error of the OW estimator is free of \( X \) and \( \beta_0 \) because

\[ \Delta_X(\hat{e}, 1 - \hat{e}) = 0, \]

\[ \hat{\tau}^{\text{OW}} - \tau = \Delta_\varepsilon(\hat{e}, 1 - \hat{e}) \]
A class of estimators

Family of RAL estimators for \( \tau \) is (Tsiatis et al. (2008))

\[
I : \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{Z_iY_i}{r} - \frac{(1-Z_i)Y_i}{1-r} - \frac{Z_i-r}{r(1-r)} \{ rg_0(X_i) + (1-r)g_1(X_i) \} \right\} + o_p(N^{-1/2}),
\]

where \( g_0(X_i), g_1(X_i) \) are scalar functions of \( X_i \)

- Unadjusted estimator: \( g_z(X_i) = 0 \)

- ANCOVA I (\( Y \sim Z + X \)): \( g_z(X_i) = E(Y_i | X_i) \forall z \)

- ANCOVA II (\( Y \sim Z + X + ZX \)): \( g_z(X_i) = E(Y_i | Z_i = z, X_i) \); fully efficient with correct model

- IPW with working logistic \( \hat{e}_i \): asymptotically equivalent to ANCOVA II when \( g_z(X_i) \) linear in \( X_i \) (Shen et al. 2013)
Large-sample properties of OW

Proposition 1 (a) If the PS is estimated by a parametric model $e(X; \theta)$ that satisfies a set of mild regularity conditions, then $\hat{\tau}^{OW}$ belongs to the class of estimators $\mathcal{I}$.

(b) Suppose $X^1$ and $X^2$ are two nested sets of baseline covariates with $X^2 = (X^1, X^*^1)$, and $e(X^1; \theta_1)$, $e(X^2; \theta_2)$ are nested smooth parametric models. Write $\hat{\tau}_1^{OW}$ and $\hat{\tau}_2^{OW}$ as two OW estimators with the weights defined through $e(X^1; \hat{\theta}_1)$ and $e(X^2; \hat{\theta}_2)$, respectively. Then

$$\text{avar}(\hat{\tau}_2^{OW}) \leq \text{avar}(\hat{\tau}_1^{OW})$$

(c) With logistic PS, $\hat{\tau}^{OW}$ is asymptotically equivalent to ANCOVA II, and is semiparametric efficient as long as the true $E(Y_i|X_i, Z_i = z)$ is linear in $X_i$. 
Large-sample properties of OW

- Extending results specific to IPW (Shen et al. 2013)
  - adjusting for the baseline covariates using OW does not adversely affect efficiency in large samples than without adjustment

- Can be fully efficient when the true outcome is linear

- When \( r = 1/2 \),

\[
\lim_{N \to \infty} N\text{Var}(\hat{\tau}^{\text{OW}}) = (1 - R^2_{\tilde{Y} \sim X}) \lim_{N \to \infty} N\text{Var}(\hat{\tau}^{\text{UNADJ}}) = 4(1 - R^2_{\tilde{Y} \sim X})\text{Var}(\tilde{Y}_i),
\]

where \( \tilde{Y}_i = Z_i(Y_i - \mu_1) + (1 - Z_i)(Y_i - \mu_0) \) is the mean-centered outcome, \( R^2_{\tilde{Y} \sim X} \) measures the proportion of explained variance after regressing \( \tilde{Y}_i \) on \( X_i \).

- Choice of \( X \)’s: chance imbalance and prognostic values
Large-sample properties of OW

The results in Proposition 1 apply more broadly

**Proposition 2**

Proposition 1 holds for the general family of estimators using balancing weights \( \{w_1(x) = h(x)/e(x), w_0 = h(x)/(1 - e(x))\} \), as long as the tilting function \( h(X) \) is a smooth function of the propensity score, where smoothness is defined by satisfying a set of mild regularity conditions.

- Smoothness requires differentiability

- Matching weights are defined with \( h(x) = \min\{e(x), 1 - e(x)\} \); results can still hold once we smooth over the non-differentiable point at \( e(x) = 1/2 \)
Binary outcomes

- Target estimand: causal risk difference, risk ratio and odds ratio

- Ratio estimands on the log scale

\[ \tau_{RR} = \log \left( \frac{\mu_1}{\mu_0} \right), \quad \tau_{OR} = \log \left\{ \frac{\mu_1/(1 - \mu_1)}{\mu_0/(1 - \mu_0)} \right\} . \]

- IPW leads to improved efficiency improvement over the unadjusted in RCTs (Williamson et al., 2014), but OW may have better finite-sample properties
Variance estimation

- Obtain the usual sandwich variance for balancing weights estimators

- Can summarize the variance estimators for $\hat{\tau}^{\text{OW}}_{\text{RD}}$, $\hat{\tau}^{\text{OW}}_{\text{RR}}$, $\hat{\tau}^{\text{OW}}_{\text{OR}}$

$$\text{Var}(\hat{\tau}^{\text{OW}}) = \frac{1}{N} \left[ \hat{V}^{\text{UNADJ}} - \hat{v}_1^T \left\{ \frac{1}{N} \sum_{i=1}^{N} \hat{e}_i (1 - \hat{e}_i) \tilde{X}_i^T \tilde{X}_i \right\}^{-1} (2\hat{v}_1 - \hat{v}_2) \right],$$

where $\text{Var}(\hat{\tau}^{\text{OW}})$ is the variance of the unadjusted estimator, $\hat{v}_1$, $\hat{v}_2$ depend on the choice of estimands

- Form of the variance estimator shed light on the variance reduction property of OW
Simulation design: continuous outcomes

- Generate $p = 10$ baseline covariates from the standard normal
- Randomize treatment based on $Z_i \sim \text{Bern}(r)$ ($r = 0.5, 0.7$)
- Simulate $Y_i(z) \sim N(z \alpha + X_i^T \beta_0 + zX_i^T \beta_1, \sigma_y^2)$
- Specify $\beta_0$ such that signal-to-noise ratio is 1 but with varying importance attached to components of $X$
- Vary $\beta_1 \in \{0, 0.25, 0.5, 0.75\}$ to represent levels of treatment effect heterogeneity
- Interested in small ($N = 50$) to moderate ($N = 200$) RCTs; 2000 simulations
Simulation design: continuous outcomes

Several comparators

- Unadjusted estimator
- IPW and OW: consider linear specification of $X$ in logistic model
- LR: ANCOVA II model
- AIPW: combining IPW with regression (the usual DR estimator applied to RCTs)

Variance estimation:

- IPW, OW and AIPW: sandwich variance via M-estimation
- LR: Huber-White estimator suggested in Lin (2013)
Relative efficiency: $r = 0.5$, constant effect

- Efficiency: $\text{OW} \geq \text{LR} \geq \text{IPW}$
- Advantage of OW in small samples $N \leq 100$
- Equivalent for large $N = 200$
- AIPW almost equal to LR (regression dominates)
Relative efficiency: $r = 0.5$, strong HTE

- Efficiency: $LR \geq OW \geq IPW$
- Correct outcome model and balanced design favors LR in small samples
- Advantage of LR decreases with reduced degree of HTE
- Equivalent for larger $N$, say over 500
- OW always beat IPW
Relative efficiency: $r = 0.7$, constant effect

- Left: correct outcome model – LR (blue) less efficient in small samples
- Right: incorrect outcome model – LR (blue) less efficient
- OW dominates in both cases
Inference for continuous outcomes

- Coverage for IPW and OW close to nominal with the sandwich variance

- Huber-White variance for LR unstable, and severely biased towards zero in small samples and leads to under-coverage when $r \neq 0.5$

  - misspecification

  - strong HTE

- AIPW variance similar to LR
Simulations with binary outcomes

Estimation

- Covariate adjustment leads to efficiency improvement likely when $N \geq 100$, except under strong HTE
- Correct outcome model can be more efficient than weighting in small samples only for common (non-rare) outcomes
- OW always better than IPW

Inference in finite-samples

- Sandwich variance for OW has smallest finite-sample bias, and the Huber-White variance for logistic regression largest bias
Best Apnea Interventions for Research (BestAIR) RCT

- **Goal**: Evaluate the effect of continuous positive airway pressure (CPAP) treatment on the health outcomes of patients with obstructive sleep apnea

- **Sample Size**: 83 patients in the active CPAP group and 86 patients in the sham control arm

- **Outcome**: SBP and daily sleepiness measured by Epworth Sleepiness Scale (ESS)

- **Covariates**: demographics (e.g. age, gender, ethnicity), BMI, Apnea-Hypopnea Index (AHI), average seated radial pulse rate (SDP), site and baseline outcome measures (e.g. baseline blood pressure and ESS)
Chance imbalance and face validity

<table>
<thead>
<tr>
<th>Baseline categorical covariates and number of units in each group.</th>
<th>All patients $N = 169$</th>
<th>CPAP group $N_1 = 83$</th>
<th>Control group $N_0 = 86$</th>
<th>ASD$^{\text{UNADJ}}$</th>
<th>ASD$^{\text{IPW}}$</th>
<th>ASD$^{\text{OW}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (male)</td>
<td>107</td>
<td>54</td>
<td>53</td>
<td>0.046</td>
<td>0.002</td>
<td>0.000</td>
</tr>
<tr>
<td>Race &amp; ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>152</td>
<td>75</td>
<td>77</td>
<td>0.051</td>
<td>0.015</td>
<td>0.000</td>
</tr>
<tr>
<td>Black</td>
<td>11</td>
<td>5</td>
<td>6</td>
<td>0.060</td>
<td>0.007</td>
<td>0.000</td>
</tr>
<tr>
<td>Other</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>0.086</td>
<td>0.034</td>
<td>0.000</td>
</tr>
<tr>
<td>Recruiting center</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Site 1</td>
<td>54</td>
<td>26</td>
<td>28</td>
<td>0.046</td>
<td>0.002</td>
<td>0.000</td>
</tr>
<tr>
<td>Site 2</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td>0.065</td>
<td>0.024</td>
<td>0.000</td>
</tr>
<tr>
<td>Site 3</td>
<td>105</td>
<td>52</td>
<td>53</td>
<td>0.073</td>
<td>0.013</td>
<td>0.000</td>
</tr>
<tr>
<td>Baseline continuous covariates, mean and standard deviation (in parenthesis).</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (years)</td>
<td>64.4 (7.4)</td>
<td>64.4 (8.0)</td>
<td>64.3 (6.8)</td>
<td>0.020</td>
<td>0.017</td>
<td>0.000</td>
</tr>
<tr>
<td>BMI (kg/m$^2$)</td>
<td>31.7 (6.0)</td>
<td>31.0 (5.3)</td>
<td>32.4 (6.5)</td>
<td>0.261</td>
<td>0.042</td>
<td>0.000</td>
</tr>
<tr>
<td>Baseline SBP (mmHg)</td>
<td>124.3 (13.2)</td>
<td>121.6 (11.1)</td>
<td>127.0 (14.6)</td>
<td>0.477</td>
<td>0.020</td>
<td>0.000</td>
</tr>
<tr>
<td>Baseline SDP (beats/minute)</td>
<td>63.1 (10.7)</td>
<td>63.0 (10.4)</td>
<td>63.2 (10.9)</td>
<td>0.020</td>
<td>0.016</td>
<td>0.000</td>
</tr>
<tr>
<td>Baseline AHI (events/hr)</td>
<td>28.8 (15.4)</td>
<td>26.5 (13.0)</td>
<td>31.1 (17.2)</td>
<td>0.348</td>
<td>0.039</td>
<td>0.000</td>
</tr>
<tr>
<td>Baseline ESS</td>
<td>8.3 (4.5)</td>
<td>8.0 (4.5)</td>
<td>8.5 (4.6)</td>
<td>0.092</td>
<td>0.010</td>
<td>0.000</td>
</tr>
</tbody>
</table>

- exemplify the “removal” of chance imbalance
Data analysis

<table>
<thead>
<tr>
<th>Method</th>
<th>Estimate</th>
<th>Standard error</th>
<th>95% Confidence interval</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Systolic blood pressure (continuous)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UNADJ</td>
<td>-5.070</td>
<td>2.345</td>
<td>(-9.667, -0.473)</td>
<td>0.031</td>
</tr>
<tr>
<td>IPW</td>
<td>-2.638</td>
<td>1.634</td>
<td>(-5.841, 0.566)</td>
<td>0.107</td>
</tr>
<tr>
<td>LR</td>
<td>-2.790</td>
<td>1.724</td>
<td>(-6.169, 0.588)</td>
<td>0.106</td>
</tr>
<tr>
<td>AIPW</td>
<td>-2.839</td>
<td>1.642</td>
<td>(-6.058, 0.380)</td>
<td>0.084</td>
</tr>
<tr>
<td>OW</td>
<td>-2.777</td>
<td>1.689</td>
<td>(-6.088, 0.534)</td>
<td>0.100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Epworth Sleepiness Scale (continuous)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UNADJ</td>
<td>-1.503</td>
<td>0.702</td>
<td>(-2.878, -0.128)</td>
<td>0.032</td>
</tr>
<tr>
<td>IPW</td>
<td>-1.232</td>
<td>0.486</td>
<td>(-2.184, -0.279)</td>
<td>0.011</td>
</tr>
<tr>
<td>LR</td>
<td>-1.260</td>
<td>0.519</td>
<td>(-2.276, -0.243)</td>
<td>0.015</td>
</tr>
<tr>
<td>AIPW</td>
<td>-1.255</td>
<td>0.479</td>
<td>(-2.193, -0.317)</td>
<td>0.009</td>
</tr>
<tr>
<td>OW</td>
<td>-1.251</td>
<td>0.491</td>
<td>(-2.214, -0.288)</td>
<td>0.011</td>
</tr>
</tbody>
</table>

- An illustration of consequences of ignoring baseline imbalance in an important baseline covariates
- Covariate adjusted results are more or less similar in BestAIR
Regression versus weighting

- Though commonly used, regression may invite a “fishing expedition” and come with additional caveats
  - rare outcomes and unbalanced randomization
  - model misspecification and efficiency
  - inference with the Huber-White sandwich variance

- We demonstrate balancing weights can serve as an alternative way to adjust for covariates and improve precision in RCTs
  - asymptotic equivalence to efficient ANCOVA
  - can do better in weighting via OW than IPW
Reflection on overlap weights

- A continuum of study designs characterized by degree of overlap
- Under weak overlap, OW move the goalpost to gain efficiency over IPW
- Under good overlap, ATO $\approx$ ATE but more efficient
- In the limit (RCT), OW estimate ATE, still more efficient than IPW
Potential misconception on weighting

- Raad et al. (2020) demonstrated superior coverage of the linear regression interval estimator over the IPW interval estimator.

- Only considered **model-based variance** when the outcome regression is correctly specified.

- Further assumes $r = 1/2$, and therefore valid (Wang et al., 2019).

- Inherit limitation to IPW but not weighting in general, since improvement can be made via OW.

- OW require a one-line change of code; more involved variance calculation included in \texttt{PSweight R} package.
R package: PSweight
(Zhou et al., 2020+)

PSweight: R package on CRAN that provides a wide range of propensity score weighting methods, incorporating:

- Overlap weighting
- Inverse probability weighting, with or without trimming
- Binary treatment and multiple treatments
- Simple weighting estimator and augmented estimator
- Continuous, binary, count (survival forthcoming) outcomes
- Diagnostic tables and graphics

Link on CRAN:
https://cran.r-project.org/web/packages/PSweight/index.html
Potential applications

▶ Chance imbalance in other study designs

▶ pre-specified subgroup analysis of RCTs (limited sample size and lower power)

▶ Multi-arm randomized trials

▶ Cluster- or group-randomized trials

▶ …


