Overlap, inverse probability, and matching weights: what are we weighting for?

Yi Liu

Department of Statistics, North Carolina State University

Contact Information:
Phone: +1 (919) 308 7746
Email: yliu297@ncsu.edu

Introduction

Propensity score (PS) methods are used in mitigating covariate imbalance of treatment groups in non-randomized studies, playing a central role in darwing causal conclusions. Controversies have emerged about the goals of different PS weights, their estimands, and targeted populations. In this poster, we present our research about:

- What to expect when using overlap weights (OW), matching weights (MW), or entropy weights (EW) and compare to IPW weights;
- What is the role of the proportion of participants in the treatment groups ($p = P(Z = 1)$) in estimating these quantities.

Why do we care?

- Either ATT or ATC can be primary interested estimand. When exposure is rare (samll p) and the treated population is of interest, ATT is encouraged and ATE reaches extreme values and larger biases[1]. Then who can also be a replacement of ATT here? What if p is large?
- How sensitive are these estimands to model misspecification under finite sample?

Method

WATE and Balancing Weights

Denote Z as indicator of treatment (1 for treatment, 0 for control), $X = (X_1, ..., X_p)'$ as covariates, $e(x) = P(Z = 1|X = x)$ as propensity score, $Y(z), z = 0, 1$ as potential outcome associated with the treatment. A class of estimand, *weighted average treatment effect (WATE)*, is defined by[2]

$$\tau_g = \frac{E[g(X)\tau(X)]}{E[g(X)]}$$

where $\tau(x) = E[Y(1) - Y(0)|X = x]$, $g(x)$ is a *selection function* which defines the target population. Table 1 gives $g(x)$ we considered[4].

Target	$g(x)$	Estimand	Method
overall	1	ATE	IPW
treated	$e(x)$	ATT	IPWT
control	$1 - e(x)$	ATC	IPWC
restricted	$\mathbf{1}\{\alpha \le e(x) \le 1 - \alpha\}$	ATE	IPW
overlap	$e(x)(1 - e(x))$	ATO	OW
overlap	$\min\{e(x), 1 - e(x)\}$	ATM	MW
overlap	$-[e(x)\ln(e(x)) + (1 - e(x))\ln(1 - e(x))]$	ATEN	EW

IPW (resp. OW, MW, EW): inverse probability (resp. overlap, matching, entropy); we choose $\alpha = 0.05, 0.1$ and 0.15

Table 1: Choices of g, corresponding target population and causal estimands

The *balancing weights* (w_0, w_1) is given by $(w_0(x), w_1(x)) \propto \left(\frac{g(x)}{1 - e(x)}, \frac{g(x)}{e(x)}\right)$, which balanced the distributions of the covariates of the two treatment groups.

Given data $\{(X_i, Y_i, Z_i), i = 1, \dots, N\}$. τ_g can be estimated by the *Hájek-type estimator*

$$\widehat{\tau}_g^{\mathrm{H}} = \frac{\sum_{i=1}^N Z_i \widehat{w}_1(X_i) Y_i}{\sum_{i=1}^N Z_i \widehat{w}_1(X_i)} - \frac{\sum_{i=1}^N (1 - Z_i) \widehat{w}_0(X_i) Y_i}{\sum_{i=1}^N (1 - Z_i) \widehat{w}_0(X_i)}$$

where $\widehat{w}_z(x),\ z = 0, 1$ is calculated by a propensity score (PS) model.

The *augmented estimator* of τ_g is given by

$$\widehat{\tau}_g^{\mathrm{aug}} = \widehat{\tau}_g^{\mathrm{H}} + \left[\frac{\sum_{i=1}^N g(X_i)\{\widehat{m}_1(X_i) - \widehat{m}_0(X_i)\}}{\sum_{i=1}^N g(X_i)} - \frac{\sum_{i=1}^N Z_i \widehat{w}_1(X_i)\widehat{m}_1(X_i)}{\sum_{i=1}^N Z_i \widehat{w}_1(X_i)} - \frac{\sum_{i=1}^N (1 - Z_i)\widehat{w}_0(X_i)\widehat{m}_0(X_i)}{\sum_{i=1}^N (1 - Z_i)\widehat{w}_0(X_i)}\right]$$

where $m_z(X) = E(Y(z)|X),\ z = 0, 1$, is an outcome regression (OR) model. The augmented estimator includes Hájek-type estimator as the first part, and has an additional "augmented part" by the outcome model.

As for variance estimation, we use the close-form sandwich variance estimator. The score equation for this estimator is given by

$$\sum_{i=1}^N \Psi_\theta(X_i, Z_i, Y_i) = \sum_{i=1}^N \begin{bmatrix} \psi_\beta(X_i, Z_i) \\ Z_i \psi_{\alpha_1}(X_i, Y_i) \\ (1 - Z_i)\psi_{\alpha_0}(X_i, Y_i) \\ g(X_i)\{m_1(X_i) - \tau_{1g}^m\} \\ g(X_i)\{m_0(X_i) - \tau_{0g}^m\} \\ Z_i w_1(X_i)(Y_i - m_1(X_i) - \mu_{1g}) \\ (1 - Z_i) w_0(X_i)(Y_i - m_0(X_i) - \mu_{0g}) \end{bmatrix} = 0$$

where $\theta = (\beta', \alpha_1', \alpha_0', \tau_{1g}^m, \tau_{0g}^m, \mu_{1h}, \mu_{0g})'$.

For ATE, ATT and ATC, we actually use their *double-robust (DR) estimator* and corresponding sandwich variance estimators. Augmented estimators of overlap estimands are not DR.

Our hunch

First, clearly ATE $= p$ATT + $(1 - p)$ATC. Second, the balancing weights of overlap weights (OW) are $(w_0, w_1) \propto (e(x), 1 - e(x))$, so

- when $e(x) \approx 0.5$, $(e(x), 1 - e(x)) \approx \left(\frac{0.25}{1 - e(x)}, \frac{0.25}{e(x)}\right)$ (IPW/ATE weights)
- when $e(x)$ is small, $(e(x), 1 - e(x)) \approx \left(\frac{e(x)}{1 - e(x)}, 1\right)$ (ATT weights)
- when $e(x)$ is large, $(e(x), 1 - e(x)) \approx \left(1, \frac{1 - e(x)}{e(x)}\right)$ (ATC weights)

Since $p = P(Z = 1) = E[e(X)]$, we conjecture that under some conditions, the first moment of the propensity score might be sufficient to reflect how overlap estimands weight ATT and ATC, in a similar way as how $e(x)$ does. Also, overlap (OW), matching (MW) and entropy weights (EW) are similar, so they should weight similarly.

Simulation

Propensity score analysis

- 7 covariates: $X_1 \sim X_7$ (same as Li and Li[3]);
- Treatment Z is generated via logistic regression by $X_1 \sim X_7$;
- Outcome model (linear regression): $Y(0) = 0.5 + X_1 + 0.6X_2 + 2.2X_3 - 1.2X_4 + (X_1 + X_2)^2 + \varepsilon$ and $Y(1) = Y(0) + \delta(X)$, with $\varepsilon \sim N(0, 4)$, $\delta(X) = 4 + 3(X_1 + X_2)^2 + X_1 X_3$ (treatment effect);
- $M = 2000$ replications, sample size $N = 1000$ for each replication.

We generate 3 PS models with $p = 20.77\%, 49.72\%$ and 79.59% respectively. The ratio of variances of propensity score in treatment group to control group (r) of these three population are all in $[0.5, 2]$, which is considered as "equal variance" case[5]. Figure 2 gives the estimated propensity score distribution of the 3 models.

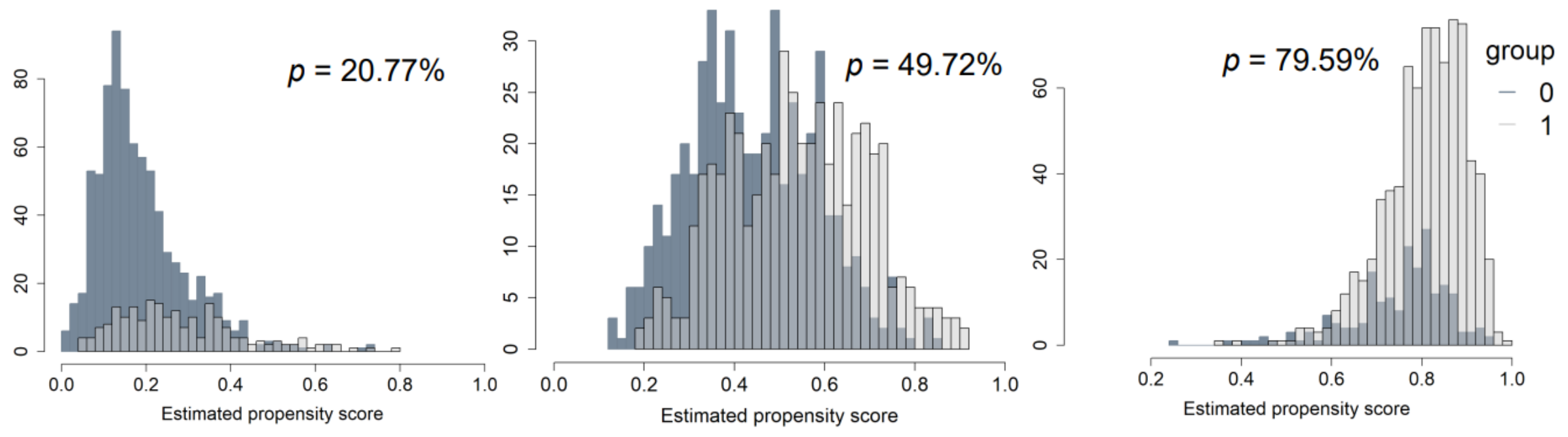
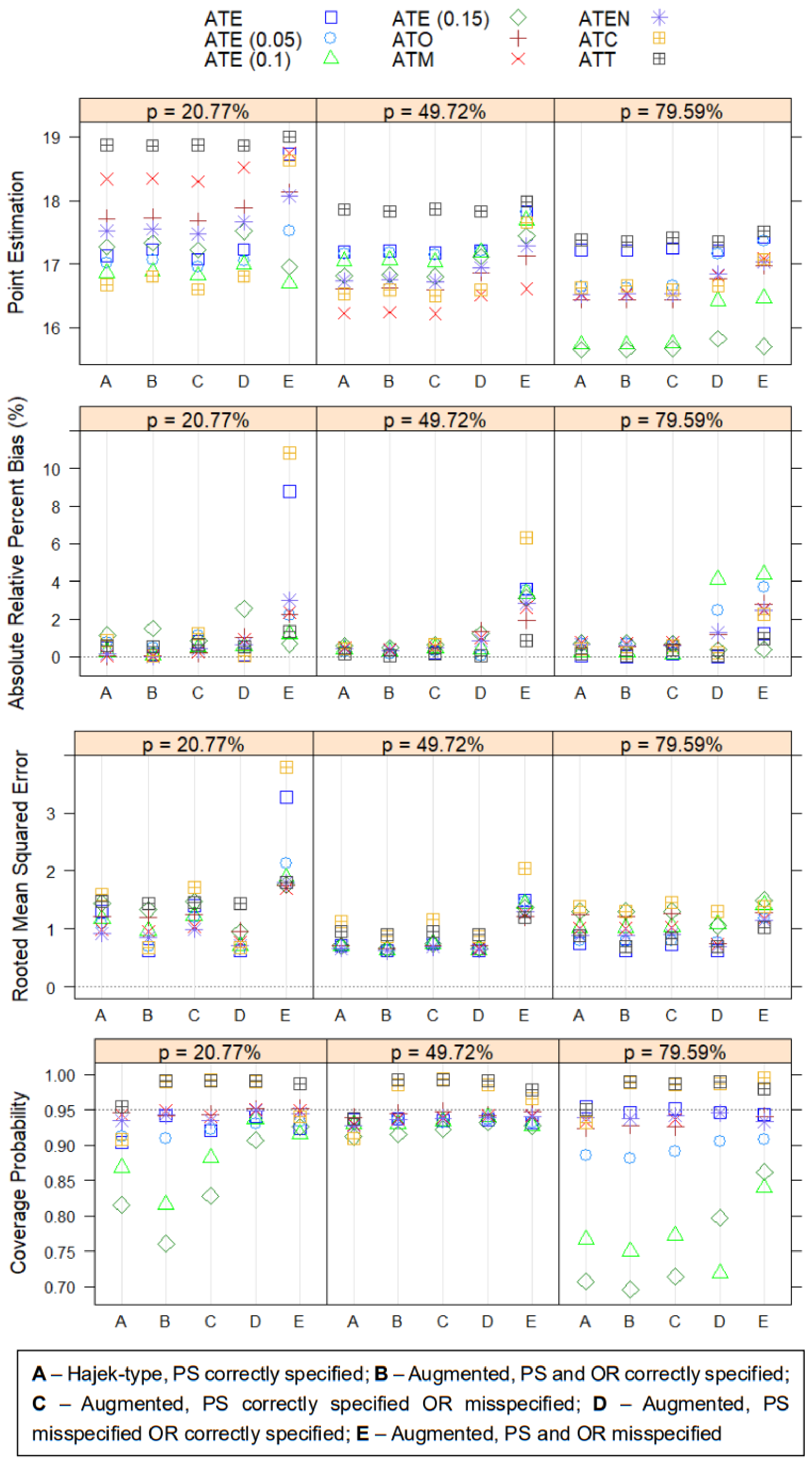


Figure 1: Estimated propensity scores distributions of models in simulation

Important results



A – Hajek-type, PS correctly specified; **B** – Augmented, PS and OR correctly specified; **C** – Augmented, PS correctly specified OR misspecified; **D** – Augmented, PS misspecified OR correctly specified; **E** – Augmented, PS and OR misspecified

Figure 2: Estimated propensity scores distributions of models in simulation

- When variances of propensity scores in treatment and control groups are considered as equal ($r \in [0.5, 2]$): (1) when p is high, overlap estimators (ATO, ATM and ATEN) weight toward (which does not mean exactly get very close to) ATC, and vice versa; (2) when $p \approx 0.5$ and no extreme weights exist, IPW and overlap estimators are similar;
- We have more results in our incoming paper when variances of propensity scores are not equal, which indeed deviate from our expectations above about the impact of p, so r also plays a role in these relationships;
- Augmented overlap estimators are more robust to model misspecifications, the differences among them are slight, and their sandwich variance estimations have better coverage rates of constructing confidence intervals. In addition, the Hájek-type estimator of overlap estimands has been proved more robust than that of ATE[6], so we did not investigate it here.

Data Example

We analyzed a right heart catheterization (RHC) data (https://hbiostat.org/data/). We investigate the effectiveness of the RHC diagnostic procedure (treatment, $Z = 1$) during the initial care of hospitalized, critically ill patients. We have in total 5735 subjects enrolled in the study, where **2184 (38%)** of them received the RHC treatment. The outcome is their log-length of stay in the intensive care unit (ICU) during the first 24 hours. 72 covariates are considered, such as age, race and some medical indices, in both PS and OR models. Table 2 gives the analysis results by different estimations.

Ratio of variances $r = 1.16$							
		Hájek-type Estimator			**Augmented Estimator**		
Estimand	Prop.	**Est.**	SE	p-value	Est.	SE	p-value
ATE	100%	**0.130**	0.032	<0.001	**0.129**	0.033	<0.001
ATE (0.05)	93%	0.099	0.030	<0.001	0.102	0.030	<0.001
ATE (0.1)	82%	0.102	0.029	<0.001	0.100	0.029	<0.001
ATE (0.15)	73%	0.078	0.029	0.008	0.079	0.029	0.007
ATO	100%	**0.095**	0.028	<0.001	**0.098**	0.028	<0.001
ATM	100%	**0.094**	0.028	<0.001	**0.095**	0.028	<0.001
ATEN	100%	**0.100**	0.028	<0.001	**0.102**	0.028	<0.001
ATC	100%	**0.156**	0.035	<0.001	**0.148**	0.037	<0.001
ATT	100%	**0.090**	0.046	0.049	**0.099**	0.043	0.021

Prop.: proportion of sample used; Est.: point estimation; SE: standard error

Table 2: RHC data analysis results

This is the case of small p ($< 40\%$) and equal variances ($r \in [0.5, 2]$), and the estimated ATE is closer to the estimated ATC, overlap estimates are closer to the estimated ATT, which is consistent to the simulation results.

Concluding Remarks

This work shows why sometimes ATE fails to identify the logical treatment effect, and provided some facts to researchers when they make analytical choices according to what they want to accomplish. p and r jointly decide the weighting direction of overlap (ATO, ATM and ATEN) estimators to ATT and ATC based on our simulations. The RHC data analysis further confirms this finding. At the same time, overlap (augmented) estimators are generally more robust to model misspecifications, and have better coverage probabilities using the close-form sandwich variance estimators.

References

[1] David Hajage, Florence Tubach, Philippe Gabriel Steg, Deepak L Bhatt, and Yann De Rycke. On the use of propensity scores in case of rare exposure. *BMC medical research methodology*, 16(1):38, 2016.

[2] Fan Li, Kari Lock Morgan, and Alan M Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

[3] Yan Li and Liang Li. Propensity score analysis methods with balancing constraints: A monte carlo study. *Statistical Methods in Medical Research*, 30(4):1119–1142, 2021.

[4] Roland A Matsouaka and Yunji Zhou. A framework for causal inference in the presence of extreme inverse probability weights: the role of overlap weights. *arXiv preprint arXiv:2011.01388*, 2020.

[5] Donald B Rubin. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3):169–188, 2001.

[6] Yunji Zhou, Roland A Matsouaka, and Laine Thomas. Propensity score weighting under limited overlap and model misspecification. *Statistical Methods in Medical Research*, 29(12):3721–3756, 2020.

Acknowledgements

The author is grateful to Roland Matsouaka (Duke Biostatistics & Bioinformatics) for his meticulous guidance to this work, and to Fan Li (Duke Statistical Science), Hwanhee Hong (Duke Biostatistics & Bioinformatics) and Yunji Zhou (Duke Global Health Insititute) for some insightful suggestions and helpful discussions.