STAT6061/STAT5008 – Causal Inference

Part 3-3. Propensity Score Methods

An-Shun Tai

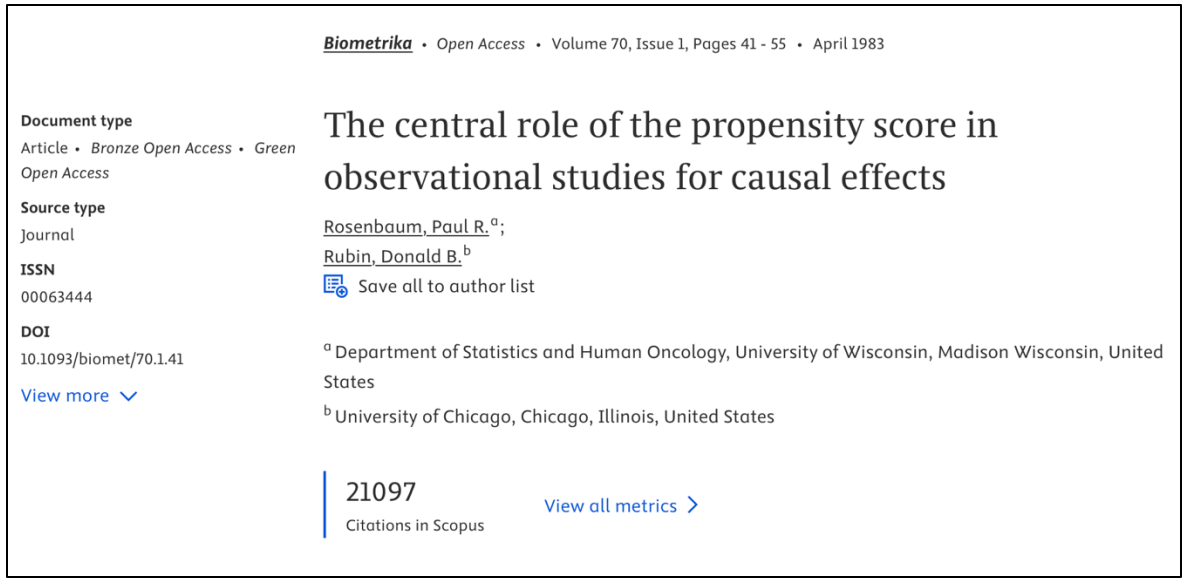
[1]*Department of Statistics*
National Cheng Kung University

[2]*Institute of Statistics and Data Science*
National Tsing Hua University

Propensity score: History

- The propensity score was introduced by Rosenbaum and Rubin in 1983, and their work has since become one of the most frequently cited papers in statistics.

- The propensity score is used to reduce bias in observational studies by balancing covariates between treatment and control groups.



Biometrika • *Open Access* • Volume 70, Issue 1, Pages 41 - 55 • April 1983

Document type
Article • *Bronze Open Access* • *Green Open Access*

Source type
Journal

ISSN
00063444

DOI
10.1093/biomet/70.1.41

View more

The central role of the propensity score in observational studies for causal effects

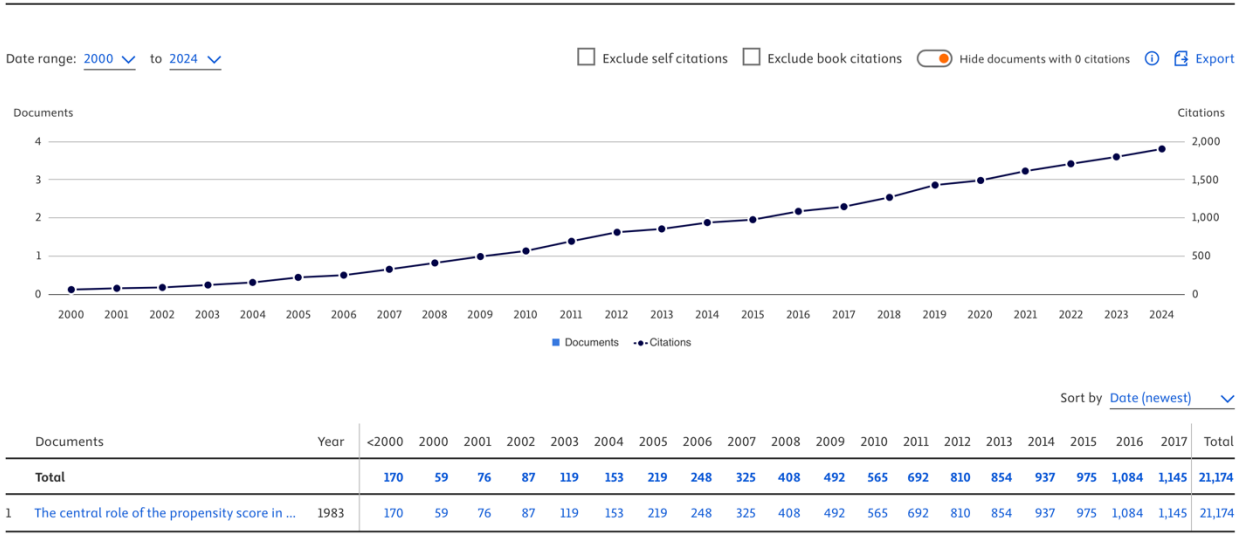
Rosenbaum, Paul R.[a];
Rubin, Donald B.[b]
Save all to author list

[a] Department of Statistics and Human Oncology, University of Wisconsin, Madison Wisconsin, United States

[b] University of Chicago, Chicago, Illinois, United States

21097
Citations in Scopus

View all metrics



Date range: 2000 to 2024 | Exclude self citations | Exclude book citations | Hide documents with 0 citations | Export

Documents | Citations

Documents | Citations

Sort by Date (newest)

	Documents	Year	<2000	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	Total
	Total		**170**	**59**	**76**	**87**	**119**	**153**	**219**	**248**	**325**	**408**	**492**	**565**	**692**	**810**	**854**	**937**	**975**	**1,084**	**1,145**	**21,174**
1	The central role of the propensity score in ...	1983	170	59	76	87	119	153	219	248	325	408	492	565	692	810	854	937	975	1,084	1,145	21,174

The 1st theorem for the propensity score: Unconfoundedness

(Rosenbaum and Rubin, 1983)

Theorem 3.1 (Unconfoundedness given the propensity score)

If the conditional exchangeability hold, that is,

$$A \perp \{Y(1), Y(0)\}|X,$$

then

$$A \perp \{Y(1), Y(0)\}|e(X),$$

where $e(X) = \Pr(A = 1|X)$.

- Conditioning on the propensity score alone is sufficient to remove confounding bias induced by observed covariates.
- Given this theorem, the propensity score serves as a tool for dimensionality reduction in confounding adjustment.
- We aim to show that $\Pr(A = 1|Y(1), Y(0), e(X)) = \Pr(A = 1|e(X))$ by applying the law of total expectation.

(**Hint:** To establish this result, demonstrate that both conditional probabilities are equal to $e(X)$)

The 2nd theorem for the propensity score: checking balance

(Rosenbaum and Rubin, 1983)

- Matching methods often target balance in the propensity score rather than directly in the covariates. Why is this approach commonly used?

Theorem 3.2 (The propensity score as a balancing score)

The propensity score satisfies

$$A \perp X | e(X).$$

Moreover, for any function $h(\cdot)$*, we have*

$$\mathbb{E}\left(\frac{A}{e(X)}h(X)\right) = \mathbb{E}\left(\frac{(1-A)}{1-e(X)}h(X)\right).$$

- The first result implies that if X distributions differ between treated and control groups, the distributions of their propensity scores $e(X)$ must also differ.

- The second result states that the any function $h(X)$ of the covariates has the same mean across the treatment and control groups, if weighted by the inverse of the propensity score.

Propensity score stratification

- **Theorem 3.1** motivates a simple method for estimating causal effects: *propensity score stratification*.

- Rosenbaum and Rubin (1983) recommend stratifying on the quintiles of the propensity score and computing the treatment effect within each quintile.

  1. Discretize the estimated propensity score by its K quantile, denoted by $\hat{e}'(X)$.
  2. Approximate exchangeability within strata:
  $$A \perp \{Y(1), Y(0)\} | e(X) = e_k \ (k = 1,2, \dots, K)$$
  3. Analyze the observational data in the same way as the SRE stratified on $\hat{e}'(X)$.

- How to choose K?

- If K is too small, exchangeability may not hold within strata; if K is too large, some strata may lack overlap between treated and control units. Therefore, there is a bias–variance trade-off in choosing.

- Based on Cochran (1968) and Rosenbaum & Rubin (1983b, 1984): $K = 5$ is often effective in reducing bias across many settings.

Propensity score weighting

Theorem 3.3

If the conditional exchangeability $(A \perp \{Y(1), Y(0)\}|X)$ *and positivity* $(0 < e(X) < 1)$ *hold, then*

$$\mathbb{E}\big(Y(1)\big) = \mathbb{E}\left(\frac{A}{e(X)}Y\right), \mathbb{E}\big(Y(0)\big) = \mathbb{E}\left(\frac{(1-A)}{1-e(X)}Y\right).$$

and

$$\tau = \mathbb{E}\big(Y(1) - Y(0)\big) = E\left(\frac{A}{e(X)}Y - \frac{1-A}{1-e(X)}Y\right).$$

- Theorem 3.3 provides the foundation for inverse probability weighting (IPW), also known as inverse probability of treatment weighting (IPTW).

Propensity score weighting: Horvitz-Thompson estimator

- Theorem 3.3 yields the Horvitz–Thompson (HT) estimator, also known as the inverse probability weighting (IPW) estimator.

$$\hat{\tau}_1^{HT} = \frac{1}{N}\left[\sum_{i=1}^{N}\frac{A_i}{e(X_i)}Y_i - \sum_{i=1}^{N}\frac{1-A_i}{1-e(X_i)}Y_i\right]$$

$$= \frac{1}{N}\left[\sum_{i=1}^{N}w_1(X_i)A_iY_i - \sum_{i=1}^{N}w_0(X_i)(1-A_i)Y_i\right]$$

where $w_1(X_i) = 1/e(X_i)$ and $w_0(X_i) = 1/[1-e(X_i)]$.

- Given a known propensity score, $\hat{\tau}_1^{HT}$ is a nonparametric unbiased estimator of ATE.

- When $e(X_i)$ is unknown, we substitute it with the estimated propensity score to obtain:

$$\hat{\tau}_2^{HT} = \frac{1}{N}\left[\sum_{i=1}^{N}\frac{A_i}{\hat{e}(X_i)}Y_i - \sum_{i=1}^{N}\frac{1-A_i}{1-\hat{e}(X_i)}Y_i\right]$$

- Under standard regularity conditions, $\hat{\tau}_2^{HT}$ is a consistently and asymptotic normal (CAN) estimator when the model of the propensity score is correctly specified.

Propensity score weighting: Hájek estimator

- Note that the weights in the HT estimator do not sum to one, which leads to a key issue**: its lack of invariance.**

- Hájek estimator (1971) with normalized weights, also known as the stabilized IPW estimator:

$$\hat{\tau}^H = \frac{\sum_{i=1}^N w_1(X_i) A_i Y_i}{\sum_{i=1}^N w_1(X_i) A_i} - \frac{\sum_{i=1}^N w_0(X_i)(1-A_i) Y_i}{\sum_{i=1}^N w_0(X_i)\,(1-A_i)}$$

- Why consider the Hájek estimator?

- Practical perspective

1. Normalizing prevents a few large weights from dominating the estimate.
2. Results are easier to interpret as weighted averages.

- Statistical perspective

The Hájek estimator trades a small amount of bias for greater stability and reduced variance in finite samples.

$$Var(\hat{\tau}_1^{HT}) > Var(\hat{\tau}^H)$$

Propensity score weighting: Different weights

(Li, Morgan, and Zaslavsky, 2018)

- Generalization of the HT and Hájek estimators:

$$\frac{\sum_{i=1}^{N} \mathcal{W}_1(X_i) A_i Y_i}{\sum_{i=1}^{N} \mathcal{W}_1(X_i) A_i} - \frac{\sum_{i=1}^{N} \mathcal{W}_0(X_i)(1-A_i) Y_i}{\sum_{i=1}^{N} \mathcal{W}_0(X_i)\,(1-A_i)}$$

where $\mathcal{W}_1(X_i) = h(X_i)/e(X_i)$ and $\mathcal{W}_0(X_i) = h(X_i)/[1-e(X_i)]$.

- Specification o $h(x)$ defines the target population and causal estimands and determines the weights.

Target population	$h(x)$	Causal Estimand	Weights $\left(\mathcal{W}_1(X_i), \mathcal{W}_0(X_i)\right)$
Combined	1	ATE	$(1/e(x), 1/[1-e(x)])$
Treated	$e(x)$	ATT	$(1, e(x)/[1-e(x)])$
Control	$1-e(x)$	ATC	$([1-e(x)]/e(x), 1)$
Overlap	$e(x)[1-e(x)]$	ATO	$(1-e(x), e(x))$
Matching*	$\min\{e(x), 1-e(x)\}$		$\left(\frac{\min\{e(x),1-e(x)\}}{e(x)}, \frac{\min\{e(x),1-e(x)\}}{[1-e(x)]}\right)$

Remark 1: The overlap weights

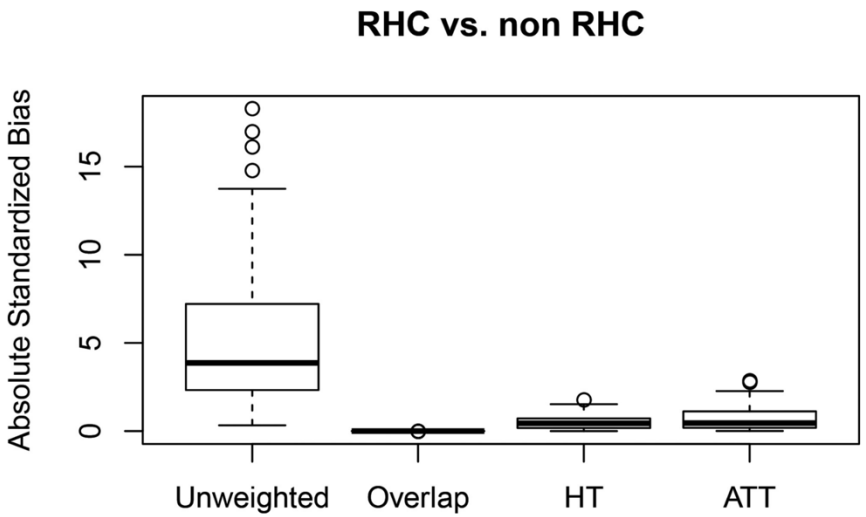
- The estimand corresponds to **the average treatment effect in the overlap population** (ATO) — a substantively meaningful target group.

- Focuses estimation on units with similar propensity scores across treatment groups (i.e., $e(x) = 1/2$) — those most comparable.

- Statistical properties

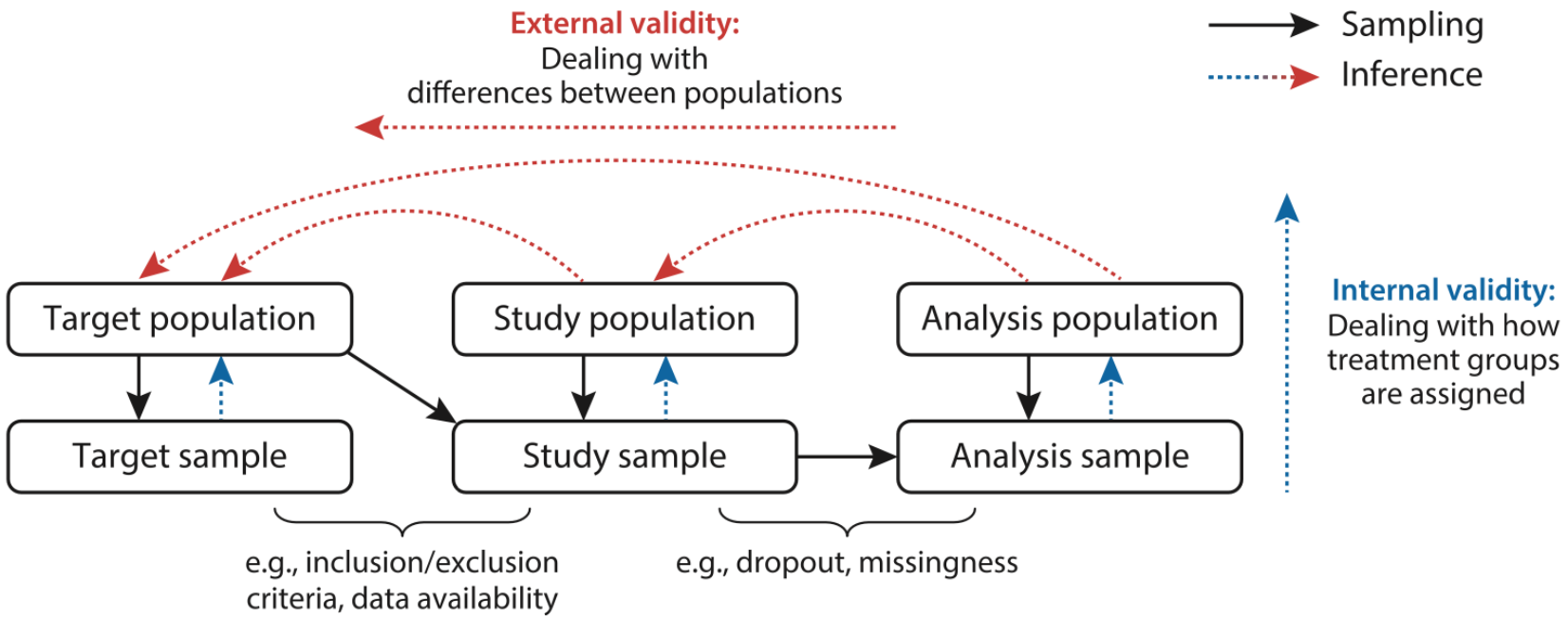
1. Avoids **extreme weights** by down-weighting units with propensity scores near 0 or 1.
2. Overlap weights automatically **achieve exact mean balance** on all covariates included in the propensity score model.

- ✓ Assessed the effect of right heart catheterization (RHC) on 30-day survival using observational data from five U.S. hospitals.
- ✓ Patients were classified by RHC use within 24 hours (treated: 2,184; control: 3,551).



Remark 2: Generalizability and Transportability

(Degtiar and Rose, 2023)



- **Generalizability**

Can the treatment effect estimated in the study sample be extended to the same population from which the sample was drawn?

- **Transportability**

Can the treatment effect be applied to a different target population?

Propensity score in regressions: As a covariate

- Theorem 3.1 reveals that if exchangeability holds conditioning on X, then it also holds conditional on e(X):

$$A \perp \{Y(1), Y(0)\} | e(X).$$

- Thus, the ATE is also nonparametrically identified by

$$\mathbb{E}\big(\mathbb{E}(Y|A=1, e(X))\big) - \mathbb{E}\big(\mathbb{E}(Y|A=0, e(X))\big)$$

$$= \int \mathbb{E}(Y|A=1, e(X)) \Pr(e(X)=e')de' - \int \mathbb{E}(Y|A=0, e(X)) \Pr(e(X)=e')de'$$

- This implies that in the outcome regression method discussed in Part 3.1, one can adjust for $e(X)$ instead of the full covariate vector X.

- The OLS estimator (from a population view)

$$\underset{\beta_0, \beta_a, \beta_e}{\arg\min}\, \mathbb{E}\{Y - (\beta_0 + \beta_a A + \beta_e e(X))\}^2$$

Propensity score in regressions: As a covariate (cont.)

- When the propensity score model is correctly specified and the outcome is linear in A and $e(X)$ (implying a homogenous ATE), the OLS estimator of β_e is consistent for ATE.

- In general, β_e corresponds to the average treatment effect for the overlap population (ATO), defined as

$$\frac{\mathbb{E}\big(h_o(X)\mathbb{E}(Y(0)-Y(1)|X)\big)}{\mathbb{E}\big(h_o(X)\big)}$$

where $h_o(X) = e(X)[1-e(X)]$.

Propensity score in regressions: Weighted least squares

- Weighting constructs a pseudo-population that mimics a randomized experiment. What are the implications of applying regression to this pseudo-randomized setting?

- The weighted least squares of Y on $(1, A)$ yields the Hájek estimator $\hat{\tau}^H$.

Proposition 3.1

The Hájek estimator $\hat{\tau}^H$ equals $\hat{\beta}$ from the following WLS

$$(\hat{\beta}_0, \hat{\beta}) = \arg\min_{\beta_0, \beta} \sum w_i \big(Y_i - (\beta_0 + \beta A_i)\big)^2$$

with weights

$$w_i = \frac{A_i}{\hat{e}(X_i)} + \frac{1 - A_i}{1 - \hat{e}(X_i)}.$$

- Proposition 3.1 establishes an analogue of a key property of regression in the context of a CRE: even without covariate adjustment, a simple linear regression yields an unbiased estimator, regardless of the true outcome model specification.

Propensity score in regressions: Weighted least squares (cont.)

- In a CRE, it has been noted that including covariates and their full interactions with A can improve efficiency.

- What happens when we include covariates and their full interactions with treatment in WLS?

⇒ This WLS estimator has the desirable property of double robustness, meaning it remains consistent if either the outcome model or the propensity score model is correctly specified (but not necessarily both). More details coming up next.

- Simulation from Kang and Schafer (2007)

Sample size	y-model	Method	Bias	% Bias	RMSE	MAE
(a) $n = 200$	Correct	*OLS*	−0.08	−3.4	2.48	1.68
	Incorrect	*OLS*	−0.57	−17.7	3.26	2.24
(b) $n = 1000$	Correct	*OLS*	−0.00	−0.1	1.17	0.79
	Incorrect	*OLS*	−0.84	−56.0	1.72	1.15

Sample size	π-model	Method	Bias	% Bias	RMSE	MAE
(a) $n = 200$	Correct	*strat-π*	−1.15	−38.1	3.22	2.17
	Incorrect	*strat-π*	−2.82	−87.7	4.28	3.13
(b) $n = 1000$	Correct	*strat-π*	−1.08	−81.5	1.71	1.18
	Incorrect	*strat-π*	−2.87	−202.7	3.19	2.83

Sample size	π-model	y-model	Method	Bias	% Bias	RMSE	MAE
(a) $n = 200$	Correct	Correct	*WLS*	−0.09	−3.4	2.48	1.68
		Incorrect	*WLS*	0.38	13.2	2.88	1.92
	Incorrect	Correct	*WLS*	−0.08	−3.4	2.48	1.68
		Incorrect	*WLS*	−2.20	−70.0	3.83	2.74
(b) $n = 1000$	Correct	Correct	*WLS*	0.00	−0.1	1.17	0.78
		Incorrect	*WLS*	0.16	12.0	1.35	0.92
	Incorrect	Correct	*WLS*	0.00	−0.1	1.17	0.78
		Incorrect	*WLS*	−2.99	−203.6	3.33	2.98

Estimate the propensity score

Step 1. Model Treatment Assignment:

- For binary treatments, use logistic regression: $logit(\Pr(A = 1|X)) = \beta X$

Step 2. Estimate Propensity Scores:

- Fit the model (e.g., with stepwise covariate selection):

$$\hat{e}(X) = \frac{e^{\hat{\beta}X}}{1 + e^{\hat{\beta}X}}$$

Step 3. Check Overlap

- Examine score distributions across treatment groups.
- If there's substantial non-overlap, consider discarding extreme observations.

Step 4. Assess Covariate Balance

- Use matching, weighting, or stratification based on.
- Choice depends on the method applied in Step 2.

Step 5. Refine Model if Needed

- If covariates remain unbalanced, re-fit the model using Higher-order terms, Splines, or Covariate-treatment interactions
- Repeat steps 2–3 until balance is achieved.

References

Degtiar, I., & Rose, S. (2023). A review of generalizability and transportability. *Annual Review of Statistics and Its Application*, *10*(1), 501-524.

Kang, J. D., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data.

Li, F., Morgan, K. L., & Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, *113*(521), 390-400.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.