HSR Health Services Research

© Health Research and Educational Trust
DOI: 10.1111/1475-6773.12020
METHODS ARTICLE

Squeezing the Balloon: Propensity Scores and Unmeasured Covariate Balance

John M. Brooks and Robert L. Ohsfeldt

Objective. To assess the covariate balancing properties of propensity score-based algorithms in which covariates affecting treatment choice are both measured and unmeasured.

Data Sources/Study Setting. A simulation model of treatment choice and outcome.

Study Design. Simulation.

Data Collection/Extraction Methods. Eight simulation scenarios varied with the values placed on measured and unmeasured covariates and the strength of the relationships between the measured and unmeasured covariates. The balance of both measured and unmeasured covariates was compared across patients either grouped or reweighted by propensity scores methods.

Principal Findings. Propensity score algorithms require unmeasured covariate variation that is unrelated to measured covariates, and they exacerbate the imbalance in this variation between treated and untreated patients relative to the full unweighted sample.

Conclusions. The balance of measured covariates between treated and untreated patients has opposite implications for unmeasured covariates in randomized and observational studies. Measured covariate balance between treated and untreated patients in randomized studies reinforces the notion that all covariates are balanced. In contrast, forced balance of measured covariates using propensity score methods in observational studies exacerbates the imbalance in the independent portion of the variation in the unmeasured covariates, which can be likened to squeezing a balloon. If the unmeasured covariates affecting treatment choice are confounders, propensity score methods can exacerbate the bias in treatment effect estimates.

Key Words. Propensity scores, covariate balance, matching, binning, assumptions, simulation

The strength of randomized controlled trials (RCTs) is the assumption that randomized treatment assignment yields a balanced distribution of covariates thought to be related to outcome between the treatment and control groups

(Rubin 2001). Published studies of RCT results traditionally report a table displaying the balance in measured covariates (e.g., patient age, gender, baseline clinical conditions, etc.) between the treatment and control groups. Demonstrated balance of measured covariates across treatment groups is intended to lend credence that such balance extends to unmeasured covariates (Berk 2004). In the context of observational (nonrandomized) data, researchers have espoused designing treatment effect studies that mimic the measured covariate balancing properties of RCTs (Rosenbaum and Rubin 1983a,b; Rubin 1997, 2001, 2007; Joffe and Rosenbaum 1999; Shah et al. 2005). The use of a propensity score (PS)—the probability a patient received treatment given the patient's measured covariate values—has become a mainstay in efforts to find measured covariate balance in observational data studies to estimate treatment effects. It has been said that PS-based methods "can be used to design observational studies in a way analogous to the way randomized experiments are designed" (Rubin 2001) with a design attempting to "assemble groups of treated and control units such that within each group the distributions of covariates is balanced" (Rubin 2001). While methodologists are quick to qualify that achieving balance in measured covariates between groups of treated and untreated patients does not "guarantee" balance in unmeasured covariates across groups, measured covariate balance often creates an "expectation" of unmeasured covariate balance as in RCTs (Ward and Johnson 2008). Indeed, a review of the PS literature noted that "many of the articles in our review" imply that "propensity scores might also balance the unknown confounders between exposure groups" (Shah et al. 2005).

Several PS-based algorithms have been suggested to create patient samples that are balanced in measured covariates between treated and untreated patients. These algorithms range from stratification (D'Agostino 1998) and matching based on propensity scores (Hall, Summers, and Oberchain 2003; Frolich 2007; Stuart 2010) to using patient-specific propensity scores to weight observations (Rosenbaum 1987; Robins, Hernan, and Brunback 2000). Treatment effect inferences are then made by contrasting average outcomes between treated and untreated patients with similar propensity scores (and correspondingly similar distributions of measured covariates). These algorithms yield unbiased treatment effect estimates only if after balancing measured

Address correspondence to John M. Brooks, Ph.D., University of Iowa, College of Pharmacy and College of Public Health, s-515 Pharmacy Bldg. 115 S. Grand Ave, Iowa City, IA 52242; e-mail: john-brooks@uiowa.edu. Robert L. Ohsfeldt, Ph.D., is with the Texas A&M Health Science Center, School of Public Health, College Station, TX.

covariates, unmeasured covariates are “ignorable” or that the remaining unmeasured covariates that affected treatment choice are independent of outcome (Rosenbaum and Rubin 1983a,b; Joffe and Rosenbaum 1999). Unmeasured covariates affecting treatment choice are ignorable if either (1) they have no relationship (either directly or indirectly) with outcome, or (2) they are balanced between treatment and control groups after balancing measured covariates. Neither of these conditions can be verified directly with data available to researchers. The condition that the unmeasured covariates affecting treatment choice have no relationship with outcome is identical to the assumption required to yield unbiased estimates in standard multivariate regression-based treatment effect estimators—treatment is orthogonal to the error term in the outcome relationship after adjusting for the measured covariates included in the regression model (Angrist and Pischke 2009). Stated differently, this condition assumes that none of the unmeasured covariates affecting treatment choice *confound* the relationship between treatment and outcome. This orthogonal assumption requires theory-based persuasion by researchers for acceptance. Therefore, the conceptual advantage of PS-based methods relative to standard regression appears to hinge on the assumption that balancing measured covariates between treated and nontreated patients leads to unmeasured covariate balance between treated and nontreated patients. If this assumption holds, unbiased treatment effect estimates can be obtained without relying on theory to support the orthogonal assumption.

However, PS-based analyses of treatment effects using observational data largely ignore what seems to be a fundamental question—why did patients with the same or similar propensity scores receive different treatments? Intuitively, it would seem that unmeasured factors not accounted for in the PS model must be different between two patients with similar propensity scores for them to receive different treatments. Let patient utility associated with treatment U(T) and no treatment U (NT) be represented in terms of measured (X_{M}) and unmeasured covariates (X_{U}):

$$U(\mathrm{T}) = \alpha_1 \cdot X_{\mathrm{M}} + \alpha_2 \cdot X_{\mathrm{U}}; \quad (1)$$

$$U(\mathrm{NT}) = \beta_1 \cdot X_{\mathrm{M}} + \beta_2 \cdot X_{\mathrm{U}}. \quad (2)$$

The measured and unmeasured covariates in equations (1) and (2) represent any factors affecting the utility of treatment versus no treatment for the patient. These covariates could represent factors related to patient preferences over the outcome changes induced by treatment choice (e.g., an actor may value facial changes from cosmetic surgery more than a construction worker) or factors affecting the relative effectiveness of treatment (e.g., a child with an ear infection and a high fever will expect more benefit from an antibiotic than a child with an ear infection and a low fever). A patient will choose treatment if the net utility gain from treatment—NG(T)—is positive:

$$NG(\mathrm{T}) = U(\mathrm{T}) - U(\mathrm{NT}) = (\alpha_1 - \beta_1) \cdot X_{\mathrm{M}} + (\alpha_2 - \beta_2) \cdot X_{\mathrm{u}} > 0. \qquad (3)$$

Based on equation (3), patient treatment choices depend on their respective values of X_{M} and X_{U}. If $(\alpha_2 - \beta_2) > 0$ treated patients will tend to have higher average values of X_{U} than untreated patients, but with X_{M} also varying across patients it may be possible to find treated patients with low values of X_{U} and untreated patients with high values of X_{U}.

If, however, two patients A and B are matched to have identical values of the measured covariate—$\overline{X}_{\mathrm{M}}$—and patient A chooses treatment and patient B does not, it must be that:

$$\begin{aligned} NG(\mathrm{T})^{\mathrm{A}} &= (\alpha_1 - \beta_1)\overline{X}_{\mathrm{M}} + (\alpha_2 - \beta_2){X_{\mathrm{u}}}^{\mathrm{A}} > 0 \\ &> (\alpha_1 - \beta_1)\overline{X}_{\mathrm{M}} + (\alpha_2 - \beta_2){X_{\mathrm{U}}}^{\mathrm{B}} = NG(\mathrm{T})^{\mathrm{B}} \qquad (4) \end{aligned}$$

where for patient i $\mathrm{NG(T)}^i$ equals the net gain of treatment and X_{u}^i equals i's value of X_{U}. With a fixed value of X_{M}, for equation (4) to hold it *must be* that $X_{\mathrm{u}}^{\mathrm{A}} \neq X_{\mathrm{u}}^{\mathrm{B}}$. If $(\alpha_2 - \beta_2) > 0$ treated patients with matched X_{M} values must have higher values of X_{U} than untreated patients. Therefore, across a set of treated and nontreated patients matched on X_{M}, we would expect greater average differences in X_{U} than the average differences in X_{U} between the population of treated and nontreated patients not matched by X_{M}.

In this study, we demonstrate the covariate balancing properties of PS-based algorithms through the lens of a simple treatment choice simulation model in which covariates affecting treatment choice are both measured and unmeasured. Prior simulation-based research showed that imbalance in unmeasured covariates related to treatment assignment remains after using PS-based algorithms (Austin, Grootendorst, and Anderson 2007). Others have described the extent in which treatment effect estimates from propensity score-based approaches are sensitive to imbalance in unobserved covariates (Rosenbaum and Rubin 1983a,b; Lin, Psaty et al. 1998). However, it has not been shown how PS-based algorithms *affect* the balance of unmeasured covariates between treated and untreated patients.

In our simulations, we find properties that are problematic for researchers hoping to make treatment effect inferences relying *only* on the expectation that balancing measured covariate implies balanced unmeasured covariates. To yield treated and untreated patients with similar propensity scores, we find that PS algorithms *require* imbalance in the portion of the variation of the unmeasured covariates that affect treatment choice that is unrelated to the measured covariates. In addition, as compared with the full unweighted sample, PS algorithms *exacerbate* the imbalance in the portion of the unmeasured covariates unrelated to the measured covariates between treated and untreated patients. This result is directly counter to the assumption often relied on in applications of propensity score methods that balancing measured covariates implies balance in the unmeasured covariates that affected treatment choice (Shah et al. 2005).

METHODS

Simulation Model Structure

We modified a simple simulation model of treatment choice and outcome that was used in previous research (Brooks and Fang 2009). In this model, covariates affecting treatment choice are divided between those measured and unmeasured by a subsequent researcher. A propensity score is estimated for each simulated patient using the measured covariate. In simulations, the unmeasured covariates affecting treatment choice are distinguished by their assumed relationship with the model outcome. For simplicity, outcome in the model is defined as being "cured" from a given condition. Patients can either choose the "treatment" or the "alternative" (e.g., another treatment, watchful waiting). Patients can be cured using the alternative, but the treatment increases the cure probability relative to the alternative. Patients value being cured, but treatment is more costly relative to the alternative. Treatment is specified as a binary variable—T, where $T = 1$ if the patient chooses treatment, and 0 if the patient chooses the alternative. The probability of a cure, P (C), is specified in the following manner:

$$P(\mathrm{C}) = \beta_o + \beta_{\mathrm{T}} \cdot T + \beta_{\mathrm{M}} \cdot X_{\mathrm{M}} + \beta_{\mathrm{U1}} \cdot X_{\mathrm{U1}} + \beta_{\mathrm{U2}} \cdot X_{\mathrm{U2}} + \varepsilon. \qquad (5)$$

β_{T} equals the increase in the probability of cure relative to the alternative for a patient that chooses treatment. The vector $\boldsymbol{\beta}$ contains the parameters β_{T}, β_{M}, β_{U1}, and β_{U2} in equation (5) that equal the changes in the probability of cure related to T, X_{M}, X_{U1}, and X_{U2}, respectively. Covariates X_{M}, X_{U1}, X_{U2}, and ε have direct effects on P(C) and are distinguished by the assumption that X_{M} is measured and available for subsequent research, whereas X_{U1}, X_{U2}, and ε are not. Specifically, ε represents the accumulated other risk factors related to cure that are not related to treatment choice.

The model assumes that patients consult with their providers to gain knowledge of equation (5) and form a treatment valuation relative to the alternative—T^*—that is based on the value patients associate with a cure, the cost of treatment, and the effect on treatment valuation from patient-specific covariates:

$$T^* = (\mathrm{V} \cdot \beta_{\mathrm{T}}) - \mathrm{S} + \alpha_{\mathrm{M}} \cdot X_{\mathrm{M}} + \alpha_{\mathrm{U1}} \cdot X_{\mathrm{U1}} + \alpha_{\mathrm{U3}} \cdot X_{\mathrm{U3}} + \alpha_{\mathrm{U4}} \cdot X_{\mathrm{U4}}, \qquad (6)$$

where V is the value patients place on being cured; S equals the incremental costs associated with the treatment relative to the alternative; X_{M} and X_{U1} are defined as in equation (5), and each was specified to have a direct effect on treatment valuation. X_{U3} and X_{U4} are additional unmeasured covariates affecting treatment valuation. X_{U3} is specified to have an indirect effect on cure through a correlation with the covariate X_{U2} that is specified in equation (5):

$$\mathrm{Corr}(X_{\mathrm{U2}}, X_{\mathrm{U3}}) \neq 0. \qquad (7)$$

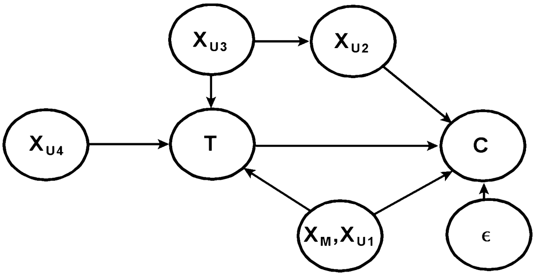
An intuitive example of an X_{U2}, X_{U3} combination could be X_{U3} (patient health preferences) and X_{U2} (healthy diet). A patient with higher health preferences may place greater value on treatment ($\alpha_{\mathrm{U3}} > 0$) and patient health preferences and healthy diet are likely positively correlated. X_{U4} affects treatment valuation but has no direct or indirect effect on cure and, given these properties, would be an instrumental variable if it was measured. The vector $\boldsymbol{\alpha}$ contains the covariate value weights α_{M}, α_{U1}, α_{U3}, and α_{U4} in equation (6) that represent changes in treatment valuation from changes in X_{M}, X_{U1}, X_{U3}, and X_{U4}, respectively.

Following standard discrete choice theory (Ben-Akiva and Lerman 1985), the patient chooses treatment (T) if T^* is greater than zero, or:

$$T = 1 \text{ if } (T^* = \mathrm{f}(\mathrm{V}, \mathrm{S}, X_{\mathrm{M}}, X_{\mathrm{U1}}, X_{\mathrm{U3}}, X_{\mathrm{U4}}; \alpha, \beta) > 0), 0 \text{ otherwise} \qquad (8)$$

Figure 1 contains a directed acyclic graph summarizing the model relationships.

Figure 1: Directed Acyclic Graph of Simulation Model Relationships



X_{U1} and X_{U2} represent unmeasured confounding variables in the estimation of the relationship between the likelihood of a cure C and T when only adjusting for X_M.

$$P(\mathrm{C}) = \beta_0 + \beta_T \cdot T + \beta_M \cdot X_M + \text{error}, \tag{9}$$

where error = ($\beta_{U1} \bullet X_{U1} + \beta_{U2} \bullet X_{U2} + \varepsilon$). Because X_{U1} and X_{U2} are in the error term of this relationship, T would not be orthogonal to the error term as X_{U1} is directly related to T and X_{U2} is indirectly related to T via its correlation with X_{U3}. As the orthogonal assumption is not valid in this model, to make inferences on β_T from equation (5) using PS algorithms, it must be assumed that balancing X_M between treated and untreated patients yields balance in X_{U1} and X_{U3} (and indirectly X_{U2}) between treated and untreated patients.

We found little discussion in the literature describing how the characteristics of the relationships among the covariates and outcomes affect the ability to balance unmeasured covariates between treated and untreated patients using PS algorithms. Here, we theorized that unmeasured covariate balance will be affected by the relative covariate value weights (α) placed on measured and unmeasured covariates in the treatment valuation equation (6). We suspected that smaller value weights on the unmeasured covariates (α_{U1}, α_{U3}, α_{U4}) relative to the measured covariate (α_M) will require greater differences in the actual unmeasured covariate values to yield treated and untreated patients with matched measured covariates. In addition, it has been suggested that relationships among the measured and unmeasured covariates may enable the measured covariates to serve as proxies for the unmeasured covariates and lead to balance in the unmeasured covariates (Schneeweiss et al. 2009; Stuart 2010). To evaluate this suggestion, the simulation model was constructed to specify relationships between the X_M and the unmeasured covariates X_{U1}, X_{U3}, and X_{U4}.

We suspect that the stronger the relationships between the measured and unmeasured covariates, the less independent influence unmeasured covariates will have on treatment choice, and *the more difficult* it will be to find matched treated and untreated patients based on propensity scores generated by the measured covariate. As a result, we expected that as the strength of the relationship between the measured and unmeasured covariates increases, we would find greater imbalance between matched treated and untreated patients in variation in each unmeasured covariate that is unrelated to the measured covariate.

Simulation Approach

Eight simulation scenarios were specified that varied with the value weights ($\boldsymbol{\alpha}$) placed on the measured and unmeasured covariates in equation (6) and the level of relationship between the measured covariate and the unmeasured covariates. In each scenario, 100,000 simulated patient observations were generated and the values of the parameters V, S, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ were identical for all simulated patients. Table 1 contains the parameter values used in the eight simulation scenarios. Scenarios 1, 2, and 3 were constructed with no relationships among the measured and unmeasured covariates in equation (6), and the simulations varied only by the value weights ($\boldsymbol{\alpha}$) assigned to each covariate. Value weights were assigned to the covariates so that relative value of the unmeasured covariates fell relative to the measured covariate moving from scenarios 1 through 3 (α_{U1}, α_{U3}, and α_{U4} equaled 200, 100, and 50 in scenarios 1, 2, and 3, respectively, whereas α_{M} equaled 100 across all three scenarios). To better contrast the differences in the covariate balancing properties of propensity score methods across scenarios 1, 2, and 3, the treatment cost parameter (S) was adjusted to ensure that the expected treatment valuation (T^{*}) in each scenario equaled zero so approximately 50 percent of the patients in each scenario chose treatment. In these scenarios, the values of covariates themselves—X_{M}, X_{U1}, X_{U3}, and X_{U4}—for each simulated patient were randomly

Table 1: Model Parameters across Simulation Scenarios

	Scenarios							
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
V	1,500	1,500	1,500	1,500	1,500	1,500	1,500	1,500
β_{T}	.2	.2	.2	.2	.2	.2	.2	.2
S	650	500	425	500	500	500	500	500
α_{M}	100	100	100	100	100	100	100	100
α_{U1}	200	100	50	100	100	100	100	100
α_{U3}	200	100	50	100	100	100	100	100
α_{U4}	200	100	50	100	100	100	100	100
Weight*	0	0	0	.1	.3	.5	.7	.9
$(R^2)^{\dagger}$	.042	.168	.434	.289	.559	.691	.735	.750

*The value used in equation (11) relating unmeasured covariate *i* to the measured confounder via:

$$X_{\mathrm{Ui}} = (\mathrm{weight})^{*}X_{\mathrm{M}} + (1 - \mathrm{weight})^{*}E_{\mathrm{i}},$$

where E_i is a uniform random variable (0,1).

†Percentage of variation *T* explained by X_{M} in the propensity score equation (12).

sampled from independent uniform distributions between (0, 1) using the RANUNI function within the SAS IML procedure. This approach insured the independence of each covariate across patients. X_{U2} was then estimated to ensure a correlation with X_{U3} using the following linear relationship:

$$X_{\mathrm{U2}} = (.5)^{*} X_{\mathrm{U3}} + (.5)^{*} D, \tag{10}$$

where D is an independent uniform random variable (0,1).

In scenarios 4 through 8, the covariate value weights (α) were specified as in scenario 2, but the covariate values of X_{U1}, X_{U3}, and X_{U4} for each simulated patient were estimated with relationships with X_{M} based on the following linear models:

$$X_{\mathrm{Ui}} = (\mathrm{weight})^{*} X_{\mathrm{M}} + (1 - \mathrm{weight})^{*} E_{\mathrm{i}}, \tag{11}$$

where each E_{i} was an independent uniform random variable (0,1). In scenarios 2, 4, 5, 6, 7, and 8, "weight" was set to 0, .1, .3, .5, .7, and .9, respectively, so that the strength of the relationship between the measured and unmeasured covariates increases across scenarios. In each of these scenarios, X_{U2} was then estimated using equation (10) and the covariate X_{U3}.

Propensity Score Approaches Using the Simulated Data

Once treatment choices were generated for the simulated patients in each of the eight scenarios, a treatment propensity score was estimated for each patient as the predicted probability of treatment based on a linear probability model using the measured covariate X_{M}, where X_{Mi} equals the value of the measured covariate for patient i:

$$\hat{T}_i = \hat{\delta}_0 + \hat{\delta}_{\mathrm{M}} X_{\mathrm{Mi}}. \tag{12}$$

$\hat{T}_i$ equals the predicted propensity score for patient i given X_{Mi}. Table 1 contains the R^2 value for each scenario showing the proportion of T variation described by X_{M}. Using the estimated propensity scores, three distinct PS algorithms were used to balance X_{M} between treated and untreated patients. First, simulated patients were stratified into five PS bins using the algorithm described by D'Agostino (1998). The full sample was grouped into quintiles of the propensity score, with "Bin 1" containing the fifth of the sample with the lowest probability of choosing treatment (lowest propensity scores), and "Bin 5" containing the fifth of the sample with the highest probability of choosing treatment (highest propensity scores). Second, the entire sample was reweighted using the inverse probability weighting algorithm $\hat{T}_i$ based on Imbens

(2000) and Robins et al. (2000). Third, treated patients were matched with untreated patients based $\hat{T}_i$ on the using matching without replacement and three different criteria for the required "closeness" of propensity scores to be considered a match, sometimes referred to as the match "tolerance" or "caliper" value; specifically, the match tolerance was varied from 0.1, 0.01, and 0.001 (Stuart 2010).

Within each scenario, the ability of the PS algorithms to balance covariates was assessed by comparing the means of the measured covariate X_M and the unmeasured covariates for treated and untreated patients within (1) the full unweighted sample; (2) the full sample weighted by the inverse probability weights; (3) the samples within each of the five PS-stratified bins; and (4) patient samples matched by propensity scores. In addition, for each unmeasured covariate X_{Ui}, we estimated R_i as the residual of the regression of X_M on X_{Ui}. R_i contains the portion of the X_{Ui} variance that is not "redundant" with X_M. We assessed the impact on the balance of these residuals between treated and untreated patients using each PS sampling approach in each scenario.

RESULTS

The results for all model covariates and cure rates for each simulation scenario are provided in the Appendices. Appendices A, B, C, D, and E contain results for X_{U1}, X_{U2}, X_{U3}, X_{U4}, and cure (C), respectively. For comparison purposes, each appendix contains the results for X_M. The Appendices report the means of each covariate for the treated and untreated patients and the mean difference between the treated and untreated patients for each model scenario and sampling approach. Appendices A–D also contain results for each R_i—the portion of X_{Ui} unrelated X_M that was measured as the residuals of the regression of X_M on each X_{Ui}–R_i.

General Results across Scenarios

The first row for each scenario in the appendices shows the imbalance for each model covariate between treated and untreated patients in the full unweighted sample. For example, in scenario 2 in Appendix A, the mean of X_M is .617 for treated patients and .380 for untreated patients for a difference of .237; and the mean of R_1 is .115 for treated patients and −.115 for untreated patients for a difference of .230. The remaining rows under each scenario show the means of each model covariate for treated and untreated patients for the respective PS algorithm. In scenario 2 in Appendix A, using the matching algorithm with (.01) tolerance, the mean of X_M is .494 for treated patients and .505 for untreated patients for a difference of .011, and the mean of R_1 is .138 for treated patients and −.137 for untreated patients for a difference of .275. As a result, matching by propensity scores *decreased* the imbalance in the measured covariate X_M by 95 percent (100*(.237–.011)/.237) but *increased* imbalance in R_1—the portion of X_{U1} unrelated to X_M—by 20 percent (100*(.230–.275)/.275). Table 2 contains a summary of the change in imbalance between treated and untreated patients moving from the full unweighted sample to each PS-adjusted sample for X_M and the portion of the variation in each unmeasured covariate that that is unrelated to X_M – R_1, R_2, R_3, & R_4, for the covariates X_{U1}, X_{U2}, X_{U3}, & X_{U4}, respectively. In *every scenario*, *for every unmeasured covariate related to treatment choice*, the imbalance in the portion of those covariates that is unrelated to X_M *always increases* when moving from the full unweighted sample to PS-adjusted samples.

Table 2: Percent Reduction (Increase) in Sample Mean Difference (Treated–Untreated) from PS Methods for the Measured Covariate X_M, and the Independent Portions of the Unmeasured Covariates Affecting Treatment Choice R_1, R_3, and R_4

	Scenario 1	*Scenario 2*	*Scenario 3*	*Scenario 4*	*Scenario 5*	*Scenario 6*	*Scenario 7*	*Scenario 8*
X_M								
Binning*	95.5%	95.7%	93.5%	94.9%	91.1%	n/a	n/a	n/a
Matching†	85.7	95.4	98.9	97.7	99.1	99.2%	99.0%	98.4%
Weighting	100.8	101.3	122.6	100.0	61.3	22.7	16.8	15.0
R_1								
Binning*	−4.4	−20.0	−81.5	−42.0	−157.4	n/a	n/a	n/a
Matching†	−3.9	−19.6	−76.1	−40.0	−130.2	−323.5	−720.0	−2,200.0
Weighting	−4.6	−21.9	−117.9	−53.3	−88.6	−75.0	−33.3	−66.7
R_3								
Binning*	−4.5	−20.3	−82.8	−41.5	−155.1	n/a	n/a	n/a
Matching†	−3.8	−19.7	−75.6	−28.0	−134.9	−343.8	−730.0	−2,300.0
Weighting	−4.6	−21.4	−111.0	−51.6	−101.2	−71.9	−70.0	−40.0
R_4								
Binning*	−4.2	−19.9	−83.3	−41.3	−155.3	n/a	n/a	n/a
Matching†	−4.2	−18.5	−74.4	−38.5	−131.4	−334.4	−740.0	−2,500.0
Weighting	−4.9	−21.6	−111.3	−53.8	−95.3	−68.8	−60.0	−70.0

*Average across 5 PS-quintile "bins," n/a when not all five bins contained patients.
†For match tolerance value = 0.01 (for matched portion of total sample only).

Appendix E contains differences in cure rates for each scenario between treated and untreated patients for the full unweighted sample and each PS-adjusted sample. Table 3 summarizes the treatment effect estimates across scenarios using propensity score methods and standard regression estimators. In all scenarios, the true value of β_T, the incremental effect of treatment on the probability of a cure, is 0.2. The unmeasured confounding covariates in each scenario—X_{U1} and X_{U2}—are both positively related to treatment choice and the probability of cure. Therefore, the direct estimation of the effect of treatment on cure after controlling for X_M alone should yield estimates of treatment effect that are biased high. Column I in Table 3 shows the difference in cure probabilities between treatment and untreated patients without adjusting for X_M. Column II contains the regression-based treatment effect estimate using the full unweighted sample after controlling for X_M. Regression-based estimates only come close to the true value of β_T when X_M is highly correlated with the unmeasured confounders as in scenarios 7 and 8. Columns III, IV,

Table 3: Treatment Effect Estimates by Scenario Using (1) Regression and (2) Differences between the Cure Rate for the Treated and Untreated Patients from Propensity Score Algorithms

		I	*II*	*III*	*IV*	*V*		
						Matching Tolerance Based on Propensity Score¶		
Scenario	*Truth*	*No Control**	*Regression Estimate Controlling for X_M*†	*Inverse Probability Weighting*‡	*Average across Bins*§	*.1*	*.01*	*.001*
1	.200	.263	.254	.255	.255	.263	.258	.257
2	.200	.273	.257	.257	.257	.268	.261	.260
3	.200	.271	.260	.286	.265	.265	.262	.261
4	.200	.275	.254	.255	.256	.261	.254	.253
5	.200	.282	.232	.256	.239	.240	.234	.233
6	.200	.304	.227	.291	.239	.242	.231	.233
7	.200	.318	.203	.299	.223	.220	.212	.210
8	.200	.338	.194	.316	.220	.213	.222	.225

*Difference in cure rates between treated and untreated patients without control for $\mathrm{X_M}$.

†Linear multiple regression estimate of the effect of T on cure controlling for $\mathrm{X_M}$.

‡Weighted difference in cure rates between treated and untreated patients with weights based on the propensity score.

§Average of the difference in cure rates between treated and untreated patients across the five propensity score bins.

¶Difference in cure rates between propensity score-matched treated and untreated patients.

and V contain estimates using inverse probability weighting, binning, and matching, respectively. Although differences in the extent of bias across estimation approaches were generally small, the treatment effect estimates generated via the propensity score algorithms were generally larger (more biased) than regression estimates with only one exception (the matching estimate in scenario 4 using .001 tolerance). Nonetheless, these simulation results suggest that propensity score algorithms are unlikely to reduce the bias in estimated treatment effects compared with regression estimates in the presence of unmeasured confounders.

Smaller Unmeasured Covariate Value Weights Increase Unmeasured Covariate Imbalance

Scenarios 1, 2, and 3 were constructed with variation in the value weights placed on each covariate in the treatment value relationship but with no relationships between the measured and unmeasured covariates. The treatment value weights applied to the unmeasured covariates fall relative to the value weight placed on the measured covariate moving from scenario 1 to scenario 3. Figures 2 and 3 summarize these results focusing on X_M and the portion of X_{U1} variation unrelated to X_M–R_1. Figures 2 and 3 show the percent change in

Figure 2: Percent Reduction (Increase) in Sample Mean Difference (Treated–Untreated) from PS Binning for X_M and R_1, by Unmeasured Covariate Effect Size (α_{U1})

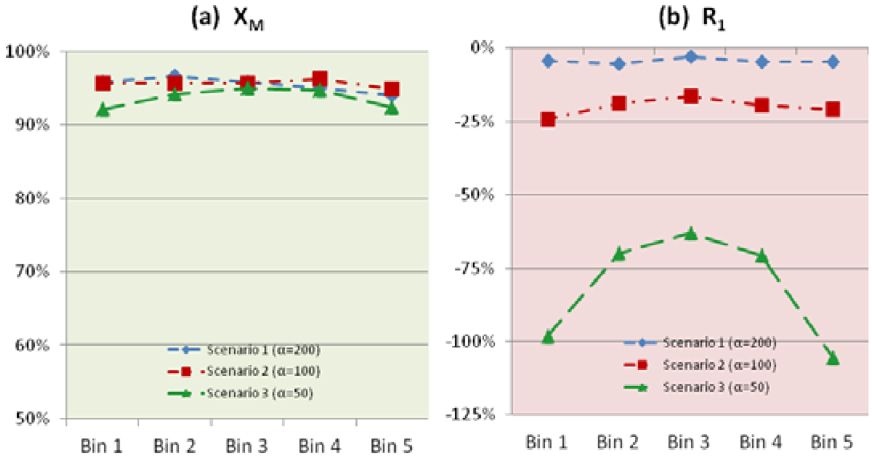
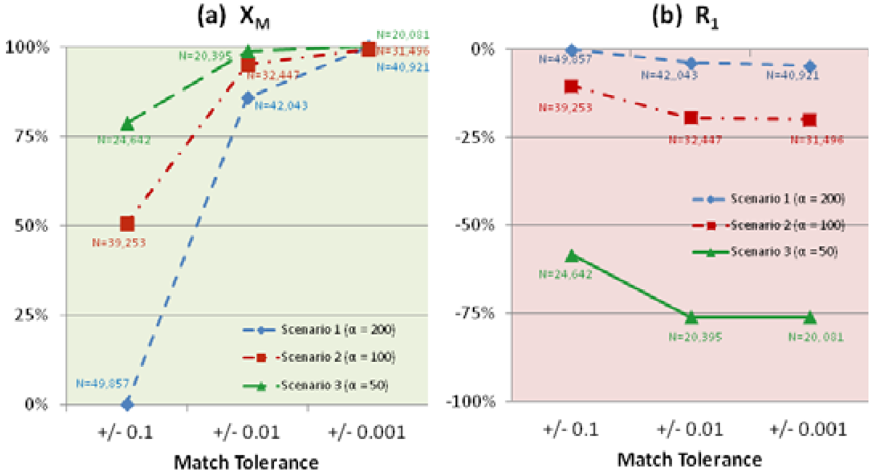


Figure 3: Percent Reduction (Increase) in Sample Mean Difference (Treated–Untreated) from PS Binning for X_M and R_1, by Unmeasured Covariate Effect Size (α_{U1})

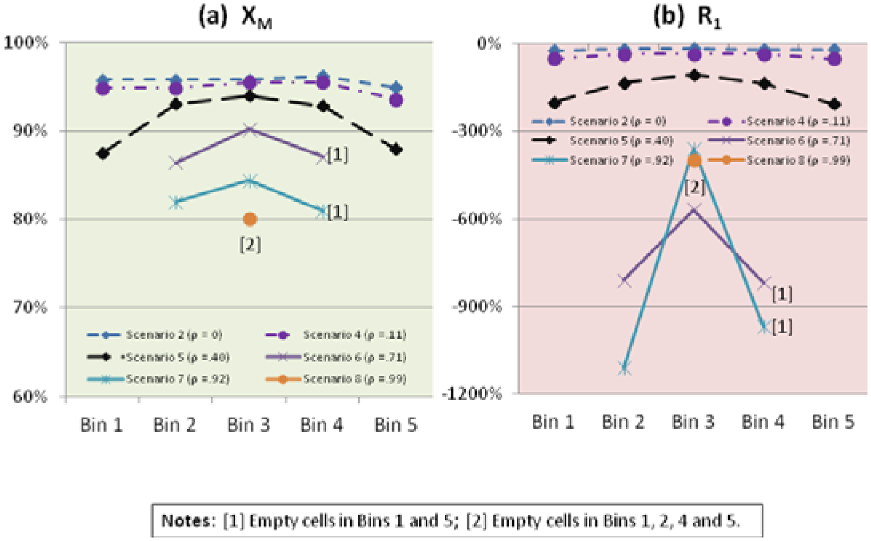


difference in X_M and R_1 balance between treated and untreated patients as a result of propensity score binning and propensity score matching, respectively. The green lines represent the balance change for the scenario with the lowest relative treatment value weight placed on the unmeasured covariates. Regardless of propensity score method used, the smaller the relative treatment weight placed on the unmeasured covariates, the greater the increase in the imbalance in unmeasured covariates when propensity score methods are applied (Figures 2b and 3b). This occurs because smaller relative value weights placed on the unmeasured requires greater differences in these covariates between treated and untreated patients for them to match on the measured covariate X_M. Interestingly, with respect to X_M, the ability to balance X_M falls with relative value weights placed on the unmeasured covariates using the propensity score binning approach (Figure 2a), whereas it increases with relative value weights placed on the unmeasured covariates using the matching approach (Figure 3a). The matching process appears to do a better job eliminating matches dissimilar in X_M when the unmeasured covariates have less weight in treatment choice. However, the number of simulated treated and untreated patients that are matched falls with relative value weights placed on the unmeasured covariates.

Greater Correlation between Measured and Unmeasured Covariates Increases Unmeasured Covariate Imbalance

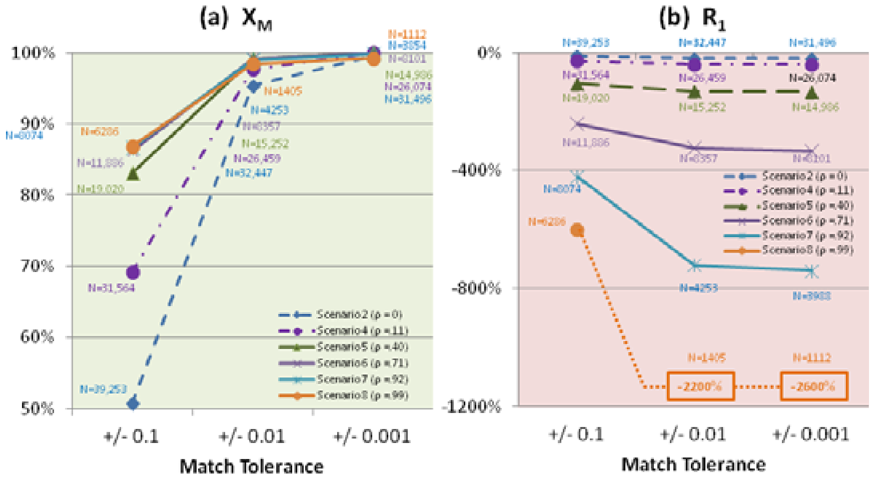
Scenarios 2, 4, 5, 6, 7, and 8 have consistent treatment value weights across covariates, but they vary with the strength of the relationships between the measured and unmeasured covariates. The strength of the relationships between X_M and the unmeasured covariates increases moving from scenarios 2 through 8. Figures 4 and 5 summarize these results, focusing on X_M and the portion of X_{U1} variation unrelated to $X_M - R_1$. Figures 4 and 5 show the percent change in difference in X_M and R_1 balance between treated and untreated patients as a result of propensity score binning and propensity score matching, respectively. Both figures show that as X_M explains more of the variation in the unmeasured covariates, the more difficult it is to find patients that made different treatment choices with similar propensity scores. For example, in scenario 8, no treated patients were found in propensity score bins 1 and 2, and no untreated patients were found in bins 4 and 5. Likewise, in scenario 8 only 1,112 treated patients were matched to an untreated patient using matching

Figure 4: Percent Reduction (Increase) in Sample Mean Difference (Treated–Untreated) from PS Binning for X_M and R_1, by Correlation between X_M, X_{U1} ($\rho_{M,U1}$)



Note. [1] Empty cells in Bins 1 and 5; [2] Empty cells in Bins 1, 2, 4, and 5.

Figure 5: Percent Reduction (Increase) in Sample Mean Difference (Treated–Untreated) from PS Matching for X_M and R_1, by Match Tolerance Factor (±0.1, ±0.01, ±0.001)



tolerance of .001. Regardless of the propensity score method used, the imbalance in R_1 between treated and untreated patients increases with the strength of the relationships between X_M and the unmeasured covariates. With respect to balancing X_M, stronger relationships between X_M and the unmeasured covariates moving from scenarios 2 through 8 are similar to decreasing the treatment value weights moving from scenarios 1–3. The ability to balance X_M falls with stronger relationships with the unmeasured covariates using the propensity score binning approach (Figure 4a), whereas it increases using the matching approach (Figure 5a). The matching process does a better job eliminating matches dissimilar in X_M when the unmeasured covariates have less independent effect on treatment choice, but the number of matched simulated treated and untreated patients falls dramatically.

DISCUSSION

It has been suggested that researchers estimating treatment effects using observational data should use propensity score-based algorithms to mimic a randomized controlled trial design. The strength of randomization, however, is the assumption that it will evenly distribute covariates (both measured and unmeasured) across treated and untreated patients. The results of randomization are often reinforced in studies by demonstrating balance in measured covariates. In addition, many researchers using propensity score methods to estimate treatment effects with observational data imply that the measured covariate balance produced by these methods results in unmeasured covariate balance (Shah et al. 2005).

Our simple simulation model suggests that balancing measured covariates between treated and untreated patients actually has opposite implications for randomized and observational studies. We first showed that to balance measured covariates, PS algorithms require the existence of unmeasured covariates with variation unrelated to the measured covariates. This independent variation is needed to ensure that patients with similar propensity scores are observed making different treatment choices. Second, whereas demonstrated measured covariate balance between treated and untreated patients resulting from ex ante randomized treatment assignment reinforces the notion that all covariates are balanced, we showed that the *forced* balance of measured covariates using PS-based algorithms based on ex post treatment selection in observational studies *exacerbates* the imbalance in the variation of the unmeasured covariates that is unrelated to the measured covariates. In addition, the greater the impact that *measured* covariates have on treatment choice relative to *unmeasured* covariates, the more that the forced balance of measured covariates increases unmeasured covariate imbalance. This can be likened to squeezing a balloon. When a set of patients making different treatment decisions are forced to be balanced on one set of covariates (measured), this must be compensated by increased imbalance in the remaining unmeasured covariates affecting treatment choice. These results have implications on the use of higher dimensional propensity scores for balancing treated and untreated patients (Rassen, Brookhart et al. 2009; Schneeweiss et al. 2009). The more the variation in treatment choice that is explained by measured covariates, the harder it is to match treated and treated patients and the more imbalance in the unmeasured covariates will occur between the treated and untreated patients that are matched by propensity scores.

Because of these results, for PS methods to yield unbiased treatment effect estimates, the remaining factors affecting treatment choice have to have the properties of X_{U4}, which would be natural instruments if they were measured. Treatment variation caused by the class of variables X_{U4} can be thought of as "good" treatment variation that is tantamount to a natural experiment, but theory is required to justify the notion that all remaining treatment variations are from the class of variables X_{U4}. This is why researchers should not put measured covariates thought to be instruments in propensity score models because this reduces the amount of "good" treatment variation used to estimate treatment effects (Bhattacharya and Vogt 2007).

Conclusion

We used a simple model of treatment choice and outcome to assess the effect of propensity algorithms that balance measured covariates between treated and untreated patients on the balance of unmeasured covariates between these patients. We found that propensity score algorithms that balance measured confounders between treated and untreated patients *exacerbate* imbalance for these same patients in a portion of the variation in unmeasured confounders that is unrelated to measured covariates. However, independent variation in the unmeasured covariates is required for propensity score algorithms to balance measured covariates between treated and untreated patients. Although our simulation model was simple, we challenge researchers to construct alternative models in which in the independent variation in the unmeasured covariates that affect treatment choice *becomes more balanced* as a result of using propensity score algorithms to balance measured covariates. As in regression-based estimation, researchers using propensity score algorithms still must provide theoretical justification for the assumption that the unmeasured covariates affecting treatment choice have no direct or indirect effects on study outcomes. Indeed, based on the results here, acceptance of this assumption appears to be even more critical when using propensity score methods as the imbalance between treated and untreated patients in the portion of the variation of the unmeasured covariates unrelated to the measured covariates will increase. As a result, if the unmeasured covariates affecting treatment choice are confounders, propensity score methods can exacerbate the bias in treatment effect estimates.

Acknowledgments

Joint Acknowledgment/Disclosure Statement: This project was supported in part by an Agency for Healthcare Research and Quality Centers for Education and Research on Therapeutics cooperative agreement (5UI8HS016094) and AHRQ grant (ROIHS018381). The authors would also like to acknowledge the extremely valuable input of Brian Dowd and the anonymous reviewers of the manuscript. Any remaining errors are the responsibility of the authors.

Disclosures: None.

Disclaimers: None.

References

Angrist, J. D., and J.-S. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, New Jersey: Princeton University Press.

Austin, P. C., P. Grootendorst, and G. M. Anderson. 2007. "A Comparison of the Ability of Different Propensity Score Models to Balance Measured Variables between Treated and Untreated Subjects: A Monte Carlo Study." *Statistics in Medicine* 26 (4): 734–53.

Ben-Akiva, M., and S. R. Lerman. 1985. *Discrete Choice Analysis.* Cambridge, Massachusetts: The MIT Press.

Berk, R. A. 2004. *Regression Analysis: A Constructive Critique.* Thousand Oaks, California: Sage Publications.

Bhattacharya, J., and W. B. Vogt. 2007. "Do Instrumental Variables Belong in Propensity Scores?" Technical Working Paper 343, National Bureau of Economic Research, Cambridge, MA, http://www.nber.org/papers/t0343.

Brooks, J. M., and G. Fang. 2009. "Interpreting Treatment-Effect Estimates with Heterogeneity and Choice: Simulation Model Results." *Clinical Therapeutics* 31 (4): 902–19.

D'Agostino, R. B. 1998. "Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-Randomized Comparison Group." *Statistics in Medicine* 17: 2265–81.

Frolich, M. 2007. "Propensity Score Matching without Conditional Independence Assumption–with an Application to the Gender Wage Gap in the United Kingdom." *Econometrics Journal* 10 (2): 359–407.

Hall, J. A., K. H. Summers, and R. L. Obenchain. 2003. "Cost and Utilization Comparisons among Propensity Score-Matched Insulin Lispro and Regular Insulin Users." *Journal of Managed Care Pharmacy* 9 (3): 263–8.

Imbens, G. W. 2000. "The Role of the Propensity Score in Estimating Dose-Response Functions." *Biometrika* 87 (3): 706–10.

Joffe, M. M., and P. R. Rosenbaum. 1999. "Invited Commentary: Propensity Scores." *American Journal of Epidemiology* 150 (4): 327–33.

Lin, D. Y., B. M. Psaty, and R. A. Kronmal. 1998. "Assessing the Sensitivity of Regression Results to Unmeasured Confounders in Observational Studies." *Biometrics* 54 (3): 948–63.

Rassen, J. A., M. A. Brookhart, et al. 2009. "Observed Performance of the High-Dimensional Propensity Score (hd-PS) Algorithm in Small Samples." *Pharmacoepidemiology and Drug Safety* 18: S15–S15.

Robins, J. M., M. A. Hernan, and B. Brumback. 2000. “Marginal Structural Models and Causal Inference in Epidemiology.” *Epidemiology* 11 (5): 550–60.

Rosenbaum, P. R. 1987. “Model-Based Direct Adjustment.” *Journal of the American Statistical Association* 82 (398): 387–94.

Rosenbaum, P. R., and D. B. Rubin. 1983a. “Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome.” *Journal of the Royal Statistical Society Series B-Methodological* 45 (2): 212–8.

———. 1983b. “The Central Role of the Propensity Score in Observational Studies for Casual Effects.” *Biometrika* 70 (1): 41–55.

Rubin, D. B. 1997. “Estimating Causal Effects from Large Data Sets Using Propensity Scores.” *Annals of Internal Medicine* 127 (8): 757–63.

———. 2001. “Using Propensity Scores to Help Design Observational Studies: Application to Tobacco Litigation.” *Health Services & Outcomes Research Methodology* 2: 169–88.

———. 2007. “The Design versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials.” *Statistics in Medicine* 26 (1): 20–36.

Schneeweiss, S., J. A. Rassen, R. J. Glynn, J. Avorn, H. Mogun, and M. A. Brookhart. 2009. “High-dimensional Propensity Score Adjustment in Studies of Treatment Effects Using Health Care Claims Data.” *Epidemiology* 20 (4): 512–22.

Shah, B. R., A. Laupacis, J. E. Hux, and P. C. Austin. 2005. “Propensity Score Methods Gave Similar Results to Traditional Regression Modeling in Observational Studies: A Systematic Review.” *Journal of Clinical Epidemiology* 58 (6): 550–9.

Stuart, E. A. 2010. “Matching Methods for Causal Inference: A Review and a Look Forward.” *Statistical Science* 25 (1): 1–21.

Ward, A., and P. J. Johnson. 2008. “Addressing Confounding Errors When Using Non-Experimental, Observational Data to Make Causal Claims.” *Synthese* 163 (3): 419 –32.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix A: Mean Measured (X_{M}) and Unmeasured Covariate (X_{U1}) Values across Model Scenarios.

Appendix B: Mean Measured (X_{M}) and Unmeasured Covariate (X_{U2}) Values across Model Scenarios.

Appendix C: Mean Measured (X_{M}) and Unmeasured Covariate (X_{U3}) Values across Model Scenarios.

Appendix D: Mean Measured (X_{M}) and Unmeasured Covariate (X_{U4}) Values across Model Scenarios.

Appendix E: Mean Measured (X_{M}) and Cure (C) Rates across Model Scenarios.

Appendix SA1: Author Matrix.