STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:2055–2061
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3245

COMMENTARY

Discussion of research using propensity-score matching: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin, *Statistics in Medicine*

Jennifer Hill*, †

Department of International and Public Affairs, Columbia University, New York, NY, U.S.A.

1. INTRODUCTION

Research using propensity-score matching has been in the literature for over two decades now. During this time, in a process akin to the way a message gets distorted and passed on in the children's game of 'telephone,' widespread dissemination has led to misunderstandings regarding the required assumptions, goals, and appropriate implementation of propensity-score matching. Thus, the bad practice that exists today is due, in large part, to degrees of separation from original sources coupled with the changing knowledge base and the time lag between new information appearing in the statistics literature and it reaching applied researchers. Another culprit more intrinsic to the nature of the method itself (at least in its current incarnation) is the 'art form' involved in proper practice [1]. Irrespective of how the current state of affairs came to be, a remedy is warranted. Peter Austin should be commended for addressing the rampant lack of good practice in propensity-score matching applications with some much needed policing and rehabilitation.

Austin provides some useful advice with regard to good practice (I avoid the term 'best practice' since there seems to be no consensus as to what this comprises). I especially appreciate his push for explicit discussion of the strategy used to create matched pairs and examination of balance across matched groups.

Austin also provides advice with which I don't agree. I am particularly at odds with his position requiring matched pairs' analyses, which is an overly narrow approach to the problem. There are many ways to address the lack of independence across samples, and methods that explicitly adjust for pairwise dependence are not always the best choice (even if, algorithmically, the dependence was created by forming matched pairs). Moreover, Austin gives this issue undue weight compared

*Correspondence to: Jennifer Hill, Department of International and Public Affairs, Columbia University, New York, NY, U.S.A.
†E-mail: jh1030@columbia.edu

with other potentially far more important considerations not even mentioned in the paper. In my opinion, there are often bigger fish to fry.

The following discussion highlights and attempts to correct some points of confusion regarding propensity-score matching implementation that I have encountered in the literature, while consulting with and teaching applied researchers, or through my own missteps. My comments regarding Austin's advice will be integrated when appropriate within this framework.

2. DEFINING CONFOUNDERS

Propensity-score matching relies strongly on the assumption of ignorability, formally, $Y(0), Y(1) \perp Z|X$, where X represents a vector of confounders, Z denotes treatment assignment, and $Y(0)$ and $Y(1)$ are the potential outcomes under control and treatment conditions, respectively (although, strictly speaking, many estimands rely on weaker forms of this assumption). Intuitively, this assumption maps to the colloquial notion of 'all confounders measured' (or the economists concept of 'selection on observables' [2]).

Some manifestations of bad practice in propensity-score matching appear to have arisen from a failure to correctly define confounders. We can define a confounder as a variable that predicts both the treatment and the outcome (see [3] for a precise, thorough, and understandable discussion of confounding including a technical exception to the definition just provided). Failure to understand this key point has led to descriptions of confounders such as 'variables that predict treatment assignment' and 'the predictors in our propensity-score estimation model.' These seemingly innocuous misunderstandings are arguably to blame for dangerous practices such as the use of stepwise regression to choose a propensity-score estimation model (also mentioned by Austin) and checking balance only on the predictors remaining in the final estimation model. A practice I find helpful is to rank order the confounders based on their importance with respect to the outcome variable. This reminds the researcher of the importance of the confounders' association with the outcome and also facilitates choice between propensity-score models that yield roughly equivalent overall balance but create trade-offs in balance on individual variables.

A final point with regard to confounders is that it can be dangerous to choose them based on simple empirical summaries (such as pairwise correlations or t-statistics in a linear regression model of the outcome on all potential covariates). Consider a simple but striking example with a potential confounder, W, distributed $\mathrm{N}(0, 1)$ and in which the mean of the observed outcome takes the form $E[Y|W] = W^2$. What would happen if sample correlation between W and the observed outcome were used as a criterion for determining whether to treat W as a confounder? The researcher might be misled into discarding W from the list of confounders (the correlation between W and Y is 0) when in fact imbalance in the standard deviation of this variable could lead to substantial bias.

3. BALANCE

One of the biggest selling points of propensity-score matching is the balance diagnostic. In theory this is a model check that is divorced from treatment effect estimation and is intended to encourage honesty. Unfortunately, balance checking is also one of the most problematic aspects of propensity-score matching because the criteria for adequate balance are ill-defined [4, 5]. Austin rightly points

DOI: 10.1002/sim

out the need for researchers to display balance checks and I agree with his admonition against using t-statistics or other sample-size-dependent tests as a means for determining adequate balance (see [4] for a nice discussion). Beyond discouraging inappropriate choices, however, we should also be advocating for additional improvements over standard practice; two such suggestions follow.

First, rather than only considering mean differences, researchers should be encouraged to examine higher-order sample moments such as variances (or standard deviations) and covariances (or correlations). If the true response surface is non-linear in the confounders (a presumption that typically drives the use of propensity-score matching), imbalance in higher-order moments could result in bias even if mean balance has been achieved. Consider again the example in which $E[Y|W]=W^2$. Even if the mean of W is perfectly balanced across samples, imbalance in the standard deviation could still result in substantial bias. Some user-written matching packages (such as *MatchIt* and *Matching* in R) now provide the option for more comprehensive balance checks such as balance checks of 'squared terms and interactions' (sample second moments) and even distributional equivalence using empirical quantile–quantile plots and associated statistics.

Second, since there is no strong theory regarding when balance is close enough, researchers should be encouraged to examine the sensitivity of their results to a range of propensity-score specifications that all yield seeming 'adequate' balance overall but that might make trade-offs in balance of some variables over others (see [1] for a nice example). More formal sensitivity analyses such as those proposed by Rosenbaum [6] and by Greenland [7] would be even better, but a discussion of these is beyond the scope of this commentary.

4. THE ESTIMAND: WHO ARE WE MAKING INFERENCES ABOUT?

Propensity-score matching typically focuses not on the average treatment effect, $E[Y(1)-Y(0)]$ but on either the effect of the treatment on the treated, $E[Y(1)-Y(0)|Z=1]$ or the effect of the treatment on the controls, $E[Y(1)-Y(0)|Z=0]$. This focus can be a strength or a weakness given the inferential goals of the paper [8]. All too often, researchers fail both to make explicit the appropriate estimand and to elucidate the relationship between this estimand and their research question and target population. While there is always a leap from empirical results to policy, practice, or scientific conclusions, clearly defining the estimand can help to better characterize that leap.

The definition of the estimand becomes still more fuzzy when there is insufficient overlap and researchers restrict their analyses to areas of 'common support.' For instance, when finding matches among the control group for treated subjects, lack of sufficient overlap across groups may encourage the researcher to drop treated subjects with propensity scores that are higher than the highest control propensity score (or, less commonly, lower than the lowest control propensity score). However, this changes our causal estimand. More work needs to be done to identify the risks of and trade-offs involved in extrapolating models beyond the common support. In the meantime, studies that are restricted to common support should make an effort to 'profile' both the excluded group and the new inferential group (that is, those treated still left or the population they represent).

Finally, researchers should be clear about whether they are intending to make sample or population inference [9]. This distinction becomes still more important when deciding how to address the lack of independence induced by matching.

DOI: 10.1002/sim

5. POST-MATCHING ANALYSIS

The goal of matching is to create a setting within which treatment effects can be estimated without making heroic parametric assumptions. If ignorability holds and the matching is able to create 'sufficient' balance across treatment groups (recognizing, again, that this criterion is not well defined), the options available for estimating treatment effects should be similar to those available in a randomized setting. In the setting that is the focus of Austin's paper (one-to-one matching without replacement), restricting to pairs effectively changes the probability of being treated in each pair to be equal; thus we can ignore the pairs and still obtain unbiased point estimates of the effect of the treatment on the treated. From the perspective of the standard error for this estimate, however, it may be inefficient to ignore the pairs, or, more generally, to ignore the dependence induced by the pair matching. Austin focuses his energies with regard to post-matching analyses primarily in advocating use of 'analyses appropriate for matched pair data' to avoid variance estimates that disregard the correlation typically induced in this framework. In this section I lay out a slightly more general framework for choices in post-matching analysis.

I discuss four considerations with regard to post-matching analyses: (1) residual bias due to imperfect matching, (2) precision, (3) lack of independence across units, and (4) collapsibility of the estimand. The second and fourth issues are also considerations in completely randomized experiments. For simplicity, I address the first three concerns assuming that the goal is to estimate the difference in mean potential outcomes for the treatment group. I then address non-collapsible treatment effect estimands.

5.1. Residual bias

In practice we don't expect that matching will remove all bias due to confounders because we are typically unable to match sufficiently closely. This can be true even if we adjust for matched pairs because the within-pair differences in covariate values may still be large [10]. Many authors have noted the advantages of performing additional covariance adjustment with respect to bias reduction, among them [9–13]. In practice, such adjustment can be useful even with a simple post-matching model such as a linear regression on the treatment indicator and (possibly some subset of) confounders using the matched data [4, 14].

5.2. Precision

Just as in randomized experiments, in the matched setting we may also wish to increase the precision of the treatment effect estimator for our matched sample. In fact, precision may be especially salient in this setting because matching (one-to-one matching in particular) has the potential to discard a substantial proportion of our original sample. Further covariance adjustments through stratification or regression modeling, for example, are likely to yield more efficient treatment effect estimates. It's also important to keep in mind that while random deviations of our estimate from the population average treatment effect translate into inefficiency, they represent bias if our goal is to estimate the sample average treatment effect.

5.3. Lack of independence across units

Matching on propensity scores induces dependence across the matched samples. Failure to address this aspect of the observational study design can result in standard errors that are overly conservative.

DOI: 10.1002/sim

Since this dependence is created through pair matching, the use of some sort of matched pairs' analysis, as Austin prescribes, is certainly an appropriate way to account for this dependence. What I take issue with is Austin's assertion that this is the only appropriate choice and his lack of emphasis on the issue of additional covariance adjustment.

Matched pairs' analysis is most powerful when it results in highly correlated outcome variables. Yet, in a paper that emphasizes the goal of propensity-score matching for producing similar sample means across treatment and control groups, Rubin and Thomas point out that 'an important feature of matching on the estimated linear propensity scores is that it does not require close pairwise matches' to do so [14]. Matched pairs' analyses in these settings can lead to loss of power.

5.3.1. Post-matching covariance adjustment. The impact on our inference of corrections for pair dependence tends to get swamped by the impact of additional covariance adjustments which, as described above, have the potential to help with both bias and precision. Moreover, given that the dependence across treatment and control groups is induced in effect by matching with respect to observed covariates, X, observations can be considered exchangeable conditional on X. Therefore, in theory, further (post-matching) adjustment on X (such as by running a regression on the treatment indicator and confounders in the matched sample) is another way to account for the lack of independence. Moreover, one could additionally perform a matched pairs' adjustment in this context (for instance, by including pair indicators).

What is the downside of this approach? First, you might use the wrong model; after all it was probably this fear that motivated use of propensity-score matching in the first place. However, the matching (and consequent balance across treatment and control groups) should allow for a fair amount of robustness to model misspecification [4]. A potentially bigger problem was alluded to (though not fully elucidated) by Austin, who goes to great length to distinguish between the desire to estimate marginal *versus* conditional treatment effects. Let's examine this closer. In the linear setting, this distinction isn't a relevant for our treatment effect estimate. Assuming ignorability (and sufficiently close matches) both marginal estimates and conditional (covariance-adjusted) estimates should unbiasedly estimate both the population average treatment effect for the treated and the sample average treatment effect for the treated because $E[Y(1)-Y(0)|Z=1]=E_X[E[Y(1)-Y(0)|Z=1,X]]$. The key distinction is in the standard errors.

The variance of the conditional estimate, on the other hand, may indeed be different from the variance of the unconditional estimate. Denoting our conditional estimator by $\hat{\tau}(x)$, we can see by the following decomposition $V[\hat{\tau}(x)|Z=1]=E_X[V[\hat{\tau}(x)|Z=1,X]]+V_X[E[\hat{\tau}(x)|Z=1,X]]$ that they will differ if the final term is not 0. They will be asymptotically equivalent if the treatment effect is constant. Hence, strictly speaking, if heterogeneous treatment effects exist we can only get correct standard errors for sample (or conditional) inference.

How limiting is this in practice? There is a fair amount of disagreement on this point (for a related discussion see [13] along with discussions [15, 16]). However, a practical solution that has many side benefits in terms of understanding the science of the phenomenon is to explicitly investigate the degree to which one's confounders moderate the treatment effect. If we stratify on the covariates that modify the treatment effect (or allow for interactions between them and the treatment), then the problem may be solved. The practical difficulty in accomplishing this could be reduced by limiting the number of covariates one adjusts for post-matching.

Further, a helpful property of the propensity score (highlighted recently in another setting [17]) is that traditional models for heterogeneous treatment effects imply that the expected outcome is

DOI: 10.1002/sim

a non-linear function of the propensity score. Therefore, a test for this non-linearity may provide a reasonable omnibus test for heterogeneous treatment effects with respect to all confounders.

5.3.2. Alternatives. Of course, many other strategies exist to deal with the lack of independence and these vary as well depending on whether the goal is sample or population inference. I will discuss a few of these choices here. Abadie and Imbens [18] recently proposed a class of estimators that use matching to calculate standard errors for matching estimates; these cover both sample and population inferences and can accommodate additional covariance adjustment as well (software available for Stata, Matlab, and R). Other proposed options that can accommodate additional covariance adjustment exist. Bootstrapping [9, 19] can be used only for population inference. Randomization-based inference [6, 13] can be used only for sample inference unless we can assume additive treatment effects or extend the strategy to stratify on covariate values that moderate the treatment effect (which entails knowing which these are). Multilevel models may be able to model the paired structure more efficiently than traditional matched pairs' estimators [20].

5.4. Non-collapsibility

What happens if the model most naturally fit to our outcome variable yields non-collapsible treatment effect estimates? A common example would be if we have a binary outcome. Further adjustments for the confounders would be facilitated by fitting a logistic regression. However, logistic regression produces coefficients that (when exponentiated) estimate the odds ratio, and the odds ratio is not collapsible. Therefore, *even if we have data from a randomized experiment*, the conditional odds ratio will not (generally speaking) be equal to the marginal odds ratio [21], and there is debate regarding which of these estimands is of primary interest [22]. Austin, who has previously written about this topic with respect to propensity matching [23], advocates estimating the marginal odds ratio in this setting for better interpretability.

My perspective on this is that in most cases it makes sense to avoid using non-collapsible parameters as causal estimands because it creates unnecessary trade-offs between interpretability of the estimands and potentially beneficial covariance adjustment. The good news is that this proscription is not, in practice, all that limiting. For instance, in the binary outcome variable case, one could still fit a logistic regression (or a probit regression or a classification tree) but then use it to estimate $E[Y(1)-Y(0)|Z=1]$ by comparing predicted probabilities. Standard errors could be estimated by bootstrapping or simulation (packages such as *Zelig* in R directly facilitate these types of calculations). There is also an argument to be made (see [24], for example) for using linear regression to model binary outcomes in many scenarios, which would be a particularly easy solution.

6. CONCLUSION

In conclusion, Peter Austin should be applauded for his work that points out clear trouble spots in applied research using propensity-score matching. I hope my discussion has raised some other pertinent concerns as well as providing additional analysis options. Moreover, I hope the dialogue between Austin and the discussants will open up further debate and research regarding best practice in propensity-score matching. There is much work to be done. Finally, I thank Joel Greenhouse for giving me this opportunity to express my views.

DOI: 10.1002/sim

REFERENCES

1. Dehejia RH. Practical propensity score matching: a reply to Smith and Todd. *Journal of Econometrics* 2005; **120**:355–364.
2. Barnow B, Cain G, Goldberger A. Issues in the analysis of selectivity bias. In *Evaluation Studies*, Stromsdorfer E, Farkas G (eds), vol. 5. Sage: San Francisco, 1980.
3. Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Statistical Science* 1999; **14**(1):29–46.
4. Ho DK, Imai K, King G, Stuart E. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 2007; **15**(3):199–236.
5. Sekhon JS. Alternative balance metrics for bias reduction in matching methods for causal inference. *Technical Report*, University of California at Berkeley, 2007.
6. Rosenbaum PR. *Observational Studies*. Springer: New York, 2002.
7. Greenland S. Basic methods for sensitivity analysis of biases. *International Journal of Epidemiology* 1996; **25**(6):1107–1116.
8. Heckman J, Robb R. Alternative methods for evaluating the impact of interventions. In *Longitudinal Analysis of Labor Market Data*, Heckman J, Singer B (eds). Wiley: New York, 1985.
9. Imbens G. Nonparametric estimation of average treatment effects under exogeneity: a review. *The Review of Economics and Statistics* 2004; **86**(1):4–29.
10. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 1985; **39**:33–38.
11. Rubin DB. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* 1973; **29**:185–203.
12. Abadie A, Imbens GW. Simple and bias-corrected matching estimators for average treatment effects. *Technical Report*, *NBER Working Paper*, 2002.
13. Rosenbaum PR. Covariance adjustment in randomized experiments and observational studies. *Statistical Science* 2002; **17**:286–327.
14. Rubin DB, Thomas N. Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association* 2000; **95**:573–585.
15. Angrist J, Imbens G. Comment on 'covariance adjustment in randomized experiments and observational studies'. *Statistical Science* 2002; **17**(3):304–307.
16. Robins J. Comment on 'covariance adjustment in randomized experiments and observational studies'. *Statistical Science* 2002; **17**(3):309–321.
17. Heckman JJ, Vytlacil E. Structural equations, treatment effects, and econometric policy evaluation. *Econometrica* 2005; **73**:669–738.
18. Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. *Econometrica* 2006; **74**(1):253–267.
19. Hill J, Reiter J. Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine* 2006; **25**(13):2230–2256.
20. Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press: New York, 2006.
21. Gail M, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regression and omitted covariates. *Biometrika* 1984; **71**:431–444.
22. Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials. *Controlled Clinical Trials* 1998; **19**(3):249–256.
23. Austin P, Grootendorst P, Normand SLT, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of the treatment effect: a Monte Carlo study. *Statistics in Medicine* 2007; **26**:754–768.
24. Hellevik O. Linear versus logistic regression when the dependent variable is a dichotomy. *Quality and Quantity* 2008; DOI: 10.1007/s11135-007-9077.

DOI: 10.1002/sim