STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:2062–2065
Published online 20 February 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.3207

COMMENTARY

Developing practical recommendations for the use of propensity scores: Discussion of 'A critical appraisal of propensity score matching in the medical literature between 1996 and 2003' by Peter Austin, *Statistics in Medicine*

Elizabeth A. Stuart*, †

Departments of Mental Health and Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, U.S.A.

I first wish to thank Dr Austin for an important paper on the use of propensity score methods in practice. It follows a series of papers by Dr Austin highlighting issues that applied researchers face in using propensity score methods—an area of research that deserves more attention. As Dr Austin shows, there has been considerable misuse and misunderstanding of propensity score methods. Papers such as his are important tools to ensure researchers use propensity scores in the most appropriate ways.

In this discussion I focus on two primary topics. First, I fully endorse and elaborate on Dr Austin's discussion of balance diagnostics of propensity score methods. Second, I discuss his statement regarding the requirement of analyses that account for the paired nature of the data after propensity score matching.

BALANCE DIAGNOSTICS

Dr Austin gave a thorough description of balance diagnostics for matching methods and emphasized the importance of having two steps: assessing the balance between the matched samples (and possibly refining the matching) and only then moving on to the analysis of the outcome. This is an idea highly related to the concept of the careful design of observational studies [1–3], which is of crucial importance but not always carefully considered. In addition, the availability of diagnostics for the success of the first step—the resulting balance between the matched samples—is an important feature.

*Correspondence to: Elizabeth A. Stuart, Department of Mental Health, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, U.S.A.
†E-mail: estuart@jhsph.edu

Balance is defined in [4] as the similarity between the multivariate empirical distributions of the covariates in the treated and control groups. An ideal situation would be to have exact matches on all covariates, such that the treated and control groups would have identical empirical distributions. This is rarely feasible, and thus researchers are left to select the matching method that yields the best balance, which involves comparing balance across many covariates at the same time.

As Dr Austin points out, non-randomized studies all too rarely check the balance of the covariates between the treated and control individuals. And even when these comparisons are done, they are often inappropriate. In particular, statistical tests are not appropriate ways to assess balance. There are a few reasons for this, two of which I discuss here. The first relates to the fact that balance is an in-sample quantity, defined without reference to a population or super-population. This idea is expressed in Dr Austin's quotes from Senn and Begg and elaborated on in [4]. Second, and perhaps more importantly, balance tests (and, for example, the resulting p-values) can lead to misleading results when comparing matched samples. This is because these tests generally conflate changes in balance with changes in sample size (or effective sample size, for weighting and subclassification approaches). This can lead researchers to select a matched sample with worse balance than another, simply because a reduction in power leads to an 'insignificant' test statistic. As illustrated in [4], balance may appear to improve when in reality the power of the test has simply decreased. This consequence of the use of test statistics in practice can be dangerous.

I propose two additional points for discussion, which were not explicitly mentioned in Dr Austin's paper. First, it is best if the standardized difference (formula shown in Section 2.1 of Dr Austin's paper) is calculated using the standard deviations of the covariates in the full (original) treated and control groups, when calculating the standardized difference in either the original or matched data. This ensures that the denominator remains the same pre- and post-matching, allowing for clear investigation of the resulting change in balance. Second, since balance involves the comparison of the multivariate empirical distributions of the covariates, balance should be checked not only on the means of the distributions, but also on their variances and important two- or three-way interactions [2]. I believe that none of the papers in Dr Austin's literature review would have met this more stringent criterion, but violating this criterion can produce major biases in causal estimates.

MATCHED-PAIR ANALYSES

Dr Austin made clear his disappointment that only 28 per cent of the studies he examined used outcome analyses that accounted for the matched-pair nature of the data. However, the statistical and methodological research is less clear regarding the need for matched-pair analyses than Dr Austin implies. There is a wide array of opinions on the topic of whether analyses of 'matched' data need to account for the matched pairs. This debate also fits into the broader discussion of variance estimation for propensity score methods.

On the one hand, some research has indicated that analyses considering the matched pairs are appropriate and outperform analyses that pool the pairs and just consider the matched groups as a whole. Early simulation work by Rubin [5] found that a matched-pair regression estimator had slightly less bias than did a pooled estimator, but that the pooled estimator performed well too. Rubin also acknowledges in that paper that the pooled estimator may be preferred in some settings because of the larger degrees of freedom. Since discussion of the matched-pair nature of the data comes down to assumptions about the underlying hypothetical assignment mechanism, many analyses that use randomization-based inference do treat the data as arising from matched pairs [6, 7].

On the other hand, many analyses of matched data in the statistical literature do not consider the paired nature of the data and instead treat the groups as two independent samples [8, 9]. As stated by Schafer and Kang [10], 'After the matching is completed, the matched samples may be compared by an *unpaired* t-test. ('Matching' erroneously suggests that the resulting data should be analyzed as if they were matched pairs. The treated and untreated samples should be regarded as independent, however, because there is no reason to believe that the outcomes of matched individuals are correlated in any way.)'. Recent simulation and empirical work by Hill and Reiter [11] show poor performance of the matched-pair variance estimator in general; I will defer to Dr Hill for further discussion of that work.

There are at least two reasons to believe that analyses of matched data do not need to account for the paired nature of the data. First, the theory behind propensity scores does not imply that any individual pairs will have similar covariate values—in fact, two individuals with the same propensity score may have very different values of the covariates. The theory of propensity scores says only that within *groups* of individuals with similar propensity scores, the *distributions* of the covariates that went into the propensity score will be similar. If matched-pair analyses will be done, arguably balance checks should also be done accounting for the matched pairs. Second, the theory underlying matching methods developed by Rubin and Thomas [9, 12] and Rubin and Stuart [13] does not rely on matched pairs—just matched samples. For example, the results showing the effects of matching on variance do not consider the individual pairings. These theoretical results are in the setting of ellipsoidally symmetric distributions (or mixtures of such distributions) and affinely invariant matching methods, but empirical work has shown that the results do hold much more broadly [12–14].

Some of the debate may come down to what researchers feel they need to account for in variance estimation; whether the matched sample can be conditioned on (and thus the covariates not modeled, as in Ho *et al.* [15]) or whether one needs to take into account the uncertainty in sampling from a population (as in Abadie and Imbens [16]). This general topic of variance estimation is certainly one that deserves further research, as demonstrated by the lack of consensus in the statistical literature.

CONCLUSIONS

I thank Dr Austin for this interesting and provocative paper. For applied researchers, it highlights the need for balance checks and careful consideration of outcome analyses. I also encourage applied researchers to think beyond simple 1:1 matching methods. Those methods are appealing because of their simplicity—of both implementation and description (and I in fact have used 1:1 matching many times)—but other methods may provide improved performance. More sophisticated methods, such as full matching [17], can make use of all of the available data and are optimal in terms of reducing bias (differences) in the propensity score. These methods are also becoming increasingly easy to use, with both R [18] and SAS [19] software available. Researchers should consider a variety of methods; one of the benefits of propensity score approaches is that a variety of methods can be attempted and the one that yields the best balance selected as the final choice [3].

Dr Austin's paper also provides a challenge to methodological researchers. Applied researchers wish to know 'best practices' for the use of propensity score methods in practice, but unfortunately clear advice does not yet exist. For example, further methodological research is needed on variance estimation after matching to clarify some of the issues brought up by Dr Austin. Similarly, I believe that part of the reason for a lack of sufficient balance checks in the applied literature is a lack of easy-to-use methods to assess balance. Methodological researchers should further develop and investigate appropriate balance measures for both univariate and multivariate comparisons.

In summary, the paper by Dr Austin, and the resulting discussion, highlights the need for further methodological work on the use of propensity scores in practice, and for further discussions such as this. A first step is to identify the areas of concern or question, which Dr Austin has done here. This sort of research will ultimately help applied researchers wade through the many propensity score methods available to determine the best approach for any particular analysis.

REFERENCES

1. Rosenbaum PR. Choice as an alternative to control in observational studies. *Statistical Science* 1999; **14**(3): 259–304.
2. Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology* 2001; **2**:169–188.
3. Stuart EA, Rubin DB. Best practices in quasi-experimental designs: matching methods for causal inference. In *Best Practices in Quantitative Methods*, Chapter 11, Osborne J (ed.). Sage Publications: Thousand Oaks, CA, 2007; 155–176.
4. Imai K, King G, Stuart EA. Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society*, *Series A* 2008; **171**:481–502.
5. Rubin DB. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* 1973; **29**:185–203.
6. Rubin DB. Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics* 1991; **47**:1213–1234.
7. Lu B, Zanutto E, Hornik R, Rosenbaum PR. Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association* 2001; **96**:1245–1253.
8. Dehejia RH, Wahba S. Causal effects in non-experimental studies: re-evaluating the evaluation of training programs. *Journal of the American Statistical Association* 1999; **94**:1053–1062.
9. Rubin DB, Thomas N. Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association* 2000; **95**:573–585.
10. Schafer JL, Kang JDY. *Average Causal Effects*: *A Practical Guide and Simulated Case Study*. Pennsylvania State University: State College, PA, 2007, unpublished manuscript.
11. Hill J, Reiter JP. Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine* 2006; **25**:2230–2256.
12. Rubin DB, Thomas N. Matching using estimated propensity scores, relating theory to practice. *Biometrics* 1996; **52**:249–264.
13. Rubin DB, Stuart EA. Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions. *The Annals of Statistics* 2006; **34**(4):1814–1826.
14. Barnard J, Frangakis CE, Hill JH, Rubin DB. Principal stratification approach to broken randomized experiments: a case study of school choice vouchers in New York City (with discussion and rejoinder). *Journal of the American Statistical Association* 2003; **98**(462):299–323.
15. Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 2007; **15**(3):199–236.
16. Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. *Econometrica* 2006; **74**(1):235–267.
17. Hansen BB. Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association* 2004; **99**:609–618.
18. Ho DE, Imai K, King G, Stuart EA. MatchIt: nonparametric preprocessing for parametric causal inference. R software package, 2007. http://gking.harvard.edu/matchit/ (accessed 14 November 2007).
19. Kosanke J, Bergstralh E. Vmatch: computerized matching of cases to controls using variable optimal matching. 2004. http://mayoresearch.mayo.edu/mayo/research/biostat/sasmacros.cfm (accessed 14 November 2007).