STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:2066–2069
Published online 20 March 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3243

REJOINDER

Discussion of 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003'

I would like to begin by thanking the editors of *Statistics in Medicine* for providing this forum to discuss my recent review of propensity-score matching in the medical literature. I would also like to thank Drs Hansen, Hill, and Stuart for their discussion of issues raised by my article and of other issues related to propensity-score matching in general. Hopefully, this focus on propensity-score matching will lead to improved practice in applied research and to the identification of areas of future research. My response to the discussants' comments is structured as follows: First, I will discuss issues concerning the comparison of balance between treated and untreated subjects in the propensity-score matched sample. Second, I will address issues relating to estimating the treatment effect and its statistical significance.

In my initial review [1], I described the need to assess the degree to which matching on the estimated propensity score resulted in a matched sample in which measured baseline covariates had a similar distribution (were balanced) between treated and untreated subjects. This allows one to assess whether matching on the propensity score has reduced or eliminated observed systematic differences between treated and untreated subjects. Matching on the true propensity score eliminates systematic differences between treated and untreated subjects. However, in non-randomized settings, the true propensity score is not known and must be estimated using the data. Balance tests can function as a test that the propensity-score model has been specified correctly. In my review, I asserted that the use of significance testing is inappropriate for assessing balance, as the resultant p-values are confounded with sample size. The matched sample is invariably smaller than the initial sample. This can result in the appearance of better balance in the matched sample solely due to the lower statistical power to detect imbalance. Both Drs Hill and Stuart concur with this proscription against the use of significance testing. Furthermore, both Drs Hill and Stuart suggest that rather than only examining means of continuous variables between treated and untreated subjects, applied researchers also examine balance in squared terms and interactions. Examining the variances of baseline variables between treated and untreated subjects can complement the comparison of means. In practice, examining all interactions between baseline variables could prove cumbersome. Further research is required into what constitutes acceptable balance in measured covariates between treated and untreated subjects in the matched sample. Currently, there are two options when researchers are faced with imbalance in the propensity-score matched sample. Researchers can modify the propensity-score model by including additional variables, higher order terms, or interactions. Alternatively, researchers can change the caliper width used in matching. By using a narrower caliper for matching, researchers may be able to achieve better balance in the matched sample. Arguably, each of these approaches could result in improved balance in the matched sample. The relative advantage of each approach requires further examination. Furthermore, there is a paucity of research into the relative strengths of different methods used in the medical literature for forming matched pairs. In a recent study, I have compared, both

empirically and via Monte Carlo simulations, the relative performance of different methods for forming matched pairs [2].

Dr Hansen is the lone dissenter on the prohibition against the use of significance testing to assess balance. He cites Cochran [3, Section 3.1] as a justification for the use of significance testing to compare the distribution of baseline covariates between treated and untreated subjects in the matched sample. However, in Section 3.1 of the cited reference, Cochran is not discussing balance in matched studies, but in observational studies in general. The proscription against the use of significance testing for assessing balance that was provided in my initial article applies only in the context of matched samples drawn from a larger sample. The prohibition was not intended to be generalized to all observational studies in general. Furthermore, Cochran, in the same section, states that '*although these checks on the x distribution are usually made by tests of significance*, *it is not clear what kind of assurance is given by the finding of a non-significant result*, *nor that a test is the appropriate criterion*' (italics mine) [3, Section 3.1, p. 242, lines 12–14]. Furthermore, some researchers erroneously fail to distinguish between confounding and significance testing. Confounding, as defined by Hill above (see Section 2 of Hill's discussion), occurs when a variable exists that predicts both treatment and the outcome. Omitting this variable from a regression model, or allowing this variable to be imbalanced between treatment groups, can result in biased estimation of the treatment effect. This effect of confounding is independent of sample size or of having sufficient statistical power to detect the presence of confounding.

The second major area of discussion centers on the need to account for the matched nature of the sample when estimating the treatment effect and its statistical significance. Among the discussants, Hansen is silent on this issue, whereas Stuart reflects ambivalence about whether creating matched pairs creates a lack of independence in the matched sample. Hill acknowledges that propensity-score matching induces a lack of independence in the sample. The reason for this is as follows. Propensity-score matching does not guarantee that matched subjects will have identical covariate values. However, within strata of subjects matched on the propensity score, subjects will have the same multivariate distribution of baseline covariates [4]. Therefore, *on average* matched subjects will be more similar, in terms of baseline characteristics, than unmatched subjects. Since the baseline covariates are also associated with the outcome (otherwise there would be no confounding), then, *on average*, matched subjects will have outcomes that are more similar than do unmatched subjects.

While Hill and I agree on the lack of independence that is induced by propensity-score matching, our disagreement centers on the nature of the post-matching analysis. Hill writes that 'the options available for estimating treatment effects should be similar to those available in a randomized setting'. I fully agree with this statement. Indeed, it is the ability of propensity-score matching to mimic some of the characteristics of a randomized controlled trial (RCT) that led me to advocate for the use of simple measures of treatment effect such as risk differences and relative risks. The *British Medical Journal* (*BMJ*) requires that, when reporting the results of clinical trials, the following information be included in both the study abstract and the results section: the absolute event rates among experimental and control groups, the relative risk reduction, and the number needed to treat or harm (NNT or NNH) along with its 95 per cent confidence interval [5] (this assumes a dichotomous outcome in the trial). The number needed to treat requires calculating the absolute risk reduction due to the treatment.

Hill advocates for the use of post-matching covariance adjustment to account for residual bias and to improve precision. However, this approach may lead one away from methods that are commonly employed in RCTs. While several authors in the statistical literature have advocated for

DOI: 10.1002/sim

covariance adjustment in RCTs [6–10], my impression is that, in RCTs in the medical literature, unadjusted analyses are more common than adjusted analyses. Furthermore, when estimating a linear treatment effect, the difference in means (the unadjusted treatment effect) and the adjusted difference in means (the adjusted treatment effect) will, on average, coincide. This is because the difference in means is a collapsible estimator, with the marginal (or population-average) effect and the adjusted (or conditional) effect being equal. Recent systematic reviews have demonstrated that, in the medical literature, propensity-score methods are used overwhelmingly to estimate the effect of treatments on dichotomous or time-to-event outcomes [11, 12]. Regression adjustment in such a context generally involves the use of logistic regression or Cox regression models. However, neither the odds ratio nor the hazard ratio is collapsible [13]. In recent research, it was demonstrated that propensity-score methods result in biased estimation of conditional odds ratios and conditional hazard ratios [14]. Furthermore, their use results in suboptimal inferences about marginal odds ratios [15]. However, propensity-score matching results in unbiased estimation of relative risks [16]. This last finding complements Rosenbaum and Rubin's result that conditioning on the propensity score allows for unbiased estimation of linear treatment effects (e.g. differences in means and risk differences) [4]. Differences in means, risk differences, and relative risks (in the absence of effect modifications) are collapsible measures of treatment effect, whereas the odds ratio and hazard ratio are not [13, 17]. My recommendation for the use of measures of effect such as risk differences and relative risks reflects my desire to employ measures of treatment effect that are collapsible and that better reflect measures that are more commonly employed in RCTs in the medical research. Importantly, the estimated risk difference or relative risk is the same whether or not one accounts for the matched sample. It is only the variance estimation that will differ depending on whether one accounts for the paired nature of the sample. In a related unpublished study, I demonstrate that there is never any benefit to conducting an unmatched unadjusted analysis compared with conducting an unadjusted analysis that accounts for the matched-pairs nature of the sample (manuscript submitted).

Dr Hill addresses other issues relating to the use of propensity-score methods in the applied literature. I agree that we must be vigilant against those who would wish to infer that propensity-score methods allow one to control for unmeasured confounders. I have addressed this issue in a prior study comparing the use of propensity-score methods with clinical and administrative data [18] and in a study examining variable selection for propensity-score models [19]. An important issue raised by Dr Hill relates to the estimand and the question of who we are making inferences about. While this issue has received greater attention in the econometrics literature, it has received less attention in the biostatistical and medical literature. Greater translational efforts are necessary in this area.

In summary, I would make the following suggestions for studies that employ propensity-score matching. First, the methods used to form the matched sets should be explicitly described, allowing other researchers to replicate the study methods. Second, applied researchers should explicitly verify that matching on the estimated propensity score has resulted in a matched sample in which measured baseline variables are balanced between treated and untreated subjects. Baseline balance should be assessed using methods such as standardized differences that are not influenced by sample size. Third, when feasible, measures of treatment effect similar to those reported in RCTs should be employed. Finally, the statistical significance of these measures of effect should account for the lack of independence induced by propensity-score matching. I would like to conclude by reiterating my thanks to the editors of *Statistics in Medicine* for facilitating this discussion, and by extending my thanks to Drs Hansen, Hill, and Stuart for the stimulating discussion that has

DOI: 10.1002/sim

been engendered. Hopefully, this discussion will continue, and new avenues of research will be pursued, leading to improved application of statistical methods in observational research.

ACKNOWLEDGEMENTS

I would like to thank Robert Platt and Thérèse Stukel for reviewing and commenting on a draft of this response.

PETER C. AUSTIN
Institute for Clinical Evaluative Sciences
Toronto, Ont., Canada M4N 3M5

REFERENCES

1. Austin PC. A critical appraisal of propensity score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*. DOI: 10.1002/sim.3150.
2. Austin PC. The performance of different propensity-score matching methods used in the medical literature. Under review.
3. Cochran WG. The planning of observational studies of human populations. *Journal of the Royal Statistical Society, Series A* (*General*) 1965; **128**:234–266.
4. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**:41–55.
5. http://resources.bmj.com/bmj/authors/types-of-article/research (accessed 11 January 2008).
6. Altman DT, Dore CJ. Randomisation and baseline comparisons in clinical trials. *The Lancet* 1990; **335**:149–153.
7. Senn S. Testing for baseline balance in clinical trials. *Statistics in Medicine* 1994; **13**:1715–1726.
8. Senn S. Baseline comparisons in randomized clinical trials. *Statistics in Medicine* 1991; **10**:1157–1160.
9. Rothman KJ. Epidemiologic methods in clinical trials. *Cancer* 1977; **39**:1771–1775.
10. Altman DG, Dore CJ. Baseline comparisons in randomized clinical trials. *Statistics in Medicine* 1991; **10**:797–802.
11. Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods give similar results to traditional regression modeling in observational studies: a systematic review. *Journal of Clinical Epidemiology* 2005; **58**:550–559.
12. Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology* 2006; **59**:437–447.
13. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 1984; **7**:431–444.
14. Austin PC, Grootendorst P, Normand SLT, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Statistics in Medicine* 2007; **26**:754–768.
15. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in Medicine* 2007; **26**:3078–3094.
16. Austin PC. The performance of different propensity score methods for estimating relative risks. *Journal of Clinical Epidemiology*. DOI: 10.1016/j.clinepi.2007.07.011.
17. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology* 1987; **125**:761–768.
18. Austin PC, Mamdani MM, Stukel TA, Anderson GM, Tu JV. The use of the propensity score for estimating treatment effects: administrative versus clinical data. *Statistics in Medicine* 2005; **24**:1563–1578.
19. Austin PC, Grootendorst P, Anderson GM. 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* 2007; **26**:734–753.