STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:2050–2054
Published online 14 April 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.3208

COMMENTARY

The essential role of balance tests in propensity-matched observational studies: Comments on ‘A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003’ by Peter Austin, *Statistics in Medicine*

B. B. Hansen[*, †]

Statistics Department, University of Michigan, 439 West Hall, Ann Arbor, MI 48109-1107, U.S.A.

Peter Austin has made an exacting, timely and eye-opening review of uses of propensity-score matching in medical research. Its Section 2.1 argues that the reports of propensity-matched analyses should include descriptive assessments of matched treatment-control differences on baseline variables. When propensity matching on covariates including BMI, for example, one should report the difference between matched cohorts’ mean BMIs, perhaps after inverse scaling by the pooled s.d. of BMIs prior to matching. The recommendation is a good one: matched differences on prognostic variables, and on variables that track selection into treatment, speak to the credibility of subsequent matched outcome analyses; and although the basic promise of propensity matching is that it should lessen such differences, the extent of the reduction varies greatly from case to case. Furthermore, since successful propensity matches or subclassifications enable comparisons similar to those which randomization would have given—in terms of *observed* covariates, at least [1], and should those covariates jointly suffice to remove confounding, then also in terms of outcomes [2]—it follows that balance is the basic mark of success of a propensity adjustment.

Austin’s review also makes a negative recommendation: When appraising balance, avoid significance tests. Having compared means of BMI and other variables, Austin would not have us go on to calculate either paired/two-sample t-tests or other tests for a treatment-control difference on BMI. His pessimism about the state of reporting in the medical propensity-matching literature stems in large part from this opinion; only 2 of 47 papers reported balance properly, Austin reports, but it turns out that another 33 were disqualified on the basis of having tested balance, rather than reporting it using purely descriptive measures. This dim view of balance testing is driven by two complaints, complaints Austin shares with Imai, King and Stuart [3]:

A. Null hypotheses typically refer to populations from which the sample was taken, yet the matched sample need not represent any background population, and if it does the balance within that population is not relevant to matched inferences based on the sample.

[*] Correspondence to: B. B. Hansen, Statistics Department, University of Michigan, 439 West Hall, Ann Arbor, MI 48109-1107, U.S.A.

[†] E-mail: ben.b.hansen@umich.edu

B. The sample-size reduction that accompanies matching can reduce the significance of baseline imbalances even when the imbalances remain the same in absolute terms.

Complaint A applies to some, but not all, balance tests. Among those to which it does not apply are the permutation tests discussed in this comment and elsewhere [4]. Complaint B is correct, but less paradoxical than it may seem. Matching aims to ensure the coverage of tests and confidence intervals for treatment effects. Ordinarily it would do this by reducing bias, but in those cases where matching reduces baseline imbalances' statistical significance but not their magnitudes, it does it by increasing standard errors. In all cases, assessing imbalances in terms of their statistical significance better promotes coverage than does assessing them in absolute terms.

Balance tests in observational studies: The use of tests to compare non-randomized treatment and control groups pre-dates propensity scores, going back at least to Cochran [5, Section 3.1], who argued that two-sample t-tests and statistics were particularly well suited to decide whether between-group differences on an x-variable were enough to necessitate adjustments for it. In Cochran's analysis, small t-statistics suggest bias small enough that the coverage of confidence intervals for outcome differences are little affected, whereas large t-statistics open the door to more damaging biases; this pattern holds irrespective of sample size. The same is true after propensity matching, against a more general model than Cochran's, and for balance statistics the calibration of which does not presuppose sampling from a hypothetical population. Non-significance of imbalance on a given baseline x suggests of one potential bias in causal effect estimates that it is too small to effect type I errors, whereas significance of the same imbalance suggests a bias large enough to inflate type I errors. This is so irrespective of sample size, and the same cannot be said for any measure of balance that is independent of sample size.

Viewed as significance tests, rather than informal diagnostics, balance assessments are simply goodness-of-fit tests of a certain kind. Cochran took them to be comparisons of background populations of treatment and control subjects, but there is another interpretation that better fits propensity analysis. Let $z \in \{0,1\}$ indicate assignment to the treatment group, let $x_1, \ldots, x_K$ be measured covariates, and let y be an outcome measurement. Propensity matching posits that treatment assignments z are independent draws of random variables $Z \sim \text{Bernoulli}(\pi(x_1, \ldots, x_K))$, π an unknown function, grouping the subjects into matched sets on the basis of estimated values of π. However these values are estimated, and however closely or loosely subjects were matched on the estimates, outcome analysis proceeds under the model that $\pi_{si} = \pi_{sj}$ whenever i, j belong to a common matched set s. Properly conducted balance tests probe the goodness of this model's fit to the data.

Such a model, call it $\tilde{H}_0$, is almost sure to be false, at least in detail. Being mindful of Box's advice—'all models are false, but some are useful' [6, p. 202]—we need not be deterred by this, only made wary of matches of which $\tilde{H}_0$ is so false as to be misleading. Let $\tilde{H}_\mathrm{a}$ stand for the actual state of affairs in a given observational study, where perhaps $\pi_{si} \doteq \pi_{sj}$ but typically $\pi_{si} \neq \pi_{sj}$ for some, perhaps all, s and $i \neq j$. Our hope is that by working under $\tilde{H}_0$, rather than $\tilde{H}_\mathrm{a}$, little harm is done to tests of hypotheses about treatment effects. Tests of matched differences at baseline show us whether the substitution does violence to a test in which the correct answer is known, since the treatment cannot have affected measurements that preceded it.

The remainder of this note adapts Cochran's argument to the setting of propensity modeling, which takes assignment to treatment, but not necessarily the generation of outcomes, to be stochastic.

Balance tests without superpopulation sampling: Assume matching to have produced m non-overlapping pairs, $p=1,\ldots,m$. Randomization-based tests of $\tilde{H}_0$ based on the mean paired difference in a baseline variable x, $d(\mathbf{z},\mathbf{x})=m^{-1}\sum_p (z_{p1}-z_{p2})(x_{p1}-x_{p2})$, have $\mathbf{Z}$ as their only random variable. It is natural to condition on $\mathscr{C}=\{(Z_{p1},Z_{p2})=(1,0) \text{ or } (0,1), p=1,\ldots,m\}$, making $d(\mathbf{Z},\mathbf{x})$ a sum of known linear transforms of Bernoulli trials. Then the mean difference $d(\mathbf{z},\mathbf{x})$ is scaled by the square root of its variance under $\tilde{H}_0$, which usually differs somewhat from the estimated variance of $d(\mathbf{z},\mathbf{X})$ under models taking x's to be drawn from a population; according to $\tilde{H}_0$, it is more nearly N(0, 1) than t-distributed. This randomization test can be generalized, first to multiple-controls matched, full-matched [7, 8] and subclassified data, and then for all of these designs to the situation where imbalance along any number of x's is to be tested [4]. These tests meet Austin's and Imai *et al.*'s stipulation that the balance assessment not presuppose sampling from a superpopulation.

Let v be a variable not affected by treatment assignment. How does $d(\mathbf{Z},\mathbf{v})$'s distribution under $\tilde{H}_0$ compare with its distribution under $\tilde{H}_\mathrm{a}$? The answer depends on how closely subjects are matched on the propensity score and the relationship of v's to propensity scores, in a manner with interesting ramifications for causal inference. Write θ for the logit-scale propensity score, i.e. $\theta_{pi}\equiv \operatorname{logit}(\pi_{pi})$. Note that θ denotes a true, typically unknown, propensity score, not an estimate of it. Then $\Pr(Z_{p1}=1|\mathscr{C})$ is not $\pi_{p1}=\{1+\exp(-\theta_{p1})\}^{-1}$ but, by the conditioning on $\mathscr{C}$, $\{1+\exp(\theta_{p2}-\theta_{p1})\}^{-1}$, from which it follows that

$$\mathbf{E}[d(\mathbf{Z},\mathbf{v})|\mathscr{C}]=m^{-1}\sum_p (v_{p1}-v_{p2})\tanh\{(\theta_{p1}-\theta_{p2})/2\}$$

$$V[d(\mathbf{Z},\mathbf{v})|\mathscr{C}]=m^{-2}\sum_p (v_{p1}-v_{p2})^2\operatorname{sech}^2\{(\theta_{p1}-\theta_{p2})/2\}$$

However, $\operatorname{sech}(x)\leqslant 1$, with $\operatorname{sech}(x)=1$ iff $x=0$; and after matching on $\hat{\theta}$, the differences $\{|\theta_{p1}-\theta_{p2}|: p\leqslant m\}$ should be small enough to warrant the first-order Taylor approximation $\tanh(x)\doteq x$. With some algebra,

$$\mathbf{E}[d(\mathbf{Z},\mathbf{v})|\mathscr{C}]\doteq s(\mathbf{v},\theta|\mathscr{C}) \tag{1}$$

and

$$V[d(\mathbf{Z},\mathbf{v})|\mathscr{C}]\leqslant 2s^2(\mathbf{v}|\mathscr{C})/m \tag{2}$$

where $s(\mathbf{v},\mathbf{w}|\mathscr{C})=m^{-1}\sum_{p=1}^{m}\sum_{i=1,2}(v_{pi}-\bar{v}_p)(w_{pi}-\bar{w}_p)$ and $s^2(\mathbf{v}|\mathscr{C})=s(\mathbf{v},\mathbf{v}|\mathscr{C})$. The bound in (2) is attained only under $\tilde{H}_0$; the true ($\tilde{H}_\mathrm{a}$) variance is inevitably smaller.

Write $\mathbf{E}_0, V_0$ for expectation and variance under $\tilde{H}_0$ and conditional on $\mathscr{C}$; $\mathbf{E}_a, V_a$ for the same conditional expectation and variance under $\tilde{H}_\mathrm{a}$. In the vicinity of the approximating model, $\tilde{H}_0$, the power of a test of balance on x to reject $\tilde{H}_0$ is determined by

$$\frac{\mathbf{E}_a d(\mathbf{Z},\mathbf{x})-\mathbf{E}_0 d(\mathbf{Z},\mathbf{x})}{V_0^{1/2} d(\mathbf{Z},\mathbf{x})}\doteq(0.707)m^{1/2}\frac{s(\mathbf{x},\theta|\mathscr{C})}{s(\mathbf{x}|\mathscr{C})} \tag{3}$$

$$=(0.707)m^{1/2}r_{x,\theta|\mathscr{C}}s(\theta|\mathscr{C}) \tag{4}$$

where $r_{x,\theta|\mathscr{C}}$ is the partial correlation $s(\mathbf{x},\theta|\mathscr{C})/(s(\mathbf{x}|\mathscr{C})s(\theta|\mathscr{C}))$ of x and θ within matched sets. The power of the test increases with the number of matched sets, with the magnitude of differences between matched subjects' propensity scores, and with the residual correlation of x and the propensity score.

Distinguishing harmful from harmless failures of $\tilde{H}_0$: Randomization tests of hypotheses about the causal effect also assume $\tilde{H}_0$. To test the strict null of no treatment effect on y, one adds to $\tilde{H}_0$ the assumptions that $y_c = y_t = y_{\text{obs}}$ and that there is no hidden bias [9], assessing $d(\mathbf{z}_{\text{obs}}, \mathbf{y}_{\text{obs}})$ against the $\tilde{H}_0$-law of $d(\mathbf{Z}, \mathbf{y}_{\text{obs}})$. Under these assumptions, (1) and (2) apply to $y_{\text{obs}} = y_c$, so that if

$$\frac{\mathbf{E}_a d(\mathbf{Z},\mathbf{y}_c) - \mathbf{E}_0 d(\mathbf{Z},\mathbf{y}_c)}{V_0^{1/2} d(\mathbf{Z},\mathbf{y}_c)} \doteq (0.707) m^{1/2} r_{y_c,\theta|\mathscr{C}} s(\theta|\mathscr{C}) \tag{5}$$

is small then the bias of $d(\mathbf{Z},\mathbf{y}_c)$ will be small—not necessarily in absolute terms, but certainly *relative to its standard error*, so that the bias won't undermine the test for a treatment effect. In parallel with (4), (5) says that the bias-to-standard-error ratio for effect estimation increases with sample size, with the magnitude of matched subjects' differences on θ, and with the residual correlation of y_c and the propensity score.

This similarity has several consequences. The first is a justification for balance tests' dependence on sample size. The presence of the same $m^{1/2}$ factor in (4) and (5) shows that if the samples were randomly reduced, balance tests' power would indeed diminish, but the decrease would be in perfect alignment with an improvement in $\tilde{H}_0$'s capacity to adequately approximate $\tilde{H}_{\text{a}}$. Another is that it links the prognostic value of a covariate to balance testing and the bias of treatment effect estimation. Suppose that x correlates with y_c: then in (3), the matched correlation of x with θ may approximate that of y_c with θ; if it is large enough to make (3) large, giving power to the balance test, then this suggests that (5) is large, perhaps large enough to skew causal inferences. As a third consequence of the similarity between (4) and (5), suppose that x is notable as a predictor of treatment assignment: x is a selection covariate. Note that $r_{x,\theta|\mathscr{C}} s(\theta|\mathscr{C})$ is bounded in magnitude by $s(\theta|\mathscr{C})$, with the bound achieved only when $r_{x,\theta|\mathscr{C}} = \pm 1$. If the matched correlation of x and θ is high, then (3) approximates $(m/2)^{1/2} s(\theta|\mathscr{C})$, which is also an upper bound for (5); if the test for balance on x lacks power, then the bias must be small. In sum, for both prognostic and selection covariates, balance tests tend to reject when bias due to inexact propensity matching is enough to undermine causal inferences and tend not to reject when that bias is small enough to be ignored. Therein lies their value.

ACKNOWLEDGEMENTS

Jake Bowers, Kosuke Imai, Paul Rosenbaum and Jasjeet Sekhon generously commented on drafts of this note.

REFERENCES

1. Braitman LE, Rosenbaum PR. Rare outcomes, common treatments: analytic strategies using propensity scores. *Annals of Internal Medicine* 2002; **137**(8):693–695.
2. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**:41–55.
3. Imai K, King G, Stuart E. Misunderstandings among experimentalists and observationalists: about causal inference. *Journal of the Royal Statistical Society, Series A* 2008; **171**(Part 2, Forthcoming):1–22.

4. Hansen BB, Bowers J. Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science* 2008; **23**. To appear.
5. Cochran WG. The planning of observational studies of human populations. *Journal of the Royal Statistical Society* 1965; **128**:234–266.
6. Box GE. Robustness in the strategy of scientific model-building. In *Robustness in Statistics*, Lauer RL, Wilkinson GN (eds). Academic Press: New York, 1979; 201–236.
7. Rosenbaum PR. A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society* 1991; **53**:597–610.
8. Hansen BB, Klopfer SO. Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics* 2006; **15**(3):609–627.
9. Rosenbaum PR. *Observational Studies* (2nd edn). Springer: Berlin, 2002.