Published in final edited form as:

Pharmacoepidemiol Drug Saf. 2011 March ; 20(3): 317–320. doi:10.1002/pds.2074.

The role of the *c*-statistic in variable selection for propensity score models

Daniel Westreich^{*}, Stephen R. Cole, Michele Jonsson Funk, M. Alan Brookhart, and Til Stürmer

Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC

Abstract

The applied literature on propensity scores has often cited the *c*-statistic as a measure of the ability of the propensity score to control confounding. However, a high *c*-statistic in the propensity model is neither necessary nor sufficient for control of confounding. Moreover, use of the *c*-statistic as a guide in constructing propensity scores may result in less overlap in propensity scores between treated and untreated subjects; this may require the analyst to restrict populations for inference. Such restrictions may reduce precision of estimates and change the population to which the estimate applies. Variable selection based on prior subject matter knowledge, empirical observation, and sensitivity analysis is preferable and avoids many of these problems.

Keywords

Propensity scores; *c*-statistic; variable selection; confounding

The use of propensity scores to reduce confounding bias in non-experimental studies has increased dramatically¹ since their introduction by Rosenbaum and Rubin.² The propensity score is the predicted probability of treatment (alternatively, exposure) conditional on selected covariates, and is used as part of a two-stage analytic process. First, the propensity score is estimated from the data, typically¹, (although not always^{3–4}) using logistic regression. Second, the effect of the treatment on the outcome is assessed in persons with the same estimated propensity score,^{5–7} for example by matching on propensity score,² stratification,^{2,8} or by weighting by a function of the propensity score.⁹

Treated and untreated subjects with the same propensity score have measured baseline covariates that come from the same distribution. Assuming no unmeasured confounders, conditioning on the propensity score allows one to obtain an unbiased estimate of the average treatment effect at that value of the propensity score.

In this commentary, we discuss the estimation of the propensity score itself and the use (and misuse) of the *c*-statistic in this process. The widespread reporting of the *c*-statistic from the propensity score model may suggest an underlying misconception about the goal of the propensity score.

As this work comprises only methodological commentary, ethics approval was not sought.

^{*}Corresponding author: D. Westreich, Duke University Medical Center, Department of Obstetrics and Gynecology, Box 3084 Medical Center, Durham, NC 27710 daniel.westreich@duke.edu.

Conflicts None declared

Propensity scores: goals and variable selection

The purpose of the propensity score is to eliminate confounding bias, which occurs when risk factors for the outcome are unequally (alternatively, non-randomly) distributed among treatment groups. Thus, to control for confounding using propensity scores, the epidemiologist estimates a score which attempts to achieve an equal distribution of observed risk factors for the outcome between treated and untreated subjects: that is, to balance those risk factors between treatment groups. The goal of the propensity score is *not* to predict treatment as well as possible. Balancing covariates so as to control confounding and the prediction of treatment are separate goals that require different considerations for variable selection. Thus, for example, a propensity score should not attempt to balance covariates unrelated to the outcome (e.g. instruments), as these variables do not lead to confounding.^{10–11} Nor is the use a single propensity score model to examine the association of a single treatment with multiple outcomes^{12–14} necessarily advisable.

Epidemiologists tend to view propensity scores (not unjustly) as an efficient means to control for a large number of covariates possibly even when an outcome is rare.^{15–16} Perhaps as a result, the literature on selection of covariates for inclusion in propensity scores is less well developed than the literature on selection of covariates for inclusion in outcome regression analysis. Nonetheless, several guidelines have emerged. Leaving a confounder out of a propensity score model will result in bias in the final effect estimate similar (in magnitude and direction) to the bias observed when this confounder is left out of an outcome regression model.¹¹ The inclusion of covariates in the propensity score which are not associated with the disease outcome (including instrumental variables) decreases precision of the treatment effect estimate without any advantage with respect to bias¹⁰ and may increase bias in the presence of unmeasured confounders.^{17–18} The inclusion of risk factors for the outcome not associated with treatment increases precision of the treatment effect estimate.¹⁰ Although not specifically assessed for propensity scores, the inclusion of colliders in a model can lead to collider stratification bias.^{19–20} These issues are not unique to propensity score models, although the instrumental variable issue will tend to be more relevant to propensity scores if they are misunderstood as treatment prediction scores. Given these guidelines, does the use of the *c*-statistic aid epidemiologists in creating propensity scores which achieve the goal of propensity scores: namely, to eliminate bias due to measured confounders?

The *c*-statistic

The *c*, or concordance, statistic is a measure of the discriminatory power of a predictive model, and is equivalent to the area under the receiver operating characteristic curve.^{21–22} For a model for a dichotomous variable (in the propensity score setting, a treatment), the *c*-statistic is calculated as the proportion of all possible pairs of subjects comprising one treated subject (A) and one untreated subject (B), in which the predicted probability of the outcome is higher in the treated than the untreated (higher in A than in B).²² The *c*-statistic takes on values between 0.5 (classification no better than a coin flip) to 1.0 (perfect classification), and can be calculated for any method that generates predicted values, including logistic regression and machine learning classification.³

The *c*-statistic is often cited as a measure of the “fit” of a propensity score model; that is, the ability of the model to predict treatment status using observed covariates. Among two large reviews of the propensity score literature, 91/224 articles (41%) reviewed reported a *c*-statistic or an equivalent.^{1,23} In each review, several studies reported *c*-statistics greater than 0.90, indicating very good ability of the propensity score model to predict treatment status. Indeed, the frequent reporting of the *c*-statistic suggests that many investigators view

propensity score estimation as a predictive model for the treatment. Not discussed is whether a high (or low) c -statistic in a propensity score model gives us any information on whether that propensity score model achieves its goal.

Propensity and positivity

If the c -statistic is used as a guide for variable selection into a propensity score model, it may lead to the inclusion of useless or even harmful variables in that model. In particular, the inclusion of covariates strongly related to treatment but unrelated to the outcome will increase the c -statistic and thus be preferentially included in the model; but the inclusion of such variables will lead to distributions of propensity scores with relatively little overlap between the treated and the untreated²³. Because the treatment-outcome effect is estimated in persons with the same propensity score, data that fall outside a common range of the propensity score distributions in treated and untreated are typically lost for the second stage of a propensity score analysis: either because these individuals cannot be matched, or because they are specifically excluded from further analysis.

We exclude subjects in the non-overlapping tails of the propensity score distribution because treatment effect cannot be estimated without variation of treatment given the propensity score. Formally, *positivity* is violated in these subjects.²⁴ Positivity requires that there are both treated and untreated subjects at every level of all covariates under consideration, and is one of the key assumptions for causal inference.^{24–25}

Propensity score distribution overlap is often considered in terms of treatment-stratified histograms or kernel-smoothed density estimates. If propensity score density estimates do not overlap, it is likely that there is non-positivity for some combination of covariates. (However, positivity is *not* guaranteed even if propensity score distributions fully overlap; smoothed curves may obscure regions of non-positivity.) Identification of populations never- or always-treated may be one of the main advantages of propensity scores¹⁶; regions of non-overlap are often trimmed in propensity scores.¹⁶ Trimming of propensity score curves reduces sample size and thus precision, and also changes the population in which inference is being made in complex ways. Consistent with this, Austin et al. found that as the c -statistic (area under the ROC curve) increased, the number of matched pairs for analysis decreased.²⁶ Creating unnecessary non-overlap by inclusion of unnecessary covariates (i.e., non-confounders) in propensity score models should be avoided on both counts.

Using a c -statistic to guide propensity score modeling is therefore ill-advised. Perhaps more to the point, a high c -statistic in the propensity score model is neither *necessary* nor *sufficient* for the control of confounding. Imagine a propensity score estimated in a randomized trial in which all risk factors for the outcome are perfectly balanced between treatment arms. The propensity score model built with these risk factors will have a c -statistic of 0.5: risk factors do not help predict the treatment assignment. But given perfect balance, there will be no confounding bias; thus a high c -statistic is not necessary. Conversely, a high c -statistic can be achieved by the inclusion of a strong instrument, independent of all confounders; thus, a high c -statistic is not sufficient.

Correspondingly, Weitzen et al. found in a simulation study that the c -statistic “had no relationship with residual confounding in...treatment effect estimates.”²⁷ Austin et al. reiterated that “there was no clear relationship between the [c -statistic] of a given propensity score model and the degree to which conditioning on the propensity score balanced prognostically important variables between treated and untreated subjects in the matched sample.”²⁶ In a third report, Austin argued that the c -statistic gives no indication as to

whether confounders have been omitted from the propensity-score model, nor as to whether the propensity-score model has been correctly specified.²⁸

Remarks

Neither *necessary nor sufficient* to ensure the control of confounding, the *c*-statistic is of limited value for covariate selection into a propensity score model and provides no certainty that all measured confounders have been balanced between treatment groups, or that interactions among covariates or higher order terms have been balanced. Indeed, preference for covariates which yield a high *c*-statistic may lead us to balance non-confounders including instrumental variables, resulting in reduced overlap of propensity score distributions between treatment arms, and reduced efficiency. This discussion is largely compatible with previous recommendations,^{2,26–29} and in general these cautions apply to other markers of model fit including the Hosmer-Lemeshow goodness-of-fit test statistic²⁷ as well as to automated model fitting procedures which rely on the *c*-statistic.

Unfortunately, the determination of what is a risk factor for the outcome is not always straightforward and there is no infallible way to separate risk factors from non-risk factors in a given study.¹⁷ While some have argued in favor of selecting covariates according to their empirical association with the outcome, more simulation studies are necessary before this approach can be recommended without reservation. Model selection approaches, such as those based on cross-validation^{30–31}, may also be able to improve the performance of a given propensity score model; however, currently these approaches require that the initial model yields a consistent estimator. Thus, these methods still require the analyst to decide which variables to include in the initial model.

Therefore, rather than letting the *c*-statistic guide selection of variables into the propensity score model, we recommend that selection of covariates into a propensity score model begin with analysis of causal diagrams based on prior subject-matter knowledge and hypotheses³². From there, the literature supports the inclusion of strong risk factors for the outcome, whether or not they are related to the exposure, and the exclusion of variables that are strong predictors for exposure but with no obvious relation to the outcome. However, as no method for variable selection is foolproof, the epidemiologist should consider reporting various results under different, sensible, and transparently explained model specifications. As in conventional analyses, such a sensitivity analysis approach allows the analyst and the reader to assess the sensitivity of the results to model assumptions that are not supported by strong subject-matter knowledge.¹⁸

Take-home messages

- The purpose of the propensity score is to balance risk factors for the outcome, in order to eliminate confounding bias
- A high *c*-statistic in the propensity model is neither necessary nor sufficient for control of confounding; reliance on the *c*-statistic in selecting a propensity score model may provide false confidence that all confounders have been balanced between treatment groups
- Variable selection for propensity scores should be based on prior subject matter knowledge and empirical observation

Acknowledgments

Funding. D.W. received funding from an unrestricted educational training grant from the UNC-GlaxoSmithKline Center of Excellence in Pharmacoepidemiology and Public Health, UNC School of Public Health; NIH/NIAID grant T32-AI-07001; and NIH/NICHHD K99-HD-06-3961. M.J.F. is supported by a career development award from the Agency for Healthcare Research & Quality (AHRQ K02 HS17950) and an unrestricted grant from the UNC-GSK Center of Excellence in Pharmacoepidemiology and Public Health. M.A.B. is supported by a career development grant from the National Institute on Aging (AG-027400). T.S. received investigator-initiated research funding and support as Principal Investigator (RO1 AG023178) and Co-Investigator (RO1 AG018833) from the National Institute on Aging at the National Institutes of Health, and as Principal Investigator of the UNC-DEcIDE center from the Agency for Healthcare Research and Quality. T.S. does not accept personal compensation of any kind from any pharmaceutical company, though he receives salary support from unrestricted research grants from pharmaceutical companies to UNC.

References

1. 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. *J Clin Epidemiol.* 2006; 59(5):437–47. [PubMed: 16632131]
2. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983; 70(1):41–55.
3. Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *J Clin Epidemiol.* 2010
4. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med.* 2010; 29(3):337–46. [PubMed: 19960510]
5. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med.* 1998; 17(19):2265–81. [PubMed: 9802183]
6. Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. *Am J Epidemiol.* 1999; 150(4):327–33. [PubMed: 10453808]
7. Stürmer T, Schneeweiss S, Brookhart MA, Rothman KJ, Avorn J, Glynn RJ. Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: nonsteroidal antiinflammatory drugs and short-term mortality in the elderly. *Am J Epidemiol.* 2005; 161(9):891–8. [PubMed: 15840622]
8. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med.* 1997; 127(8 Pt 2):757–63. [PubMed: 9382394]
9. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health.* 2006; 60(7):578–86. [PubMed: 16790829]
10. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. *Am J Epidemiol.* 2006; 163(12):1149–56. [PubMed: 16624967]
11. Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics.* 1993; 49:1231–1236.
12. Almond CS, Gauvreau K, Thiagarajan RR, et al. Impact of ABO-incompatible listing on wait-list outcomes among infants listed for heart transplantation in the United States: a propensity analysis. *Circulation.* 2010; 121(17):1926–33. [PubMed: 20404257]
13. Seliger S, Fox KM, Gandra SR, et al. Timing of erythropoiesis-stimulating agent initiation and adverse outcomes in nondialysis CKD: a propensity-matched observational study. *Clin J Am Soc Nephrol.* 2010; 5(5):882–8. [PubMed: 20299377]
14. Subramaniam GA, Ives ML, Stitzer ML, Dennis ML. The added risk of opioid problem use among treatment-seeking youth with marijuana and/or alcohol problem use. *Addiction.* 2010; 105(4):686–98. [PubMed: 20403020]
15. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol.* 2003; 158(3):280–7. [PubMed: 12882951]

16. Glynn RJ, Schneeweiss S, Stürmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic Clin Pharmacol Toxicol*. 2006; 98(3):253–9. [PubMed: 16611199]
17. Robins JM. Data, Design, and Background Knowledge in Etiologic Inference. *Epidemiology*. 2000; 11(3):313–320.
18. Brookhart MA, Sturmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. *Medical Care*. 2010; 48(6 Suppl 1)
19. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*. 2003; 14(3):300–6. [PubMed: 12859030]
20. Cole SR, Platt RW, Schisterman EF, et al. Illustrating bias due to conditioning on a collider. *Int J Epidemiol*. 2010; 39(2):417–20. [PubMed: 19926667]
21. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982; 143(1):29–36. [PubMed: 7063747]
22. Harrell, FEJ. *Regression Modeling Strategies*. New York City: Springer-Verlag; 2001.
23. Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiol Drug Saf*. 2004; 13(12):841–53. [PubMed: 15386709]
24. Westreich D, Cole SR. Invited commentary: positivity in practice. *Am J Epidemiol*. 2010; 171(6): 674–7. discussion 678–81. [PubMed: 20139125]
25. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008; 168(6):656–64. [PubMed: 18682488]
26. 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. *Stat Med*. 2007; 26(4):734–53. [PubMed: 16708349]
27. Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Weaknesses of goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder. *Pharmacoepidemiol Drug Saf*. 2005; 14(4):227–38. [PubMed: 15386700]
28. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med*. 2009; 28(25):3083–107. [PubMed: 19757444]
29. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med*. 2007; 26(1):20–36. [PubMed: 17072897]
30. van der Laan, MJ.; Dudoit, S. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite sample oracle inequalities and examples. U C Berkeley Division of Bio- statistics Working Paper Series, paper 130. 2003. <http://www.bepress.com/ucbbiostat/paper130>
31. Brookhart MA, van der Laan MJ. A semiparametric model selection criterion with applications to the marginal structural model. *Comput Stat Data Anal*. 2006; 50:475–98.
32. Hernán MA, Hernández-Díaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am J Epidemiol*. 2002; 155(2):176–84. [PubMed: 11790682]