American Journal of
EPIDEMIOLOGY

Volume 150
Number 4
August 15, 1999

Sponsored by the Society for Epidemiologic Research

COMMENTARY

Invited Commentary: Propensity Scores

Marshall M. Joffe[1] and Paul R. Rosenbaum[2]

The propensity score is the conditional probability of exposure to a treatment given observed covariates. In a cohort study, matching or stratifying treated and control subjects on a single variable, the propensity score, tends to balance all of the observed covariates; however, unlike random assignment of treatments, the propensity score may not also balance unobserved covariates. The authors review the uses and limitations of propensity scores and provide a brief outline of associated statistical theory. They also present a new result of using propensity scores in case-cohort studies. *Am J Epidemiol* 1999;150:327–33.

causality; confounding; observational study

DEFINITION AND MOTIVATION

An observational study is an attempt to estimate the effects of a treatment or an exposure by comparing outcomes for subjects who were not assigned at random to treatment or control, which would have happened in a randomized, controlled trial (1, 2). A variable, such as age or gender, that is measured prior to the start of treatment, and hence is unaffected by the treatment, will be called a covariate. Random assignment of subjects to treatment or control tends to balance covariates, so that treated and control groups are comparable in the sense that they have similar distributions of covariates, for instance, similar age distributions or similar proportions of women. Absent random assignment, the propensity score is a device for constructing matched pairs or matched sets or strata that balance numerous observed covariates (3). Adjustment for an estimated propensity score tends to balance observed covariates that were used to construct the score, but, unlike random assignment of treatments, the propensity score typically does not balance covariates that were not observed. Imbalances in unobserved covariates must be addressed by using additional methods (2, 4–6).

Each subject has observed covariates, $\mathbf{X}$, and an indicator of treatment, $Z = 1$ if treated and $Z = 0$ if control. The $\mathbf{X}$ for one person might record dozens of pretreatment measurements describing that person. The propensity score, $e(\mathbf{X})$, is the chance that a person with covariates $\mathbf{X}$ will be exposed to treatment, that is, $e(\mathbf{X}) = \text{prob}(Z = 1|\mathbf{X})$. The propensity score has a number of theoretical properties that have been verified in both simulated and practical situations. Before these properties are presented, it is useful to consider some informal but suggestive motivation. In the simplest randomized trial, subjects are assigned to treatment or control by the flip of a fair coin, so $e(\mathbf{X}) = \text{prob}(Z = 1|\mathbf{X}) = \frac{1}{2}$ for every $\mathbf{X}$; therefore, subjects with different patterns of covariates all have the same chance of receiving the treatment, and each possible value of $\mathbf{X}$ is as likely to turn up in the treated group as in the control group. Quite often, the published report of a randomized trial includes a table documenting that the randomization was effective, that the treated and control groups were comparable in terms of the distributions of important covariates.

In contrast, in an observational study, some subjects are more likely than others to receive the treatment, so

Received for publication February 3, 1999, and accepted for publication February 18, 1999.

[1] Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA.
[2] Department of Statistics, The Wharton School of the University of Pennsylvania, Philadelphia, PA.
Reprint requests to Dr. Marshall M. Joffe, Department of Biostatistics and Epidemiology, Room 602 Blockley Hall, 423 Guardian Drive, University of Pennsylvania, Philadelphia, PA 19104-6021.

$e(\mathbf{X}) \neq \frac{1}{2}$ for some persons, and the pattern of covariates **X** often helps to predict which treatment a subject will receive. However, suppose that we compare two subjects who have the same chance of receiving the treatment given their observed covariates **X**, say two subjects with $e(\mathbf{X}) = \frac{1}{4}$. These two subjects may be very different in terms of **X**, but their differences do not help predict which subject is more likely to receive the treatment. Given only the information in the observed covariates **X**, both subjects have the same chance, namely ¼, of receiving the treatment. So, the first subject with his **X** and the second subject with her perhaps very different **X** appear to have the same chance of ending up in the treated group; his **X** is as likely to be found in the treated group as hers is.

This suggests, and both theory and experience confirm, that if we pair or group subjects with the same propensity score $e(\mathbf{X})$, then treated and control subjects in these groups will have similar patterns or distributions of **X**. For instance, if **X** records (among other things) age, race, and gender, then the stratum consisting of all subjects with $e(\mathbf{X}) = \frac{1}{4}$ will contain people of varied ages and races and both genders. However, the mean age of subjects in the treated group will be similar to the mean age of those in the control group, the proportions of women will be similar, and the proportions of Black women older than age 50 years will be similar. The balance on the observed covariates **X** that is obtained by matching or stratifying on an estimated propensity score is of course imperfect, but it is typically somewhat better than the balance on **X** obtained by random assignment of treatments. Again, random assignment does not consider the observed **X** in balancing **X**, so this method also balances unobserved covariates. On the other hand, matching on the estimated propensity score does use **X** in balancing **X** and may do nothing to balance unobserved covariates.

The propensity score complements model-based procedures and is not a substitute for them. Matching or stratifying on propensity scores is often used in conjunction with further model-based adjustments using regression (7) or log-linear models (8). In addition, propensity scores are often used directly as the foundation for inference (9–12).

In this paper, we first review some of the literature on propensity scores in cohort studies and then present new results on their use in case-cohort studies. Related to propensity scores are discriminant matching for multivariate Normal covariates (13) and Miettinen's confounder scores (14), but the theoretical development of the properties of confounder scores is less complete (refer to Rosenbaum and Rubin (3) for a discussion of these associations).

EMPIRICAL RESULTS

In an observational study comparing coronary bypass surgery ($Z = 1$) with medical treatment ($Z = 0$), five strata formed from an estimated propensity score balanced 74 covariates that characterized patient health prior to treatment (8). The covariates described left ventricle function, ejection fraction, degree and location of coronary stenosis, age, performance status, and other characteristics. Before stratification, the surgical and medical groups differed significantly in terms of all 74 covariates. The propensity score $e(\mathbf{X}) = \text{prob}(Z = 1|\mathbf{X})$ was estimated by using logit regression of the decision to perform or withhold surgery Z on the covariates **X**, including needed interactions and quadratics. Patients were then divided into five strata by using the estimated probabilities of surgery; each stratum contained 20 percent of the patients.

Each stratum was heterogeneous in the covariates **X**, but, within the same stratum, the medical and surgical patients had similar distributions of the covariates. For instance, the stratum containing the patients who had the lowest probabilities of surgery included both many of the sickest and many of the healthiest patients. Poor left ventricle function made a patient a poor candidate for surgery, but so did limited coronary stenosis. Still, within this heterogeneous stratum of patients unlikely to receive surgery, the medical and surgical patients had similar distributions of the observed covariates. In fact, the balance on these 74 covariates was substantially greater than would have been expected by random assignment of treatments. When the stratified surgical and medical groups were compared on the basis of each of the 74 covariates, two tests of equality of the distributions were performed, thus yielding a total of 148 tests. After stratification, only one of the 148 tests of covariate balance was significant at the 0.05 level; in a randomized trial, 5 percent or $0.05 \times 148 = 7.4$ significant differences would be expected to occur by chance alone. Patient outcomes were then compared in two ways, adjusting just for the five strata and using a log-linear model relating patient outcomes to treatment, two key covariates from the 74 in **X**, and the five strata. Thus, it was possible to compare medicine and surgery for a few key, specific subgroups of patients while balancing many other less interesting covariates.

In a study of the effects of prenatal exposures to barbiturates, Rosenbaum and Rubin used the propensity score to match 221 exposed children to 221 unexposed controls drawn from a reservoir of 7,027 potential controls (15). The study involved extensive follow-up of these 442 children, so it was not practical to follow all 7,027 potential controls. The propensity score was estimated by using a logit model with 20 covariates together with selected interactions and quadratics. The

matching removed not only most of the initial bias in the 20 covariates but also most of the bias in the squares of the 20 covariates and the 190 interactions formed by multiplying two covariates. By removing bias not only in the mean value of the covariates but also in their squares and interactions, the matching also removed bias in quadratic response surfaces that might have been formed from these covariates.

Matching on the propensity score is much better than dividing each covariate into two categories and then matching exactly on the categories. If the latter form of category matching had been used, matches would have been found for only 126 of the 221 exposed children, yet the balance it would have produced on the original variables would have been poorer than the balance obtained by propensity matching with all 221 exposed children (16). Moreover, the 126 matched exposed children were very different from the remaining 95 unmatched exposed children, which would have introduced a substantial bias due to incomplete matching (16). The propensity score is also used when multiple controls are matched (17).

A simulation study found that matching by using propensity scores was vastly superior to using several other matching methods when there were 20 covariates (18). In particular, with 20 covariates, propensity matching often removed more than twice the bias removed by Mahalanobis metric matching. Propensity matching may be combined to advantage with other matching techniques, such as Mahalanobis metric matching (15, 19), optimal matching (20), and full matching (21). The Institute for Scientific Information (ISI) *Scientific Citation Index* lists some 170 citations of the six initial propensity score papers (3, 8–10, 15, 16), including numerous applications in medicine and the social sciences.

THEORETICAL PROPERTIES

The empirical results just discussed are consistent with the theory of propensity scores. Briefly, there are three theoretical issues: 1) propensity scores balance observed covariates; 2) if it suffices to adjust for covariates $\mathbf{X}$, then it suffices to adjust for their propensity score $e(\mathbf{X})$; and 3) estimated propensity scores are better than true propensity scores at removing bias, because they also remove some chance imbalances in $\mathbf{X}$.

Stated more precisely, a balancing score is a summary or function $\mathbf{b}(\mathbf{X})$ of the observed covariates such that treatment Z and the observed covariates $\mathbf{X}$ are conditionally independent given $\mathbf{b}(\mathbf{X})$. In a stratum that is homogeneous in a balancing score, treated $Z = 1$ and control $Z = 0$ subjects will have the same distribution of the observed covariates $\mathbf{X}$. The propensity score $e(\mathbf{X})$ is the simplest or coarsest balancing score, the covariates $\mathbf{X}$ themselves are the most complex or the finest balancing score, and every balancing score contains the information in the propensity score $e(\mathbf{X})$ and some additional information from $\mathbf{X}$. Therefore, to balance covariates, strata must be homogenous in the propensity score but may control for other aspects of $\mathbf{X}$ as well (refer to Rosenbaum and Rubin (3), theorem 2).

Adjustments, such as matching or stratification, for observed covariates $\mathbf{X}$ are commonly performed in observational studies, but it is a familiar fact that these adjustments may not suffice to yield appropriate estimates of treatment effects. Subjects who appear similar in terms of observed covariates $\mathbf{X}$ may differ regarding important covariates not observed. It is useful to be able to formally describe the circumstances under which adjustments for $\mathbf{X}$ will suffice; that formal condition is known as *strongly ignorable treatment assignment* (3). Imagine, for example, that subjects are assigned randomly to treatment or control with unequal, nonzero probabilities that are a function of $\mathbf{X}$ alone but that the function itself is unknown. In this instance, treatment assignment is strongly ignorable, and adjustments for $\mathbf{X}$ do suffice to estimate treatment effects. Strong ignorability, defined formally in the Appendix, implies that no systematic, unobserved, pretreatment differences exist between treated and control subjects that are related to the response under study. Strong ignorability is related closely to the absence of residual confounding by unmeasured factors under comparability-based definitions of confounding (for example, refer to Greenland and Robins (22)).

A key fact (Rosenbaum and Rubin (3), theorem 3) about balancing scores is that if it suffices to adjust for $\mathbf{X}$, then it suffices to adjust for any balancing score $\mathbf{b}(\mathbf{X})$. Formally stated, ignorability given $\mathbf{X}$ implies ignorability given $\mathbf{b}(\mathbf{X})$. For instance, in the coronary bypass surgery example mentioned in the previous section, if it would have been sufficient to adjust for the 74 patient covariates in $\mathbf{X}$, then it also would have been sufficient to adjust for the single propensity score $e(\mathbf{X})$. While it is difficult to stratify on 74 covariates at once, it is straightforward to use strata to remove bias in a single covariate (23). However, if this covariate is the propensity score, then bias due to all 74 covariates is removed.

In practice, the propensity score is unknown and must be estimated, for instance, by using a logit model, as described in the previous section. When an estimate is used in place of a true value, it is usually expected that the estimate will not perform as well; surprisingly, however, estimated propensity scores perform better than true propensity scores, as confirmed by theory (9–11), simulation (18), and practical experience (8). An estimated propensity score cannot distinguish sys-

tematic bias from an imbalance in covariates that arises from bad luck, and adjustment for an estimated propensity score tends to remove both types of imbalance. On the other hand, adjustment for the true propensity score removes systematic bias only.

In studies in which a rare disease is the outcome and a comparatively common exposure Z is under investigation, modeling the relation between exposure Z and covariates $\mathbf{X}$, that is, modeling the propensity score $e(\mathbf{X}) = \text{prob}(Z = 1|\mathbf{X})$, may be more feasible than modeling the relation between the disease outcome and $\mathbf{X}$. For instance, to study survival given the 74 covariates $\mathbf{X}$ in the example of surgery versus medical treatment presented previously, the maximum likelihood estimate for the very simplest logit model with a constant and just these covariates and no interactions would require a minimum of 75 deaths. This condition is necessary but not sufficient for the mere existence of the maximum likelihood estimate and is generally insufficient to yield good performance of this estimate. In contrast, a model for the propensity score would not require many deaths for a stable estimate of $e(\mathbf{X})$, just many patients assigned to both bypass surgery $Z = 1$ and drug treatment $Z = 0$. Matching or stratifying on the estimated propensity score could be used to adjust for all 74 covariates, and a model for responses could be confined to a small subset of the covariates in $\mathbf{X}$, as illustrated by Rosenbaum and Rubin (8).

EXTENSIONS OF PROPENSITY SCORE METHODS

This section briefly outlines several extensions of propensity scores and is slightly more technical than previous sections. Topics include the following: 1) methods of inference using propensity scores directly, without matching or stratification; 2) propensity scores in case-cohort studies; and 3) propensity scores with doses of treatment.

Inference using models for propensity scores

Although originally intended as an aid to multivariate matching and stratification, the propensity score can be used directly in inference, without matching or stratification. The propensity score provides estimates and tests of treatment effects (9–12), which sometimes simplify modeling of the effects of time-varying treatments (24, 25). Roughly speaking, when the treatment has an effect, treatment assignment is associated with subsequent outcome; therefore, a model for the propensity score that predicts assignment from covariates and outcomes exhibits an association with the outcomes, and this association will not be spurious if treatment assignment is strongly ignorable. In the simplest case, the null hypothesis of no treatment effect might be tested by using a logit regression of the assigned treatment on the outcome and covariates (9). This strategy quickly generalizes in several directions, yielding, for instance, exact and approximate tests of the null hypothesis of no treatment effect and, by inversion, confidence intervals for the magnitude of the effect (9, 11). Also, when the treatment a person receives changes over time, it is often simpler and more robust to model the treatment assignments without modeling responses (25). Estimates of population quantities are sometimes obtained by weighting observations inversely as their estimated propensity scores (10), and this approach has also been generalized by Robins (26) to time-varying treatments.

Propensity scores in case-cohort studies

The simplest, somewhat idealized case-cohort or case-base study uses all of the incident cases of a disease in a cohort and a simple random sample from the cohort as a whole (27–31). Although propensity scores are not typically applied to case-cohort studies, they are directly applicable in principle.

This simple case-cohort design is used most commonly when some data concerning the entire cohort are available from computerized records but costly additional data, such as genetic information (32), must be collected from subjects before an analysis is possible. Sampling of cohort members drastically reduces the cost of obtaining information about the cohort as a whole. Random sampling avoids the need to weigh the subtle trade-offs between confounding and selection bias that arise without random samples; refer to Rosenbaum (2), section 6.3.

The procedure is straightforward. First, the entire cohort is used to estimate the propensity score or any other balancing score $\mathbf{b}(\mathbf{X})$ from the binary exposures Z and the observed covariates $\mathbf{X}$. After the propensity score has been estimated, all cases and a random sample of cohort members are then selected, the required additional information is obtained, and cases and cohort members are compared within strata defined by the balancing score $\mathbf{b}(\mathbf{X})$. Within a single stratum, the risk ratio is estimated in the usual way, as described by Kupper et al. (28), whereas if the risk ratio is considered constant across strata, the stratum-specific results may be combined by using the methods described by Greenland (30) and Sato (31). If it suffices to adjust for the covariates $\mathbf{X}$ observed in the entire cohort, that is, if treatment assignment is strongly ignorable in the cohort, then this estimate of the stratum-specific risk ratio is free of confounding due to the observed covariates $\mathbf{X}$. Refer to the Appendix for a precise statement and a proof.

Propensity scores with ordered doses

A single variable, the propensity score, can always act as a balancing score when a treatment is compared with a control, but the situation is more complex when there are doses of treatment, say $Z = 2$ = high, $Z = 1$ = low, and $Z = 0$ = none = control. In this case, generally it is not sufficient to model and stratify on the expected dose $E(Z|\mathbf{X})$ given covariates $\mathbf{X}$. Unlike the case of two doses, the expected dose given covariates $\mathbf{X}$ need not fully describe the distribution of doses Z.

However, under special circumstances, a single-variable balancing score is available with more than two doses. This situation happens when there is a single variable, say $b(\mathbf{X})$, that determines not just the expected dose given $\mathbf{X}$ but the entire distribution of doses given $\mathbf{X}$. More precisely, if the entire distribution of doses Z depends on covariates $\mathbf{X}$ only through $b(\mathbf{X})$, so that $\text{prob}(Z|\mathbf{X}) = \text{prob}(Z|b(\mathbf{X}))$, then $b(\mathbf{X})$ is a balancing score and persons with the same balancing score in different dose groups have the same distribution of the covariates $\mathbf{X}$, that is, $\text{prob}(\mathbf{X}|b(\mathbf{X}), Z = z) = \text{prob}(\mathbf{X}|b(\mathbf{X}), Z = z')$ for each z, z'. For instance, this situation would be true if the distribution of doses given $\mathbf{X}$ was described accurately by McCullagh's ordinal logit model (33, 34); then, stratifying on a single variable, $b(\mathbf{X}) = \mathbf{X}^T\beta$, would balance $\mathbf{X}$ across several dose groups.

More generally, under certain models (35), the distribution of doses Z given many covariates $\mathbf{X}$ depends on the covariates only through a small number of linear functions of $\mathbf{X}$, say $\mathbf{X}\Lambda$ for some matrix Λ, in which case $\mathbf{X}\Lambda$ is a balancing score and strata that control for all of the several variables $\mathbf{X}\Lambda$ will tend to balance the many variables in $\mathbf{X}$.

SUMMARY

Propensity scores are used to create matched pairs or matched sets or strata that balance many observed covariates. The resulting matched sets are heterogeneous in the covariates, but the covariates tend to have similar distributions in treated and control groups; therefore, the groups as a whole appear comparable. Unlike random assignment of treatments, adjustment for the propensity score does little to balance unobserved covariates. If adjustments for the many observed covariates are sufficient to remove the bias in estimated treatment effects, then adjustments for the single variable, the propensity score, also are sufficient to remove bias. In cohort studies, estimated propensity scores usually perform better than true propensity scores because they remove some chance imbalances in covariates that the true propensity score leaves behind. Adjustments by matching or stratification on the propensity score are often combined with model-based analyses within matched pairs or strata.

ACKNOWLEDGMENTS

Dr. Joffe's work was supported by a grant from the National Heart, Lung and Blood Institute (R29 HL59184). Dr. Rosenbaum's work was supported by a grant from the National Science Foundation (SBR-9808261).

REFERENCES

1. Cochran WG. The planning of observational studies of human populations. J Royal Stat Soc 1965;182:234–55.
2. Rosenbaum PR. Observational studies. New York, NY: Springer-Verlag, 1995.
3. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41–55.
4. Rosenbaum PR. Sensitivity analysis for matched case-control studies. Biometrics 1991;47:87–100.
5. Rosenbaum PR. Quantiles in nonrandom samples and observational studies. J Am Stat Assoc 1995;90:1424–31.
6. Rosenbaum PR. From association to causation in observational studies: the role of tests of strongly ignorable treatment assignment. J Am Stat Assoc 1984;79:41–8.
7. Rosenbaum PR. Dropping out of high school in the United States: an observational study. J Educ Stat 1986;11:207–24.
8. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc 1984;79:516–24.
9. Rosenbaum PR. Conditional permutation tests and the propensity score in observational studies. J Am Stat Assoc 1984;79:565–74.
10. Rosenbaum PR. Model-based direct adjustment. J Am Stat Assoc 1987;82:387–94.
11. Robins JM, Mark SD, Newey WK. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. Biometrics 1992;48:479–95.
12. Robins JM, Ritov Y. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semiparametric models. Stat Med 1997;16:285–319.
13. Cochran WG, Rubin D. Controlling bias in observational studies: a review. Sankya 1973;35:417–46.
14. Miettinen OS. Stratification by a multivariate confounder score. Am J Epidemiol 1976;104:609–20.
15. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat 1985;39:33–8.
16. Rosenbaum PR, Rubin DB. The bias due to incomplete matching. Biometrics 1985;41:103–16.
17. Smith HL. Matching with multiple controls to estimate treatment effects in observational studies. Soc Methodol 1997;27:325–53.
18. Gu XS, Rosenbaum PR. Comparison of multivariate matching methods: structures, distances, and algorithms. J Comput Graph Stat 1993;2:405–20.
19. Rubin DB. Bias reduction using Mahalanobis metric matching. Biometrics 1980;36:293–8.
20. Rosenbaum PR. Optimal matching for observational studies. J Am Stat Assoc 1989;84:1024–32.
21. Rosenbaum PR. A characterization of optimal designs for observational studies. J Royal Stat Soc 1991;53:597–610.
22. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol 1986;15:412–18.
23. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics 1968;24:295–313.
24. Robins JM, Blevins D, Ritter G, et al. G-estimation of the effect of prophylaxis therapy for pneumocystic carinii pneumonia on the survival of AIDS patients. Epidemiology 1992;3:319–66.

25. Joffe MM, Hoover DR, Jacobson LP, et al. Estimating the effect of zidovudine on Kaposi's sarcoma using a rank preserving structural failure-time model. Stat Med 1998;17: 1073–102.
26. Robins JM. Correction for non-compliance in equivalence trials. Stat Med 1998;17:269–302.
27. MacMahon B. Prenatal exposure and childhood cancer. J Natl Cancer Inst 1962;28:1173–91.
28. Kupper LL, McMichael AJ, Spirtas R. A hybrid epidemiologic study design useful in estimating relative risk. J Am Stat Assoc 1975;70:524–8.
29. Miettinen O. Design options in epidemiologic research: an update. Scand J Work Environ Health 1982;8(suppl 1):7S–14S.
30. Greenland S. Adjustment of risk ratios in case-base studies (hybrid epidemiologic designs). Stat Med 1986;5:579–84.
31. Sato T. Estimation of a common risk ratio in stratified case-cohort studies. Stat Med 1992;11:1599–605.
32. Umbach DM, Weinberg CR. Designing and analysing case-control studies to exploit independence of genotype and exposure. Stat Med 1997;16:1731–43.
33. McCullagh P. Regression models for ordinal data. J Royal Stat Soc B 1980;42:109–42.
34. Armstrong BG, Sloan M. Ordinal regression models for epidemiologic data. Am J Epidemiol 1989;129:191–204.
35. Joffe MM, Greenland S. Standardized estimates from categorical regression models. Stat Med 1995;14:2131–41.
36. Neyman J. On the application of probability theory to agricultural experiments. Essay on principles. (Translated by D. M. Dabrowska and edited by T. P. Speed). Stat Sci 1990;5:465–72.
37. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 1974;66: 688–701.
38. Hamilton MA. Choosing the parameter for a 2 × 2 table or 2 × 2 × 2 table analysis. Am J Epidemiol 1979;109:362–75.
39. Holland PW, Rubin DB. Causal inference in retrospective studies. Eval Review 1988;12:203–31.
40. Rosenbaum PR. Propensity score. In: Armitage P, Colton T, eds. Encyclopedia of biostatistics. New York, NY: John Wiley, 1999:3551–5.

APPENDIX

Proof of Properties of Propensity Scores in Case-Cohort Studies

As the property of propensity scores in case-cohort studies is new, this appendix provides a proof. Recall the following terms and facts from Rosenbaum and Rubin (3). As explained by Kupper et al. (28), during the time period under study, each subject either becomes a new case of the disease, signified by $R = 1$, or does not, signified by $R = 0$. Each subject in the cohort has two potential responses, only one of which is observed, namely, a response $r_T = 1$ if the disease occurs when the subject is exposed to treatment or $r_T = 0$ if the disease does not occur when the subject is exposed to treatment, and a response $r_C = 1$ if the disease occurs when the subject is exposed to control or $r_C = 0$ if the disease does not occur when the subject is exposed to control (36–39). Then, $R = r_T$ if the subject was actually exposed to the treatment, $Z = 1$, whereas $R = r_C$ if $Z = 0$. A function $\mathbf{b}(\mathbf{X})$ is a *balancing score* (3) if treatment Z is independent of observed covariates $\mathbf{X}$ given $\mathbf{b}(\mathbf{X})$, that is, if $\text{prob}(Z = 1|\mathbf{X}) = \text{prob}(Z = 1|\mathbf{b}(\mathbf{X}))$. Treatment assignment is *strongly ignorable* given the observed covariates $\mathbf{X}$ if treatment Z is not determined by $\mathbf{X}$ and is independent of the potential responses (r_T, r_C) given the covariates $\mathbf{X}$, that is, if both $0 < \text{prob}(Z = 1|\mathbf{X}) < 1$ and $\text{prob}(Z = 1|\mathbf{X}, r_T, r_C) = \text{prob}(Z = 1|\mathbf{X})$ for all $\mathbf{X}$ and, in this case, appropriate adjustment, such as matching or stratification, for the observed covariates $\mathbf{X}$ yields consistent and often unbiased estimates of treatment effects; refer to Rosenbaum and Rubin (3), theorem 3. If treatment assignment is strongly ignorable given $\mathbf{X}$, then it is also strongly ignorable given any balancing score $\mathbf{b}(\mathbf{X})$. Therefore, if $\text{prob}(Z = 1|r_T, r_C, \mathbf{X}) = \text{prob}(Z = 1|\mathbf{X})$, then $\text{prob}(Z = 1|r_T, r_C, \mathbf{b}(\mathbf{X})) = \text{prob}(Z = 1|\mathbf{b}(\mathbf{X}))$ for every balancing score $\mathbf{b}(\mathbf{X})$; refer to Rosenbaum and Rubin (3), theorem 3.

Let $\mathbf{b}(\mathbf{X})$ be any balancing score, and consider the subpopulation of the cohort with a particular value of the balancing score. If all subjects in this subpopulation had been exposed to the treatment, then the population proportion of disease in this subpopulation would have been $\text{prob}(r_T = 1|\mathbf{b}(\mathbf{X}))$. On the other hand, if all subjects in this subpopulation had escaped exposure to the treatment, then the proportion of disease would have been $\text{prob}(r_C = 0|\mathbf{b}(\mathbf{X}))$ and the causal risk ratio $\rho_{\mathbf{b}(\mathbf{X})}$ in this subpopulation is $\rho_{\mathbf{b}(\mathbf{X})} = \text{prob}(r_T = 1|\mathbf{b}(\mathbf{X}))/\text{prob}(r_C = 0|\mathbf{b}(\mathbf{X}))$. When can $\rho_{\mathbf{b}(\mathbf{X})}$ be estimated from a case-cohort study?

In a case-cohort study, the proportion of cases exposed in the subpopulation defined by $\mathbf{b}(\mathbf{X})$ estimates the frequency of exposure to treatment, $Z = 1$, among cases, $R = 1$, in the subpopulation defined by $\mathbf{b}(\mathbf{X})$. That is, it consistently estimates $\text{prob}(Z = 1|R = 1, \mathbf{b}(\mathbf{X}))$. The study also provides an estimate of the frequency of exposure to treatment in the entire subpopulation, namely, $\text{prob}(Z = 1|\mathbf{b}(\mathbf{X}))$, which equals the propensity score $e(\mathbf{X})$ according to theorem 2 of Rosenbaum and Rubin (3). Combining these estimates yields a consistent estimate of

$$\lambda_{\mathbf{b}(\mathbf{X})} = \left(\frac{\text{prob}(Z = 1|R = 1, \mathbf{b}(\mathbf{X}))}{1 - \text{prob}(Z = 1|R = 1, \mathbf{b}(\mathbf{X}))}\right) \left(\frac{1}{\text{prob}(Z = 1|\mathbf{b}(\mathbf{X}))} - 1\right) \quad (1)$$

$$= \left(\frac{\text{prob}(R = 1|Z = 1, \mathbf{b}(\mathbf{X}))}{\text{prob}(R = 1|Z = 0, \mathbf{b}(\mathbf{X}))}\right)$$

$$= \left(\frac{\text{prob}(r_T = 1|Z = 1, \mathbf{b}(\mathbf{X}))}{\text{prob}(r_C = 1|Z = 0, \mathbf{b}(\mathbf{X}))}\right) \quad (2)$$

where the step from expression 1 to the left of expression 2 parallels expression 2.2 of Kupper et al. (28), and the last equality follows from the fact that $R = r_T$ when $Z = 1$ and $R = r_C$ when $Z = 0$. If adjustments for $\mathbf{X}$ suffice to remove bias, that is, if treatment assignment is strongly ignorable given $\mathbf{X}$, then $\text{prob}(r_T = 1|Z = z, \mathbf{b}(\mathbf{X})) = \text{prob}(r_T = 1|\mathbf{b}(\mathbf{X}))$ and $\text{prob}(r_C = 1|Z = z, \mathbf{b}(\mathbf{X})) = \text{prob}(r_C = 1|\mathbf{b}(\mathbf{X}))$, so that $\rho_{\mathbf{b}(\mathbf{X})} = \lambda_{\mathbf{b}(\mathbf{X})}$, and the causal risk ratio $\rho_{\mathbf{b}(\mathbf{X})}$ in the subpopulation defined by $\mathbf{b}(\mathbf{X})$ may be estimated by stratifying on $\mathbf{b}(\mathbf{X})$. In short, adjustments for a balancing score permit estimation of the causal risk ratio $\rho_{\mathbf{b}(\mathbf{X})}$ from a case-cohort study when treatment assignment is strongly ignorable, as asserted.

Strong ignorability implies that the bias due to nonrandom selection into treated and control groups may be removed by analytical adjustments, so that $\rho_{\mathbf{b}(\mathbf{X})} = \lambda_{\mathbf{b}(\mathbf{X})}$ for every balancing score $\mathbf{b}(\mathbf{X})$. However, strong ignorability does not imply that risk ratios are homogeneous (40); that is, in general, different balancing scores $\mathbf{b}(\mathbf{X})$ yield different true stratum-specific risk ratios $\rho_{\mathbf{b}(\mathbf{X})}$. For instance, this is true even if the entire population were included in a randomized experiment, where stratification on age or gender is possible but not necessary for an unbiased comparison, but the risk ratio for a person aged 50 years may differ from the risk ratio for women.