Estimating Causal Effects from Large Data Sets Using Propensity Scores

Donald B. Rubin, PhD

The aim of many analyses of large databases is to draw causal inferences about the effects of actions, treatments, or interventions. Examples include the effects of various options available to a physician for treating a particular patient, the relative efficacies of various health care providers, and the consequences of implementing a new national health care policy. A complication of using large databases to achieve such aims is that their data are almost always observational rather than experimental. That is, the data in most large data sets are not based on the results of carefully conducted randomized clinical trials, but rather represent data collected through the observation of systems as they operate in normal practice without any interventions implemented by randomized assignment rules. Such data are relatively inexpensive to obtain, however, and often do represent the spectrum of medical practice better than the settings of randomized experiments. Consequently, it is sensible to try to estimate the effects of treatments from such large data sets, even if only to help design a new randomized experiment or shed light on the generalizability of results from existing randomized experiments. However, standard methods of analysis using available statistical software (such as linear or logistic regression) can be deceptive for these objectives because they provide no warnings about their propriety. Propensity score methods are more reliable tools for addressing such objectives because the assumptions needed to make their answers appropriate are more assessable and transparent to the investigator.

Ann Intern Med. 1997;127:757-763.

From Harvard University, Cambridge, Massachusetts. For the current author address, see end of text.

Many observational studies based on large databases attempt to estimate the causal effects of some new treatment or exposure relative to a control condition, such as the effect of smoking on mortality. In most such studies, it is necessary to control for naturally occurring systematic differences in background characteristics between the treatment group and the control group, such as age or sex distributions, that would not occur in the context of a randomized experiment. Typically, many background characteristics need to be controlled.

Propensity score technology, introduced by Rosenbaum and Rubin (1), addresses this situation by reducing the entire collection of background characteristics to a single composite characteristic that appropriately summarizes the collection. This reduction from many characteristics to one composite characteristic allows the straightforward assessment of whether the treatment and control groups overlap enough with respect to background characteristics to allow a sensible estimation of treatment versus control effects from the data set. Moreover, when such overlap is present, the propensity score approach allows a straightforward estimation of treatment versus control effects that reflects adjustment for differences in all observed background characteristics.

Subclassification on One Confounding Variable

Before describing the use of propensity scores in the statistical analysis of observational studies with many confounding background characteristics, I begin with an example showing how subclassification adjusts for a single confounding covariate, such as age, in a study of smoking and mortality. I then show how propensity score methods generalize subclassification in the presence of many confounding covariates, such as age, region of the country, and sex.

The potential for a large database to suggest causal effects of treatments is indicated in **Table 1**, adapted from Cochran's work (2), which concerns mortality rates per 1000 person-years for nonsmokers, cigarette smokers, and cigar and pipe smokers drawn from three large databases in the United States, the United Kingdom, and Canada. The treatment factor here involves three levels of smoking. The unadjusted mortality rates in **Table 1** make it seem that cigarette smoking is good for health, especially relative to cigar and pipe smoking; clearly, this result is contrary to current wisdom. A problem with this naive conclusion is exposed in Table 1, where the average ages of the subpopulations are given. Age correlates with both mortality rates and smoking behavior. In this example, age is a confounding covariate, and conclusions about the effects of smoking should be adjusted for its effects.

A straightforward way of adjusting for age is to 1) divide the population into age categories of approximately equal size (such as younger and older if two categories are appropriate; younger, middle-aged, and older if three are appropriate; and so on), 2) compare mortality rates within an age category (for example, compare mortality rates for the three treatment groups within the younger population and similarly for the older population), and 3) average the age-group–specific comparisons to obtain overall

© 1997 American College of Physicians

Table 1. Comparison of Mortality Rates for Three Smoking Groups in Three Databases*

Variable	Canadian Study			United Kingdom Study			United States Study		
	Nonsmokers	Cigarette Smokers	Cigar and Pipe Smokers	Nonsmokers	Cigarette Smokers	Cigar and Pipe Smokers	Nonsmokers	Cigarette Smokers	Cigar and Pipe Smokers
Mortality rates per 1000 person-years, %	20.2	20.5	35.5	11.3	14.1	20.7	13.5	13.5	17.4
Average age, *y*	54.9	50.5	65.9	49.1	49.8	55.7	57.0	53.2	59.7
Adjusted mortality rates using subclasses, %									
2 subclasses	20.2	26.4	24.0	11.3	12.7	13.6	13.5	16.4	14.9
3 subclasses	20.2	28.3	21.2	11.3	12.8	12.0	13.5	17.7	14.2
9–11 subclasses	20.2	29.5	19.8	11.3	14.8	11.0	13.5	21.2	13.7

* Adapted from Tables 1–3 in Cochran (2).

estimates of the age-adjusted mortality rates per 1000 person-years for each of the three groups. **Table 1** shows the results for different numbers of age categories where the subclass-age boundaries were defined to have equal numbers of nonsmokers in each subclass. These results align better than the unadjusted mortality rates with our current understanding of the effects of smoking, especially when 9 to 11 subclasses are used. Incidentally, having approximately equal numbers of nonsmokers within each subclass is not necessary, but if the nonsmokers are considered the baseline group, it is a convenient and efficient choice because then the overall estimated effect is the simple unweighted average of the subclass-specific results. That is, the mortality rates in all three groups are being standardized (3) to the age distribution of nonsmokers as defined by their subclass counts.

Cochran (2) calls this method *subclassification* and offers theoretical results showing that as long as the treatment and exposure groups overlap in their age distributions (that is, as long as a reasonable number of persons from each treatment group are in each subclass), comparisons using five or six subclasses will typically remove 90% or more of the bias present in the raw comparisons shown in **Table 1**. More than five subclasses were used for the adjusted mortality rates because the large size of the data sets made it possible to do so.

A particular statistical model, such as a linear regression (or a logistic regression model; or in other settings, a hazard model) could have been used to adjust for age, but subclassification has two distinct advantages over such models, at least for offering initial trustworthy comparisons that are easy to communicate. First, if the treatment or exposure groups do not adequately overlap on the confounding covariate age, the investigator will see it immediately and be warned. Thus, if members of one group have ages outside the range of another group's ages, it will be obvious because one or more age-specific subclasses will consist almost solely of members exposed to one treatment. In contrast, nothing in the standard output of any regression modeling software will display this critical fact; the reason is that models predict an outcome (such as death) from regressors (such as age and treatment indicators), and standard regression diagnostics do not include careful analysis of the joint distribution of the regressors (such as a comparison of the distributions of age across treatment groups). When the overlap on age is too limited, the database, no matter how large, cannot support any causal conclusions about the differential effects of the treatments. For example, comparing 5-year survival rates among 70-year-old smokers and 40-year-old nonsmokers gives essentially no information about the effect of smoking or nonsmoking for either 70-year-old or 40-year-old persons.

The second reason for preferring subclassification to models concerns situations such as that found in **Table 1**, in which the groups overlap enough on the confounding covariate to make a comparison possible. Subclassification does not rely on any particular functional form, such as linearity, for the relation between the outcome (death) and the covariate (age) within each treatment group, whereas models do. If the groups have similar distributions of the covariate, such specific assumptions like linearity are usually harmless, but when the groups have different covariate distributions, model-based methods of adjustment are dependent on the specific form of the model (for example, linearity or log linearity) and their results are determined by untrustworthy extrapolations.

If standard models can be so dangerous, why are they commonly used for such adjustments when large databases are examined for estimates of causal effects? One reason is the ease with which automatic data analysis can be done using existing, pervasive software on plentiful, speedy hardware. A second reason is the seeming difficulty of using subclassification when many confounding covariates need adjustment, which is the common case. Standard modeling software can automatically handle many regressor variables and produce results, although they can be remarkably misleading. With many confounding covariates, however, the issues of

lack of adequate overlap and reliance on untrustworthy model-based extrapolations are even more serious than with only one confounding covariate. The reason is that small differences in many covariates can accumulate into a substantial overall difference. For example, if members of one treatment or exposure group are slightly older, have slightly higher cholesterol levels, and have slightly more familial history of cancer, that group may be substantially less healthy. Moreover, although standard comparisons of means between the groups like those in **Table 1**, or comparisons of histograms for each confounding covariate among groups are adequate with one covariate, they are inadequate with more than one. The groups may differ in a multivariate direction to an extent that cannot be discerned from separate analyses of each covariate. This multivariate direction is closely related to the statistical concept of the best linear discriminant and intuitively is the single combination of the covariates on which the treatment groups are farthest apart.

Subclassification techniques can be applied with many covariates with almost the same reliability as with only one covariate. The key idea is to use propensity score techniques, as developed by Rosenbaum and Rubin (1). These methods can be viewed as important extensions of discriminant matching techniques, which calculate the best linear discriminant between the treatment groups and match on it (4).

Since their introduction approximately 15 years ago, propensity score methods have been used in various applied problems in medical and other research disciplines (5–23) but not nearly as frequently as they should have been relative to model-based methods.

Propensity Score Methods

Propensity score methods must be applied to groups two at a time. Therefore, an example with three treatment or exposure conditions will generally yield three distinct propensity scores, one for each comparison (for the example in **Table 1**, nonsmokers compared with cigarette smokers, nonsmokers compared with cigar and pipe smokers, and cigarette smokers compared with cigar and pipe smokers). To describe the way propensity scores work, I first assume two treatment conditions. Cases with more than two treatment groups are considered later.

The basic idea of propensity score methods is to replace the collection of confounding covariates in an observational study with one function of these covariates, called the propensity score (that is, the propensity to receive treatment 1 rather than treatment 2). This score is then used just as if it were the only confounding covariate. Thus, the collection of predictors is collapsed into a single predictor. The propensity score is found by predicting treatment group membership (that is, the indicator variable for being in treatment group 1 as opposed to treatment group 2) from the confounding covariates, for example, by a logistic regression or discriminant analysis. In this prediction of treatment group measurement, it is critically important that the outcome variable (for example, death) play no role; the prediction of treatment group must involve only the covariates. Each person in the database then has an estimated propensity score, which is the estimated probability (as determined by that person's covariate values) of being exposed to treatment 1 rather than treatment 2. This propensity score is then the single summarized confounding covariate to be used for subclassification.

Subclassification into about five groups on the basis of the propensity score then has the rather remarkable property of adjusting for all of the covariates that went into its estimation, no matter how many there are. This is a large-sample claim that relies on certain conditions dealt with in technical statistical publications, but it is nevertheless an extremely useful guide for practice. The intuition behind the validity of this claim is fairly straightforward and proceeds as follows.

If two persons, one exposed to treatment 1 and the other exposed to treatment 2, had the same value of the propensity score, these two persons would then have the same predicted probability of being assigned to treatment 1 or treatment 2. Thus, as far as we can tell from the values of the confounding covariates, a coin was tossed to decide who received treatment 1 and who received treatment 2. Now suppose that we have a collection of persons receiving treatment 1 and a collection of persons receiving treatment 2 and that the distributions of the propensity scores are the same in both groups (as is approximately true within each propensity subclass). In subclass 1, the persons who received treatment 1 were essentially chosen randomly from the pool of all persons in subclass 1, and analogously for each subclass. As a result, within each subclass, the multivariate distribution of the covariates used to estimate the propensity score differs only randomly between the two treatment groups.

The formal proof of this result appears in Rosenbaum and Rubin (1). Research on how well this theoretical result is satisfied when using estimated rather than true propensity scores is the topic of technical statistical publications (24–28). Generally, the conclusion is that using estimated propensity scores in place of true propensity scores works very well.

Table 2. Estimated 5-Year Survival Rates for Node-Negative Patients in Six Randomized Experiments*

Study	Treatment	Women	Estimated Survival Rate
		n	%
US-NCI†	Breast conservation	74	93.9
	Mastectomy	67	94.7
Milanese†	Breast conservation	257	93.5
	Mastectomy	263	93.0
French†	Breast conservation	59	94.9
	Mastectomy	62	95.2
Danish‡	Breast conservation	289	87.4
	Mastectomy	288	85.9
EORTC‡	Breast conservation	238	89.0
	Mastectomy	237	90.0
US-NSABP‡	Breast conservation	330	89.0
	Mastectomy	309	88.0

* Adapted from Table 2 in reference 3.
† Single-center trial.
‡ Multicenter trial.

Propensity Subclassification

Several years ago, the U.S. Government Accounting Office (29) summarized results from randomized experiments comparing mastectomy (removal of the breast but not the pectoral muscle with nodal dissection but no radiation) and breast conservation therapy (lumpectomy, nodal dissection, and radiation) for the treatment of breast cancer in node-negative patients. The results, shown in **Table 2** (29), provide no evidence of differential treatment effect, at least for the type of women who participated in these informed consent clinical trials and who received the kind of care dispensed at the centers participating in these trials. The question remained, however, how broadly these results could be generalized to other node-negative women and other medical facilities. The U.S. Government Accounting Office used the National Cancer Institute's SEER (Surveillance, Epidemiology and End Results) observational database to address this question. Restrictions (including node-negative diagnosis, 70 years of age or younger, and tumor size ≤4 cm [29]) were applied to match criteria for the randomized experiments. These restrictions reduced the database to 1106 women who received breast conservation therapy and 4220 who received mastectomy, for a total of 5326 women.

The U.S. Government Accounting Office used propensity score methods on the SEER database to compare the two treatments for breast cancer. First, approximately 30 potential confounding covariates and interactions were identified: year of diagnosis (1983–1985), age category (4 levels), tumor size, geographical registry (9 levels), race (4 levels), marital status (4 levels), and interactions of year and registry. A logistic regression was then used to predict treatment (mastectomy compared with conservation therapy) from these confounding covariates on the basis of data from the 5326 women. Each woman was then assigned an estimated propensity score, which was her probability, on the basis of her covariate values, of receiving breast conservation therapy rather than mastectomy. The group was then divided into five subclasses of approximately equal size on the basis of the womens' individual propensity scores: 1064 in the most mastectomy-oriented subclass, 1070 in the next subclass, 1059 in the middle subclass, 1067 in the next subclass, and 1066 in the most breast conservation–oriented subclass.

Before examining any outcomes (5-year survival results), the subclasses were checked for balance with respect to the covariates. Propensity score theory claims that if the propensity scores are relatively constant within each subclass, then within each subclass, the distribution of all covariates should be approximately the same in both treatment groups. This balance was found to be satisfactory. If important within-subclass differences between treatment groups had been found on some covariates, then either the propensity score prediction model would need to be reformulated or it would have been concluded that the covariate distributions did not overlap sufficiently to allow subclassification to adjust for these covariates. This process of cycling between checking for balance on the covariates and reformulating the propensity score model is described by Rosenbaum and Rubin (18) in the context of a study investigating coronary bypass surgery. For example, when the variances of an important covariate were found to differ importantly between treatment and control groups, then the square of that covariate was included in the revised propensity score model. For another example, if the correlations between two important covariates differed between the groups, then the product of the covariates was added to the propensity score model.

The estimates of 5-year survival rates made on basis of the resulting propensity score subclassification are given in **Table 3** (29). Total rates and rates excluding deaths unrelated to cancer are shown. Several features of **Table 3** are particularly striking, especially when compared with the results of the randomized experiments shown in **Table 2**. First, the general conclusion of similar performance of both treatments is maintained. Second, although overall survival is similar across treatment groups, the results indicate that survival in general practice may be slightly lower than suggested by data from the population of women and types of clinics participating in the randomized clinical trials, especially in the single-clinic studies.

Third, results slightly indicate that, in general practice, women and their physicians may be mak-

Table 3. Estimated 5-Year Survival Rates for Node-Negative Patients in the SEER Database within Each of Five Propensity Score Subclasses*

Propensity Score Subclass	Treatment	Women	Estimated Survival Rate for Women	Omitting Women Whose Deaths Were Unrelated to Cancer	Estimated Survival Rates Omitting Women Whose Deaths Were Unrelated to Cancer
		n	%	*n*	%
1	Breast conservation	56	85.6	54	88.8
	Mastectomy	1008	86.7	966	90.5
2	Breast conservation	106	82.8	102	86.0
	Mastectomy	964	83.4	917	87.7
3	Breast conservation	193	85.2	184	89.4
	Mastectomy	866	88.8	841	91.4
4	Breast conservation	289	88.7	279	92.0
	Mastectomy	978	87.3	742	91.5
5	Breast conservation	462	89.0	453	90.7
	Mastectomy	604	88.5	589	90.7

* Adapted from Tables 5 and 7 in reference 3. SEER = Surveillance, Epidemiology, and End Results.

ing beneficial choices. More precisely, women in propensity subclasses 1 to 3, composed of patients whose characteristics (including age, size of tumor, and region of country) make them relatively more likely to receive mastectomy than breast conservation therapy, seem to show better 5-year survival with mastectomy than with breast conservation therapy. In contrast, for women in propensity subclasses 4 and 5 (whose characteristics make them relatively more likely to receive breast conservation therapy than mastectomy), there seems to be no advantage to mastectomy and possibly a slight advantage to breast conservation therapy. Of course, this last interpretation is subject to two caveats. First, we only adjusted for the covariates that were used to estimate the propensity score; hence, other hidden covariates may alter this interpretation. In a randomized experiment, the effects of these hidden covariates are reflected in the SEs of the estimates, but in an observational study, these effects can create bias not reflected in the SEs. Second, the sampling variability (that is, SEs) of the results do not permit firm conclusions, even if the collection of confounding covariates was sufficient to remove bias in this observational study.

Although there is no randomized assignment in the SEER database, the propensity score analyses seem to provide useful suggestive results, especially when coupled with the results of the randomized experiments, with which they are consistent.

More Than Two Treatment Conditions

With more than two treatment conditions, the propensity score usually differs for each pair of treatment groups being compared (that is, with three treatment groups labelled A, B, and C, there are three propensity scores: A compared with B, A compared with C, and B compared with C). At first, this may seem to be a limitation of propensity score technology relative to a model-based analysis, but in fact it is an important strength and points to further weaknesses in a model-based approach. We show this by exploring a range of hypothetical modifications to Cochran's (2) smoking example.

First, consider what we could have learned if the nonsmokers and cigarette smokers had had adequately overlapping age distributions, but the cigar and pipe smokers had been substantially older than persons in either of the other groups, with essentially no overlap with the cigarette smokers or the nonsmokers. Even with only one covariate, with more than two groups, the groups in one two-group comparison (nonsmokers compared with cigarette smokers) may overlap adequately, whereas for all other comparisons (in this example, those involving cigar and pipe smokers), the overlap may be inadequate. A typical model-based analysis would use all the data to provide estimates for all three two-group comparisons, even using the data from the cigar and pipe smokers to influence the comparison between the nonsmokers and the cigarette smokers, with no warning of either the extreme extrapolations involved in two of the three two-group comparisons or the use of data on cigar and pipe smokers to help estimate the comparison of nonsmokers and cigarette smokers.

Let us again modify the Cochran (2) smoking example but now include an additional covariate: an index of socioeconomic status. We assume that nonsmokers and cigarette smokers have adequate overlap in their age distributions but not much overlap in their socioeconomic status distributions, with nonsmokers having higher socioeconomic status values. In contrast, we suppose that nonsmokers and cigar and pipe smokers have substantial overlap in their socioeconomic distributions but have essentially no overlap in their age distributions. This scenario illustrates that with more than two groups and more than one covariate, the comparison of one

pair of groups can be compromised by one covariate and the comparison of another pair of groups can be compromised by a different covariate. As discussed earlier, typical model-based analyses provide no warning that comparisons may be based on extreme extrapolations, nor do they show that the extrapolations include data from groups that are not in the pair of groups being compared.

Now suppose that the nonsmokers and cigarette smokers have the same age distributions and adequately overlapping socioeconomic status distributions. For this comparison, age needs no adjustment but socioeconomic status does need to be adjusted. The propensity score for the comparison would essentially equal socioeconomic status because it, and not age, would predict being a cigarette smoker as opposed to being a nonsmoker. Thus, for this comparison, adjusting for the propensity score would be the same as adjusting for socioeconomic status. Now also assume that the nonsmokers and cigar and pipe smokers have the same socioeconomic status distributions, so that socioeconomic status needs no adjustment, and have adequately overlapping age distributions that need adjustment. Then the propensity score for this comparison would equal age, and therefore, adjusting for the propensity score would be the same as adjusting for age. Thus, the propensity score for a comparison of one pair of groups generally needs to be different from that for a comparison of a different pair of groups. To complete the current scenario, assume that cigarette smokers and cigar and pipe smokers have adequate overlap in both age and socioeconomic status and that both need adjustment. The propensity score for this comparison would involve both age and socioeconomic status because both help to predict cigarette group membership, as opposed to cigar and pipe smoking group membership, and adjusting for this propensity score would adjust for both age and socioeconomic status. Clearly, different propensity score models are needed to adjust appropriately for different comparisons. Estimating all effects by using one model in our example with three groups and adequate overlap on all covariates can be even more deceptive than estimation in the two-group setting because the model being used to compare one pair of groups (for example, nonsmokers compared with cigarette smokers) is affected by the data from the third group (here cigar and pipe smokers), which probably has covariate values that differ from those in either one of the other two groups being compared.

Limitations of Propensity Scores

Despite the broad utility of propensity score methods, when addressing causal questions from nonrandomized studies, it is important to keep in mind that even propensity score methods can only adjust for observed confounding covariates and not for unobserved ones. This is always a limitation of nonrandomized studies compared with randomized studies, where the randomization tends to balance the distribution of all covariates, observed and unobserved.

In observational studies, confidence in causal conclusions must be built by seeing how consistent the obtained answers are with other evidence (such as results from related experiments) and how sensitive the conclusions are to reasonable deviations from assumptions, as illustrated by Connors and colleagues (20), who used techniques from Rosenbaum and Rubin's work (30). Such sensitivity analyses suppose that a relevant but unobserved covariate has been left out of the propensity score model. By explicating how this hypothetical unmeasured covariate is related to treatment assignment and outcome, we can obtain an estimate of the treatment effect that adjusts for it as well as for measured covariates and hereby investigate how answers might change if such a covariate were available for adjustment. Of course, medical knowledge is needed when assessing whether the posited relations involving the hypothetical unmeasured covariate are realistic or extreme. Clarifications of nomenclature and extended sensitivity analyses reported by Lin and colleagues (31) moderate the initial conclusions of Connors and colleagues (20).

Another limitation of propensity score methods is that they work better in larger samples for the following reason. The distributional balance of observed covariates created by subclassifying on the propensity score is an expected balance, just as the balance of all covariates in a randomized experiment is an expected balance. In a small randomized experiment, random imbalances of some covariates can be substantial despite randomization; analogously, in a small observational study, substantial imbalances of some covariates may be unavoidable despite subclassification using a sensibly estimated propensity score. The larger the study, the more minor are such imbalances.

A final possible limitation of propensity score methods is that a covariate related to treatment assignment but not to outcome is handled the same as a covariate with the same relation to treatment assignment but strongly related to outcome. This feature can be a limitation of propensity scores because inclusion of irrelevant covariates reduces the efficiency of the control on the relevant covariates. However, recent work (28) suggests that, at least in modest or large studies, the biasing effects of leaving out even a weakly predictive covariate dominate the efficiency gains from not using such a

covariate. Thus, in practice, this limitation may not be substantial if investigators use some judgment.

Conclusion

Large databases have tremendous potential for addressing (although not necessarily settling) important medical questions, including important causal questions involving issues of policy. Addressing these causal questions using standard statistical (or econometric, psychometric, or neural net) models can be fraught with pitfalls because of their possible reliance on unwarranted assumptions and extrapolations without any warning. Propensity score methods are more reliable; they generalize the straightforward technique of subclassification with one confounding covariate to allow simultaneous adjustment for many covariates. One critical advantage of propensity score methods is that they can warn the investigator that, because of inadequately overlapping covariate distributions, a particular database cannot address the causal question at hand without relying on untrustworthy model-dependent extrapolation or restricting attention to the type of person adequately represented in both treatment groups. Because of this advantage, any causal questions put to a large database should be first approached using propensity score methods to see whether the question can be legitimately addressed. If so, subclassification on a well-estimated propensity score can be used to provide reliable results, which are adjusted for the covarites used to estimate the propensity score and which can be clearly displayed. After that, modeling can play a useful role. For example, standard statistical models, such as lest-squares regression, can be safely applied within propensity score subclasses to adjust for minor within-subclass differences in covariate distributions between treatment groups. This was done in the example of the study by the U.S. Government Accounting Office (29). Of course, it always must be remembered that propensity scores only adjust for the observed covariates that went into their estimation.

Grant Support: In part by a grant from the National Science Foundation (SES-9207456).

Acknowledgments: The author thanks Jennifer Hill and Frederick Mosteller for helpful editorial comments on an earlier draft of this article.

Requests for Reprints: Donald B. Rubin, PhD, Harvard University, Department of Statistics, Science Center, 6th Floor, 1 Oxford Street, Cambridge, MA 02138.

References

1. **Rosenbaum P, Rubin DB.** The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41-55.
2. **Cochran WG.** The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics. 1968;24:295-313.
3. **Finch PE.** Standardization. In: Kotz S, Johnson NL, eds. Encyclopedia of Statistical Sciences. v 8. New York: Wiley; 1988:629-32.
4. **Rubin DB.** Bias reduction using Mahalanobis' metric matching. Biometrics. 1980:36;295-8.
5. **Aiken LH, Smith HL, Lake ET.** Lower Medicare mortality among a set of hospitals known for good nursing care. Med Care. 1994;32:771-87.
6. **Cook EF, Goldman L.** Asymmetric stratification. An outline for an efficient method for controlling confounding in cohort studies. Am J Epidemiol. 1989; 127:626-39.
7. **Cook EF, Goldman L.** Performance of tests of significance based on stratification by a multivariate confounder score or by a propensity score. J Clin Epidemiol. 1989;42:317-24.
8. **Eastwood EA, Fisher GA.** Skills acquisition among matched samples of institutionalized and community-based persons with mental retardation. Am J Ment Retard. 1988;93:75-83.
9. **Fiebach NH, Cook EF, Lee TH, Brand DA, Rouan GW, Weisberg M, et al.** Outcomes in patients with myocardial infarction who are initially admitted to stepdown units: data from the Multicenter Chest Pain Study. Am J Med. 1990;89:15-20.
10. **Harrell FE Jr, Marcus SE, Layde PM, Broste SK, Cook EF, Wagner DP, et al.** Statistical methods in SUPPORT. J Clin Epidemiol. 1990;43(Suppl):89S-98S.
11. **Kane R, Garrad J, Buchanan J, Rosenfeld A, Skay C, McDermott S.** Improving primary care in nursing homes. J Am Geriatr Soc. 1991;39:359-67.
12. **Lavori PW, Keller MB, Endicott J.** Improving the validity of FH-RDC diagnosis of major affective disorder in uninterviewed relatives in family studies: a model based approach. J Psychiatr Res. 1988;22:249-59.
13. **Lavori PW, Keller MB.** Improving the aggregate performance of psychiatric diagnostic methods when not all subjects receive the standard test. Stat Med. 1988;7:727-37.
14. **Myers WO, Gersh BJ, Fisher LD, Mock MB, Holmes DR, Schaff HV, et al.** Medical versus early surgical therapy in patients with triple-vessel disease and mild angina pectoris: a CASS registry study of survival. Ann Thorac Surg. 1987;44:471-86.
15. **Stone RA, Obrosky DS, Singer DE, Kapoor WN, Fine MJ.** Propensity score adjustment for pretreatment differences between hospitalized and ambulatory patients with community-acquired pneumonia. Pneumonia Patient Outcomes Research Team (PORT) Investigators. Med Care. 1995;33(4 Suppl): AS56-66.
16. **Willoughby A, Graubard BI, Hocker A, Storr C, Vietze P, Thackaberry JM, et al.** Population-based study of the developmental outcome of children exposed to chloride-deficient infant formula. Pediatrics. 1990;85:485-90.
17. **Drake C, Fisher L.** Prognostic models and the propensity score. Int J Epidemiol. 1995;24:183-7.
18. **Rosenbaum PR, Rubin DB.** Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association. 1984;79:516-24.
19. **Reinisch J, Sanders SA, Mortensen EL, Rubin DB.** In utero exposure to phenobarbital and intelligence deficits in adult men. JAMA. 1995;274:1518-25.
20. **Connors AF Jr, Speroff T, Dawson NV, Thomas C, Harrell FE Jr, Wagner D, et al.** The effectiveness of right heart catheterization in the initial care of critically ill patients. SUPPORT Investigators. JAMA. 1996;276:889-97.
21. **Gu XS, Rosenbaum PR.** Comparison of multivariate matching methods: structures, distances, and algorithms. Journal of Computational and Graphical Statistics. 1993;2:405-520.
22. **Rubin, DB, Rosenbaum PR.** Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician. 1985;39:33-8.
23. **Rosenbaum PR, Rubin DB.** The bias due to incomplete matching. Biometrics. 1985;41:103-16.
24. **Rubin DB.** Assessing the fit of logistic regressions using the implied discriminant analysis. Discussion of "Graphical Methods for Assessing Logistic Regression Models" by Landwehr, Pregibone, and Smith. Journal of the American Statistical Association. 1984;79:79-80.
25. **Drake C.** Effects of misspecification of the propensity score on estimators of treatment effect. Biometrics. 1993;49:1231-36.
26. **Rubin DB, Thomas N.** Affinely invariant matching methods with ellipsoidal distributions. The Annals of Statistics. 1992;20:1079-93.
27. **Rubin DB, Thomas N.** Characterizing the effect of matching using linear propensity score methods with normal covariates. Biometrika. 1992;79:797-809.
28. **Rubin DB, Thomas N.** Matching using estimated propensity scores: relating theory to practice. Biometrics. 1996;52:249-64.
29. **U.S. General Accounting Office.** Breast conservation versus mastectomy: patient survival in day-to-day medical practice and randomized studies: report to the chairman, Subcommitee on Human Resources and Intergovernmental Relations, Committee on Government Operations, House of Representatives Washington, DC: U.S. General Accounting Office; 1994. Report #GAO-PEMD-95-9.
30. **Rosenbaum PR, Rubin DB.** Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. The Journal of the Royal Statistical Society. Series B. 1983;45:212-8.
31. **Lin DY, Psaty BM, Kronmal RA.** Assessing the sensitivity of regression results to unmeasured confounders in observational studies. Seattle: Univ Washington School of Public Health; 1997. Technical Report #144.