Nonrandomized Comparative Clinical Studies -
Proceedings of the International Conference on Nonrandomized
Comparative Clinical Studies in Heidelberg, April 10 -11,1997
Printed volume

Estimation from nonrandomized treatment comparisons using subclassification on propensity scores

(This article is a modification and expansion of an article in *Annals of Internal Medicine, 127*, 8(2), pp. 757-763.

D. B. Rubin

Abstract The aim of many analyses of medical data sets is to draw causal inferences about the relative effects of treatments, such as different methods of treating cancer patients. The data available to compare many such treatments are not based on the results of carefully conducted randomized clinical trials, but rather are collected while observing 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 data sets, even if only to help design a new randomized experiment or shed light on the generalizability of results from existing randomized experiments. Standard methods of analysis using routine statistical software (e.g., linear or logistic regressions), however, can be quite 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. Subclassification on propensity scores is a particularly straightforward technique and is the topic of this article.

Propensity score technology in observational studies

The objective of many medical studies is the estimation of the causal effects of some new treatment or exposure relative to a control condition (e.g., the effect of smoking on mortality). In the vast majority of such studies, there is the need to control for naturally occurring systematic differences in background characteristics between the treatment group and the control group (e.g., in age or sex distributions), systematic differences which would not occur in the context of a randomized experiment. Typically, there are many background characteristics that need to be controlled.

Propensity score technology, introduced by Rosenbaum and Rubin (1983a), 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 treated and control groups overlap enough on background characteristics to allow sensible estimation of treatment versus control effects from this data set. Moreover, when such overlap is present, the propensity score approach allows straightforward calculation of estimated treatment versus control effects that reflect adjustment for differences in all observed background characteristics. Subclassification on the propensity score is a particularly straightforward technique for such adjustment.

Subclassification on One Confounding Variable

Before describing how subclassification on propensity scores can be used in the statistical analysis of an observational study with many confounding background characteristics, we begin with an example showing how subclassification can be used to adjust for a single confounding covariate, such as age, in a study of smoking and mortality. We then show how propensity scores methods can be used to generalize subclassification on a single confounding covariate to the case with many confounding covariates, such as age, region of the country, and sex. The potential for an observational data base (i.e., not from a randomized experiment) to suggest causal effects of treatments is indicated by Table 1, adapted from Cochran (1968), which concerns mortality rates per thousand in three large data bases from the U.S., the U.K., and Canada for nonsmokers, cigarette smokers, and cigar and pipe smokers. The treatment factor here involves the three levels of smoking. It appears from the death rates in Part A of Table 1 that cigarette smoking is good for health, especially relative to cigar and pipe smoking, clearly a result contrary to current wisdom. A problem with the naive conclusion from Part A is exposed in Part B of Table 1, which gives the average ages of the subpopulations: age is correlated with both death rates and smoking behavior. Age in this example is a "confounding" covariate, and conclusions regarding the effects of smoking should be adjusted for differences in age distributions across subpopulations. A straightforward way of adjusting for age is to: (1) divide the population into age categories of approximately equal size (e.g., 2 categories = younger, older; or 3 categories = young, middle-age, old; or 4 categories, etc.); (2) compare death rates within an age category (e.g., within the younger population, compare death rates for the three treatment groups and similarly for the older population); and (3) average over the age-group-specific comparisons to obtain overall estimates of the age-adjusted death rates per 1000 for each of the three treatment groups. Part C of Table 1 shows the results for different numbers of categories of age, where the subclass age boundaries were defined to have equal numbers of nonsmokers in each subclass. These results, especially with 9-11 subclasses, align better than Part A with our current understanding of the effects of smoking. 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" (Finch, 1988) to the age distribution of nonsmokers as defined by their subclass counts. Cochran (1968) calls this method "subclassification" and offers theoretical results showing that as long as the treatment groups overlap in their age distributions (i.e., as long as there are reasonable numbers of subjects from each treatment condition in each subclass), comparisons using 5 or 6 subclasses will typically remove 90% or more of the bias present in the raw comparisons in Part A. More than five subclasses were used in the final rows of Part C in Table 1 because the large sizes of the data sets made it possible to do so.

A particular statistical model such as a linear regression (or a logistic regression, 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.

Tabel 1: Comparing Death Rates for Three Smoking Groups in each of Three Data Bases from Tables 1-3 in Cochran (1968)

	Canadian Study			UK Study			US Study		
	No Smoke	Cigarette	Cigar & Pipe	No Smoke	Cigarette	Cigar & Pipe	No Smoke	Cigarette	Cigar & Pipe
A	Death Rates per 1,000 Person Years								
	20.2	20.5	35.5	11.3	14.1	20.7	13.5	13.5	17.4
B	Average Age in Years								
	54.9	50.5	65.9	49.1	49.8	55.7	57.0	53.2	59.7
C	Adjusted Death Rates Using K Subclasses								
K=2	20.2	26.4	24.0	11.3	12.7	13.6	13.5	16.4	14.9
K=3	20.2	28.3	21.2	11.3	12.8	12.0	13.5	17.7	14.2
K=9-11	20.2	29.5	19.8	11.3	14.8	11.0	13.5	21.2	13.7

First, if the treatment groups do not adequately overlap on the confounding covariate age, the investigator will see it immediately and be warned. Thus, if members of one treatment group have ages outside the range of another group's ages, it will be obvious, because one or more age-specific subclasses will consist solely of members exposed to one treatment (or nearly so). In contrast, there is nothing in the standard output of any regression modeling software that will display this critical fact. The reason for this apparent omission is that such models predict an outcome (e.g., mortality) from regressors (e.g., age and treatment indicators), and standard regression diagnostics do not include the careful analysis of the joint distribution of the regressors (e.g., a comparison of the distributions of age across treatment groups). When the overlap on age distributions across treatment groups is too limited, the data base, no matter how large, cannot support causal conclusions about the differential effects of the treatments. For an extreme example, if the data base consists of 70 year-old smokers and 40 year-old nonsmokers, the comparison of 5-year survival rates among 70 year-old smokers and 40-year old nonsmokers provides essentially no information about the effect of smoking versus nonsmoking for either 70 year-olds or 40-year olds, or any other age group.

The second reason for preferring subclassification to models concerns more promising situations like that in Table 1, where the treatment groups overlap enough on the confounding covariate so that a comparison is possible. When estimating the treatment effect, subclassification does not rely on any particular functional form (e.g., linearity) for the relationship between the outcome (mortality) and covariate (age) within each treatment group, whereas models do rely on such assumptions. If the treatment groups have similar distributions of the covariate, common assumptions like linearity are usually harmless, but when the treatment groups have rather different covariate distributions, model-based methods of adjustment are dependent on the specific form of the model (e.g., linearity, log-linearity), and their answers are influenced by untrustworthy extrapolations. Simulations documenting the fragility of linear regression methods appear in Rubin (1973) for the case of one covariate.

If standard models can be so dangerous, why are they so commonly used for such adjustments when examining data bases for estimates of causal effects? One reason is the ease of automatic data analysis using existing, pervasive software on plentiful, speedy hardware. Nevertheless, although standard modeling software can automatically "handle" many regressor variables and produce results,

these results can be remarkably misleading. In fact, when there are many confounding covariates, the issues of lack of adequate overlap and reliance on untrustworthy model-based extrapolations are even more serious than with only one confounding covariate, as documented by simulations in Rubin (1979, Table 2). One reason for the increased problem is that small differences on many covariates can accumulate into a substantial overall difference. For example, if one treatment group is a little older, has a little higher cholesterol, has a little more familial history of cancer, and so on, that group may be substantially less healthy. Another reason for the increased problem with many covariates rather than one covariate is that diagnosing nonlinear relationships between outcomes and many covariates is more complicated. Moreover, standard comparisons of means between the groups, like those in Table 1B, or even comparisons of histograms for each confounding covariate between the treatment groups, although adequate with one covariate, 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.

A second reason for the dominance of modeling over subclassification is the seeming difficulty of using subclassification when many confounding covariates, rather than one, need adjustment, which is the common case. Fortunately, subclassification techniques can be applied with many covariates with nearly the same reliability as with only one covariate. The key idea is to use "propensity score" techniques introduced by Rosenbaum and Rubin (1983a); these can be viewed as important extensions of discriminant matching techniques, which calculate the best linear discriminant between the treatment groups and match on it (Rubin, 1980). Since their introduction a decade and a half ago, propensity score methods have been used in a variety of applied problems in medical and other research disciplines (Aiken, Smith and Lake, 1994; Connors et alia, 1996; Cook and Goldman, 1988; Cook and Goldman, 1989; Drake and Fisher, 1995; Eastwood and Fisher, 1988; Fiebach et alia, 1990; Gu and Rosenbaum, 1993; Harrell et alia, 1990; Kane et alia, 1991; Lavori and Keller, 1988; Lavori, Keller and Endicott, 1988; Malloy et alia, 1990; Myers et alia, 1987; Reinisch, Sanders, Mortensen and Rubin, 1995; Rosenbaum and Rubin, 1984; Rosenbaum and Rubin, 1985a; Stone et alia, 1995; Willoughby et alia, 1990;). Nevertheless, propensity score methods have not been used nearly as frequently as they should have been relative to model-based methods.

Propensity Score Methods

Propensity score methods generally have to be applied to treatment groups two at a time. Therefore in an example with three treatment conditions, there are generally three distinct propensity scores, one for each two-group treatment comparison (e.g., for the example of Table 1, nonsmokers versus cigarette smokers, nonsmokers versus cigar and pipe smokers, and cigarette smokers versus cigar and pipe smokers). To describe the way propensity scores work, we therefore assume two treatment conditions. Situations 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 the observational study with one function of these covariates, called the propensity score (i.e., the propensity to receive treatment 1 rather than treatment 2), and then to use this score just as if it were the only confounding covariate. Thus the collection of predictors is collapsed into a single composite predictor. The propensity score is found by predicting treatment group membership (i.e., the indicator variable for being in treatment 1 versus treatment 2) from the confounding covariates, for example by a logistic regression or a discriminant analysis. In this prediction of treatment group membership, it is critically important that the outcome variable (e.g., mortality) plays no role; the prediction of treatment group only involves the covariates. Each subject in the data base then has an estimated propensity score, which is the estimated probability, as determined by that subject's covariate values, of being exposed to treatment 1 versus treatment 2. This propensity score is then the single summarized confounding covariate to be used for subclassification.

Subclassification into 5 or more groups on 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! This is a "large-sample" claim that relies on certain conditions addressed in technical statistical publications (Rosenbaum and Rubin, 1983a; Rubin and Thomas, 1992a, 1992b), but nevertheless it is an extremely useful guide for practice (Rubin and Thomas, 1996). The intuition behind the claim's validity is fairly straightforward and proceeds as follows: Suppose that two subjects, one exposed to treatment 1 and the other exposed to treatment 2, were presented to us with the same value of the propensity score. These two subjects would then have the same predicted probability of being assigned to treatment 1 versus treatment 2, and thus, as far as we can tell from their values of the confounding covariates, a coin was tossed to decide which one received treatment 1 and which one received treatment 2. Now suppose that we have a collection of treatment 1 subjects and a collection of treatment 2 subjects, such that the distributions of the propensity scores are the same in both groups, as is approximately true within each propensity subclass. Then in subclass 1, the subjects who received treatment 1 were essentially randomly chosen from the pool of all subjects 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 with true propensity scores appears in Rosenbaum and Rubin (1983a). Research on how well this theoretical result is satisfied when using estimated rather than true propensity scores is the topic of technical statistical publications (Drake, 1993; Rubin, 1984; Rubin and Thomas, 1992a, 1992b, 1996). Generally, the conclusion is that using estimated propensity scores in place of true propensity scores works very well.

Example - Propensity Subclassification

Several years ago the U.S. Government Accounting Office (GAO, 1994) summarized results from randomized experiments comparing mastectomy (removal of breast, but not the pectoral muscle, plus nodal dissection but no radiation) and breast-conservation therapy (lumpectomy, nodal dissection and radiation) for the treatment of breast cancer for node-negative patients. Table 2 is adopted from their Table 2, and the results there provide no evidence of any differential treatment effect, at least for the type of women who participated in these informed-consent clinical trials and received the kind of care dispensed at the centers participating in these trials. The question remained, however, how broadly these results could be generalized, i.e., to other node-negative women and other medical facilities. The GAO used the National Cancer Institute's SEER (Surveillance, Epidemiology and End Results) observational data base to address this question. Restrictions (e.g., node-negative diagnosis, age 70 or younger, tumor 4 cm or smaller, etc., as detailed in GAO (1994) in its Tables 4 and I.3) were applied to correspond to criteria for the randomized experiments, and these reduced the data base to 1,106 women receiving breast-conservation therapy and 4,220 receiving mastectomy. GAO 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 versus conservation therapy) from these confounding covariates based on the data from the 5,326 (1,106 + 4220) women. Each woman was then assigned an estimated propensity score—her estimated probability, based on her covariate values, of receiving breast conservation therapy rather than mastectomy. The group of 5,326 was then divided into FIVE approximately equal-size subclasses based on their individual propensity scores, just as if these propensity scores comprised the only covariate: 1,064 were in the most mastectomy-oriented subclass, 1,070 in the next subclass, 1,059 in the middle subclass, 1,067 in the next subclass, and 1,066 were in the most breast-conservation-oriented subclass.

Before examining any outcomes (i.e., any 5-year survival results) — and the "before" is critical, the subclasses were checked for balance on the covariates. Recall that 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 in Rosenbaum and Rubin (1984) in the context of a study investigating coronary bypass surgery. For example, if the variances of an important covariate were found to differ importantly between treatment and control groups, then the square of that covariate would have been 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 would have been added to the propensity score model. If "checking for balance" had been allowed to include the examination of estimated causal effects, then the selection of a particular propensity score model could have been used to bias the estimate of the causal effect in a "preferred" direction. This point is critical: the unbiased design of an observational study requires us to check for balance in covariates without allowing the influence of the associated estimates of causal effects.

For the GAO study, the estimates of 5-year survival rates based on the resulting propensity score subclassification are given in Table 3, taken from Tables 5 and 7 in GAO (1994); both total rates and rates excluding deaths unrelated to cancer are presented. Several features of this table are particularly striking, especially when compared to the randomized experiments' results in Table 2.

Table 2: Estimated 5-year Survival Rates for Node-Negative Patients in Six Randomized Experiments; from Table 2 in U.S. GAO Report (1994).

Study	Treatment	n	Estimate
US-NCI[1]	Breast Conservation	74	93.9%
	Mastectomy	67	94.7%
Milan[1]	Breast Conservation	257	93.5%
	Mastectomy	263	93.0%
French[1]	Breast Conservation	59	94.9%
	Mastectomy	62	95.2%
Danish[2]	Breast Conservation	289	87.4%
	Mastectomy	288	85.9%
EORTC[2]	Breast Conservation	238	89.0%
	Mastectomy	237	90.0%
US-NSABP[2]	Breast Conservation	330	89.0%
	Mastectomy	309	88.0%

[1] single center [2] multicenter

Table 3: Estimated 5-year Survival Rates for Node-Negative Patients in SEER Data Base Within Each of Five Propensity Score Subclasses; from Tables 5 and 7 in U.S. GAO Report (1994).

Propensity Score Subclass	Treatment	n	Estimate	n*	Estimate*
1	Breast Conservation	56	85.6%	54	88.8%
	Mastectomy	1,008	86.7%	966	90.5%
2	Breast Conservation	106	82.8%	102	86.0%
	Mastectomy	964	82.8%	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%

* omitting patients whose deaths were unrelated to cancer.

First, the general conclusion of similar performance of both treatments is maintained. Second, although overall survival is quite similar across treatment groups, there is an indication that survival in general practice may be slightly lower than suggested from the population of women and type of clinic participating in the randomized clinical trials, especially in the single clinic studies.
Third, there is a slight indication that in general practice, women and their doctors may be making efficacious choices. More precisely, women in propensity subclasses 1-3, which are 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 under mastectomy than under breast conservation surgery. In contrast, for propensity subclasses 4-5, composed of patients whose characteristics make them relatively more likely to receive breast conservation therapy than mastectomy, there appears to be no advantage to mastectomy, and possibly a slight advantage to breast conservation therapy. Of course, this third interpretation is subject to two caveats. First, we have only adjusted for the covariates that were used to estimate the propensity score and hence other hidden covariates might alter this interpretation; in a randomized experiment, the effects of these "hidden" covariates are reflected in the standard errors of the estimates, but in an observational study these effects create bias not reflected in standard errors. Second, the sampling variability (i.e., standard errors) of the results do not permit firm conclusions about this point, even if the collection of confounding covariates used to estimate the propensity score were sufficient to remove all bias in this observational study.

The basic conclusion of the GAO analyses is, however clear: Even though there is no randomized assignment in the SEER data base, the propensity score analyses do appear to provide useful suggestive results, especially when coupled with the results of the randomized experiments, with which they are scientifically consistent.

More Than Two Treatment Conditions

With more than two treatment condition, there is generally a different propensity score for each pair of treatment groups being compared (i.e., with three treatment groups labeled A, B, and C, there are three propensity scores: A vs. B, A vs. C, and B vs. 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 not a limitation but an important strength and points to further weaknesses in a model-based approach. We see this by exploring a range of hypothetical modifications to Cochran's (1968) smoking example.

First consider what we could have learned if the nonsmokers and cigarette smokers had adequately overlapping age distributions, but the cigar/pipe smokers were substantially older than either of the other groups, with essentially no overlap with either the cigarette smokers or the nonsmokers. When there are more than two groups, one particular two-group comparison (nonsmokers versus cigarette smokers in this example) may have adequate overlap, whereas the other comparisons (those involving cigar/pike smokers in this example) may have inadequate overlap. A typical model-based analysis would use all the data to provide estimates for all three two-group comparisons, even using cigar/pipe smokers' data to influence the nonsmokers versus cigarette smokers comparison, with no warning of either (a) the extreme extrapolations involved in two of the three two-group comparisons or (b) the use of the cigar/pipe smokers data to help compare the nonsmokers and cigarette smokers, even though the cigar/pipe smokers are substantially older than both the nonsmokers and the cigarette smokers.

Let us again modify the Cochran smoking example, but now include an additional covariate, some index of socio-economic status, SES. Also suppose that nonsmokers and cigarette smokers have adequate overlap in their age distributions but not much overlap in their SES distributions, with nonsmokers typically having higher SES values. In contrast, suppose that nonsmokers and cigar/pipe smokers have substantial overlap in their SES 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 earlier, typical model-based analyses provide no warning that comparisons may be based on extreme extrapolations, nor that the extrapolations are using data from groups not in the pair of groups being compared. Now suppose that the nonsmokers and cigarette smokers have the same age distributions and adequately overlapping SES distributions. For this comparison, age needs no adjustment but SES needs to be adjusted. The propensity score for the comparison would essentially equal SES because it, and not age, would predict being a cigarette smoker versus being a nonsmoker; thus, for this comparison, adjusting for the propensity score would be the same as adjusting for SES. Also suppose that the nonsmokers and cigar/pipe smokers have the same SES distributions, so SES needs no adjustment, and adequately overlapping age distributions that need adjustment. The propensity score for this comparison would equal age, and so adjusting for it 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 than for a comparison of a different pair of groups. To complete the current scenario, suppose cigarette and cigar/pipe smokers had adequate overlap in both age and SES, and both needed adjustment. The propensity score for this comparison would involve both age and SES, because both help to predict cigarette group versus cigar/pipe group membership, and adjusting for this propensity score would adjust for both age and SES. Clearly, in general, different propensity scores models are needed to adjust appropriately for different comparisons. Estimating all effects using one model in this case 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 (e.g., nonsmokers versus cigarette smokers) is affected by the third group's data (e.g., cigar/pipe smokers), which possibly has covariate values rather different from either of the two groups being compared.

Limitations of Propensity Scores

Despite the broad utility of propensity score methods, it is important when addressing causal questions from nonrandomized studies to keep in mind that even propensity score methods can only adjust for observed confounding covariates and not unobserved ones. This is always a limitation of nonrandomized studies relative to 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 from related experiments) and how sensitive the conclusions are to reasonable deviations from assumptions, as illustrated in Connors et alia (1996) using techniques from Rosenbaum and Rubin (1983b). Such sensitivity analyses suppose that a relevant but unobserved covariate has been left out of the propensity score model. By explicating how this hypothetical unmeasurable covariate is related to treatment assignment and to outcome, we can obtain an estimate of the treatment effect that adjusts for it as well as measured covariates, and thereby investigate how answers might change if such a covariate were available for adjustment. Of course, medical knowledge is needed when accessing whether the posited relationships involving the hypothetical unmeasured covariate are realistic or extreme. Of particular relevance to Connors et alia (1996), clarifications of nomenclature and extended sensitivity analysis reported in Lin, Psaty and Kronmal (1997) moderate the initial conclusions in Connors et alia (1996).

Another limitation of propensity score methods is that they work better in larger samples for the same reason that completely randomized experiments work better in large samples. 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 the randomization, and 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. One way to create better balance in randomized experiments is to randomize within blocks of patients who are similar on prognostically important covariates. Just as blocking on such covariates can be beneficial in a randomized

experiment, blocking or matching on them in special ways can be used with propensity score methods (Rubin and Thomas, 1997).

Another possible limitation of propensity score methods is its handling of prognostically weak covariates included in the propensity score estimation. A covariate related to treatment assignment, but not to outcomes, is treated the same as a covariate with the same relationship with treatment assignment, but strongly related to outcomes. This feature can be a limitation of propensity scores in that the inclusion of irrelevant covariates reduces the efficiency of the control on the relevant covariates. Recent work, however, 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 (Rubin and Thomas, 1996). Thus, in practice, this limitation may not be substantial if investigators use some judgment. Finally, a current limitation in the application of propensity score methods concerns how to handle missing data in the covariates. In such a situation, the general objective is to achieve balance on the observed values of covariates and the observed patterns of missing data. The computational software required to achieve this objective is far more complex than for the case without missing data. Fortunately, progress is being made as described in D'Agostino and Rubin (1997).

Conclusion

Observational data bases can address, although not necessarily settle, important medical questions concerning causal effects of treatments. Addressing these causal questions using standard statistical (or econometric or psychometric, or neural net, etc.) models can be fraught with pitfalls because of their possible reliance on unwarranted assumptions and extrapolations without any warning. Subclassification on propensity scores is more reliable; it generalizes 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 data base cannot address the causal question at hand without either (a) relying on untrustworthy model-dependent extrapolations, or (b) restricting attention to the type of subject adequately represented in both treatment groups. Because of this advantage, any causal questions put to a data base should be first attacked using propensity score methods to see if the question can be legitimately addressed. If so, then subclassification on a well-estimated propensity score can be used to provide reliable results, which are adjusted for the covariates used to estimate the propensity score and which can be displayed in a transparent manner. After that, modeling can play a useful role. For example, standard statistical models, such as least squares regression, can be safely applied within propensity score subclasses to adjust for minor within-subclass differences in covariate distributions between treatment groups. This, in fact, was done in the U.S. GAO (1994) example. Of course, it always must be remembered that propensity scores only adjust for the observed covariates that went into their estimation.

Acknowledgements

Extremely helpful editorial comments on an earlier version of this article were provided by Jennifer Hill, Frederick Mosteller, the editorial staff of the Annals of Internal Medicine, and anonymous reviewers. The work for this article was partially supported by a grant from the National Science Foundation, Grant #SES-9207456.

References

[1]

Aiken, L., Smith, H. and Lake, E. (1994). “Lower Medicare mortality among a set of hospitals known for good nursing care.” Medical Care, 32, pp. 771-787.Aiken, L., Smith, H. and Lake, E. (1994). “Lower Medicare mortality among a set of hospitals known for good nursing care.” Medical Care, 32, pp. 771-787.

[2]

Cochran, W.G. (1968). “The effectiveness of adjustment by subclassification in removing bias in observational studies.” Biometrics, 24, pp. 295-313.

[3]

Connors, A.F. et alia (1996). “The effectiveness of right heart catheterization in the initial care of critically ill patients.” Journal of the American Medical Association, 276, pp. 889-897.

[4]

Cook, E.F. and Goldman, L. (1988). “Asymmetric stratification: an outline for an efficient method for controlling confounding in cohort studies.” American Journal of Epidemiology, 127, pp. 626-639.

[5]

Cook, E.F. and Goldman, L. (1989). “Performance of tests of significance based on stratification by a multivariate confounder score or by a propensity score.” Journal of Clinical Epidemiology, 42, pp. 317-324.

[6]

D’Agostino, R., Jr. and Rubin, D.B. (1997). “Estimation and Use of Propensity Scores with Incomplete Data.” Revision to appear in Journal of the American Statistical Association.

[7]

Drake, C. (1993). “Effects of misspecification of the propensity score on estimators of treatment effect.” Biometrics, 49, pp. 1231-1236.

[8]

Drake, C. and Fisher, L. (1995). “Prognostic Models and the Propensity Score.” International Journal of Epidemiology, 24, pp. 185-187.

[9]

Eastwood, E.A. and Fisher, G.A. (1988). “Skills acquisition among matched samples of institutionalized and community-based persons with mental-retardation.” American Journal on Mental Retardation, 93, pp. 75-83.

[10]

Fiebach, N.H. et alia (1990). “Outcomes in patients with myocardial-infarction who are initially admitted to stepdown units—data from the multicenter chest pain study.” American Journal of Medicine, 89, pp. 15-20.

[11]

Finch, P.E. (1988). “Standardization.” Kotz, S. and Johnson, N.L. (eds.), Encyclopedia of Statistical Sciences Volume 8. New York, Wiley. pp. 629-632.

[12]

Gu, X.S. and Rosenbaum, P.R. (1993). “Comparison of multivariate matching methods: structures, distances, and algorithms.” Journal of Computational and Graphical Statistics, 2, pp. 405-520.

[13]

Harrell, F.E. et alia (1990). “Statistical-methods in SUPPORT.” Journal of Clinical Epidemiology, 43, pp. S89-S98.

[14]

Kane, R. et alia (1991). “Improving primary care in nursing homes.” Journal of the American Geriatric Society, 39, pp. 359-367.

[15]

Lavori, P.W. and Keller, M.B. (1988). “Improving the aggregate performance of psychiatric diagnostic methods when not all subjects receive the standard test.” Statistics in Medicine, 7, pp. 723-737.

[16]

Lavori, P.W., Keller, M.B. and Endicott, J. (1988). “Improving the validity of Rh-Rdc diagnosis of major affective-disorder in uninterviewed relatives in family studies—A model based approach.” Journal of Psychiatric Research, 22, pp. 249-259.

[17]

Lin, D.Y., Psaty, B.M. and Kronmal, R.A. (1997). “Assessing the sensitivity of regression results to unmeasured confounders in observational studies.” Seattle, University of Washington School of Public Health, Department of Biostatistics, Technical Report #144.

[18]

Malloy, M. et alia (1990). “Exposure to a Chloride-Deficient Formula During Infancy: Outcome at Ages and 10 Years.” Pediatrics, 86, pp. 601-610.

[19]

Myers, W.O. et alia (1987). “Medical versus early surgical therapy in patients with triple-vessel disease and mild angina pectoris: A CASS registry of survival.” Annals of Thoracic Surgery, 44.

[20]

Reinisch, J., Sanders, S., Mortensen, E. and Rubin, D.B. (1995). “In utero exposure to Phenobarbital and intelligence deficits in adult men.” Journal of the American Medical Association, 274, pp. 1518-1525.

[21]

Rosenbaum, P. and Rubin, D.B. (1983a). “The central role of the propensity score in observational studies for causal effects.” Biometrika, 70, pp. 41-55.

[22]

Rosenbaum, P.R. and Rubin, D.B. (1983b). “Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome.” The Journal of the Royal Statistical Society, Series B, 45, pp. 212-218.

[23]

Rosenbaum, P.R. and Rubin, D.B. (1984). “Reducing bias in observational studies using subclassification on the propensity score.” Journal of the American Statistical Association, 79, pp. 516-524.

[24]

Rosenbaum, P.R. and Rubin, D.B. (1985a). “Constructing a control group using multivariate matched sampling methods that incorporate the propensity score.” The American Statistician, 39, pp. 33-38.

[25]

Rosenbaum, P.R. and Rubin, D.B. (1985b). “The bias due to incomplete matching.” Biometrics, 41, pp. 103-116.

[26]

Rubin, D.B. (1973). “The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies.” Biometrics, 29, 1, pp. 184-203.

[27]

Rubin, D.B. (1979). “Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies.” The Journal of the American Statistical Association, 74, 366, pp. 318-328.

[28]

Rubin, D.B. (1980). “Bias Reduction Using Mahalanobis’ Metric Matching.” Biometrics, 36, pp. 295-298.

[29]

Rubin, D.B. (1984). “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, 79, pp. 79-80.

[30]

Rubin, D.B. and Thomas, N. (1992a). “Affinely invariant matching methods with ellipsoidal distributions.” The Annals of Statistics, 20, pp. 1079-93.

[31]

Rubin, D.B. and Thomas, N. (1992b). “Characterizing the effect of matching using linear propensity score methods with normal covariates.” Biometrika, 79, pp. 797-809.

[32]

Rubin, D.B. and Thomas, N. (1996). “Matching using estimated propensity scores: relating theory to practice.” Biometrics, 52, pp. 249-264.

[33]

Rubin, D.B. and Thomas, N. (1997). “Combining propensity score matching with additional adjustments for prognostic covariates.” Submitted to The Journal of the American Statistical Association, Applications.

[34]

Stone, R.A. et alia (1995). “Propensity score adjustment for pretreatment differences between hospitalized and ambulatory patients with community-acquired pneumonia.” Medical Care, 33, pp. 56-66.

[35]

Willoughby, A. et alia (1990). “Population-based study of the development outcome of children exposed to chloride-deficient infant formula.” Pediatrics, 85, pp. 485-490.

[36]

General Accounting Office. (1994). “Breast conservation versus mastectomy: patient survival in day-to-day medical practice and randomized studies.” Washington D.C., U.S. General Accounting Office, Report #GAO-PEMD-95-9.