Health Services & Outcomes Research Methodology 2:169–188, 2001

Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation

DONALD B. RUBIN
Department of Statistics, Harvard University, 1 Oxford St., 6th Fl., Cambridge, MA 02138
E-mail: rubin@stat.harvard.edu

Received January 26, 2001; revised December 18, 2001; accepted December 19, 2001

Abstract. Propensity score methodology can be used to help design observational studies in a way analogous to the way randomized experiments are designed: without seeing any answers involving outcome variables. The typical models used to analyze observational data (e.g., least squares regressions, difference of difference methods) involve outcomes, and so cannot be used for design in this sense. Because the propensity score is a function only of covariates, not outcomes, repeated analyses attempting to balance covariate distributions across treatment groups do not bias estimates of the treatment effect on outcome variables. This theme will the primary focus of this article: how to use the techniques of matching, subclassification and/or weighting to help design observational studies. The article also proposes a new diagnostic table to aid in this endeavor, which is especially useful when there are many covariates under consideration. The conclusion of the initial design phase may be that the treatment and control groups are too far apart to produce reliable effect estimates without heroic modeling assumptions. In such cases, it may be wisest to abandon the intended observational study, and search for a more acceptable data set where such heroic modeling assumptions are not necessary. The ideas and techniques will be illustrated using the initial design of an observational study for use in the tobacco litigation based on the NMES data set.

Keywords: balance, matching, subclassification

1. Introduction—The Importance of Designing an Observational Study

In many contexts, the objectivity of a statistical design is of critical importance. For example, consider Phase III randomized trials, which are required before marketing pharmaceuticals. These studies are carefully designed and their structure is subject to the prior approval of the FDA (Food and Drug Administration); the initial design, which includes data collection and organization, and primary analyses are established before any outcome data from the trials are available. There still can be issues of analysis after the outcome data are available, such as conflicting results in subgroups or surprising patterns of side effects, but the design of these randomized trials attempts to minimize the impact of such complications. Arguably, the most important feature of experiments is that we must decide on the way data will be collected before observing the outcome data. If we could try hundreds of designs and for each see the resultant answer, we could capitalize on random variation in answers and choose the design that generated the answer we wanted! The lack of availability of outcome data when designing experiments is a tremendous stimulus for "honesty" in experiments and can be in well-designed observational studies as well.

Thus analogous care should be exercised in the design of important observational studies, such as ones attempting to assess the causal effects of cigarette smoking on health outcomes or of the causal effects of the conduct of the tobacco industry on medical expenditure outcomes. Of course, it is easier to control statistical bias in a randomized trial than in an observational study, where extraneous factors are not balanced by the randomization, but it is still possible to duplicate one crucial feature of a randomized experiment: one can design an observational study without access to the outcome data. In fact, one of the common maladies of observational studies for causal effects is that there is no real design, just repeated model-based analyses that produce answers. It is essentially impossible to be objective when a variety of analyses are being done, each producing an answer, either favorable, neutral, or unfavorable to the investigator's interests. But this situation is the norm in observational studies where typically the data base includes the outcome data as well as covariates and treatment indicators.

Study design, as conceptualized in this article, includes the organization of data, such as by matching and/or subclassification, that can be done without any examination of outcome data whatsoever. Thus, study design explicitly excludes doing any analyses that require access to any outcome data, e.g., least squares analyses of health-care-expenditure outcomes in the tobacco litigation example (Rubin, 2000a), or difference of differences estimates of schemes in a study of the effect of the minimum wage (Card and Kreuger, 1994), or the methods described in Heckman and Hotz (1989), or covariance adjustment for the "propensity score" (Rosenbaum and Rubin, 1983a). *Initial* study design, as used here, also excludes the specification of any analyses that use outcome data. *Entire* study design, in contrast, should include the specification of the analyses to be performed once outcomes are available. These specifications are important for objectivity but are different from the efforts discussed here, which can be completed and evaluated before outcomes are available.

A key tool for implementing initial observational-study design is the use of propensity score methods, whether for the purpose of constructing matched samples of treated-control pairs, subclasses of similar groups of treated and control units, or weighting adjustments. Propensity score methods use solely the covariates, not the outcomes, and propensity scores will be the topic of this article. The focus will be on bias reduction in design because in observational studies that typically is more important than variance reduction. Similar ideas based on the propensity score can be used to increase the precision of the design, as shown in the context of the design of a randomized experiment in education in Hill, Rubin, Thomas (1999).

After motivating this perspective on using propensity scores to help design observational studies, it will be illustrated using the primary data set appearing in much of the tobacco litigation, NMES (the National Medical Examination Survey, AHCPR, 1992), which is a large nationally representative data base of nearly 30,000 adults. NMES has many features that make it a natural candidate for estimating quantities that play a role in estimating the causal effects of smoking and the effect of the tobacco companies' alleged misconduct (see Rubin, 2000a,b, 2001). Central to the design of an objective observational study on these effects of smoking is the assembling of groups of never smokers and smokers with similar distributions of covariates, without allowing any examination of

health outcomes or health-care expenditure outcomes. In particular, because I have been involved in the tobacco litigation (Rubin, 2000a,b, 2001), all of the analyses and results of the design phase have been conducted without access to any outcome data.

2. Propensity Scores—No Outcome Variables in Sight

Propensity score methodology was introduced by Rosenbaum and Rubin (1983a). The propensity score is the probability of being treated ($W_i = 1$ vs. $W_i = 0$) given the observed value of a vector of observed covariates X_i, where i indexes the units in the study ($i = 1, \ldots, N$), and W_I is the indicator for received treatment. An extension also allows conditioning on the observed pattern of missingness in covariates (Rosenbaum and Rubin, 1984; D'Agostino and Rubin, 2000), but this extension was not used here. In a randomized experiment, the propensity scores are known, whereas in an observational study, they must be estimated from the data on W_i and X_i. No outcome data are required or desired; even if available in the data set, they should be set aside when designing the study, as was done here.

With no missing data in the covariates $\{X_i\}$, the propensity score e_i is defined as the probability that the ith unit is treated given that its vector of covariates is X_i,

$$e_{\mathrm{i}} \equiv e(X_i) \equiv Pr(W_i = 1/X_i), \tag{1}$$

a scalar summary of vector X_i. Thus, the mapping from X_i to e_i is generally a many-one function. The central result in Rosenbaum and Rubin (1983a) is that if a group of treated units and control units have the same value of the propensity score, e_i, then they have the same distribution of multivariate X_i, no matter what the dimension of X_i. Thus, having these groups of treated and control units with matching propensity scores automatically controls for all the observed covariates, at least in big samples: if there are differences in outcomes between the treated and control units, these differences cannot be due to these observed covariates. At the risk of being overly repetitive, if treatment and control groups have the same distribution of propensity scores, they have the same distribution of all observed covariates, just like in a randomized experiment.

Of course, propensity score technology can only attempt to achieve balance in observed covariates whereas randomization in experiments can stochastically balance all covariates, both observed and unobserved. Also, a randomized experiment offers the advantage that it provides an unambiguous definition of what constitutes a proper covariate—a variable that can be measured before the actual assignment of treatments. Variables measured after treatment assignment may "proxy for" proper covariates, but sometimes their status is rather ambiguous, and decisions whether to include them or not as covariates challenging.

With respect to NMES and the tobacco litigation, there is reason to include more than the usual list of proper covariates because the role of improper covariates in analyses involving outcome data will be different from the role of proper covariates; see Rubin (2000a, Section 4) for the logic and Section 5 of Rubin (2001) for a simple illustration. These two general topics, adjusting for unmeasured proper covariates and correctly adjusting for measured but improper covariates, are both beyond the scope of this paper because they require analyses involving outcome data.

3. History of Propensity Scores—Discriminant Matching in 1973 and Beyond; Observed and Unobserved Covariates

Given the simplicity of reducing a large space of covariates, X_i, to a one-dimensional summary, the probability of treatment assignment, e_i, it is interesting to review the development of the idea. The most direct path comes from matched sampling, which can also be accomplished without having outcomes available. Although there exists an extensive literature of its early use in applications (e.g., in sociology or education, Peters, 1941), statisticians seem to have mostly eschewed the topic for many years. An early formal statistical investigation in the context of treatment and control groups with different distributions was Rubin (1970), followed by Cochran and Rubin (1973), Rubin (1973a, 1976a,b), and Carpenter (1977).

A key bridge between matching and propensity scores is the use of a one-dimensional summary for matching, specifically "discriminant matching" (Cochran and Rubin, 1973; Rubin, 1976a,b, 1979, 1980). Under normality with a common covariance matrix in the treatment and control groups, the best linear discriminant is the linear version of the propensity score ($\text{logit}(e_i) = \log[e_i/(1-e_i)]$). Discriminant matching summarizes the covariates by the discriminant and uses this scalar variable for matching.

Of historical importance, the propensity score is not in the same class as any of the "confounder scores" of Miettinen (1976) nor the "selection models" of Heckman (1976). These attempts at dealing with the probability of treatment assignment either directly involve the observed dependent variable or indirectly through instrumental variables assumptions, and use this probability in a different way. In particular, they cannot be used to create balance in the observed covariates.

To be clear, I am most certainly not condemning parametric modeling based on relating the outcome variable to observed covariates or even hidden ones. Typically, analyses based on models for the outcome variables are important for obtaining the best final inferences. After the initial design stage, and after obtaining measured outcomes, models relating outcomes to observed covariates are very likely to reach improved (reduced bias, more precise) inferences, in analogy with covariance adjustments in randomized block experiments (Rubin, 1970, 1973b, 1979; Roseman, 1998; and Rubin and Thomas, 2000).

Differences due to unobserved covariates should be addressed after the balancing of observed covariates in the initial design stage, using models for sensitivity analyses (e.g., Rosenbaum and Rubin, 1983b) or models based on specific structural assumptions. When the list of observed covariates is rich, adjustment for hidden covariates may not be necessary. When the list of observed covariates is less rich, the conclusion after the initial design stage may be that the samples still require substantial adjustment for unobservables to reach scientifically plausible conclusions (e.g., Smith and Todd, 2002). This topic too is beyond the scope of this article.

4. Major Techniques for Initial Design—Matching, Subclassification, and Weighting

The initial design of an observational study attempts to assemble groups of treated and control units such that within each group the distribution of covariates is balanced. This

balance allows the initial attribution of any observed difference in outcomes to the effect of the treatment vs. control rather than differences in covariates. The techniques can be usefully, although imperfectly, classified into one of three types: matching, subclassification, and weighting. All rely on the propensity score as a fundamental component.

Propensity score matching refers to the pairing of treatment and control units with similar values of the propensity score, and possibly other covariates, and the discarding of all unmatched units (see Rosenbaum and Rubin (1983a, 1985); Rosenbaum (1989, 1991); Gu and Rosenbaum (1993); Rubin and Thomas (1992a,b, 1996, 2000); also see Cochran and Rubin (1973); and Rubin (1976a,b, 1979, 1980)) on discriminant matching. Typically, the matching finds for each treated unit one control unit, but sometimes more than one match is found (Rosenbaum, 1989, 1991; Rubin and Thomas, 2000). One-one Mahalanobis metric matching within propensity score calipers (Rosenbaum and Rubin, 1985) is a popular method illustrated here in Section 7.

Subclassification on the propensity score ranks all units by their propensity score and then uses boundaries to create subclasses with treated and control units with similar values of the propensity score—typically five or six subclasses are used, with approximately the same total number of units within each subclass (Rosenbaum and Rubin, 1983a, 1984; GAO, 1995). We illustrate the use of this technique on matched samples in Section 8.

Weighting methods use the inverse of the propensity score as a weight to apply to each treated unit and the inverse of one minus the propensity score as the weight to apply to each control unit (Czajka et al., 1992; Imbens, 2000). Such weighting methods are sometimes viewed as a final method of analysis based on ideas of Horvitz-Thompson (1952) estimation. However, this weighting can also be viewed as the limit of subclassification as the number of observations and subclasses tend to infinity. This technique is illustrated in Section 9.

The reasons for working this hard in the initial design stage of an observational study are twofold. First, as already mentioned, since no outcome data are available, none of these design efforts can inappropriately slant estimation of treatment effects on outcomes. Second, these design efforts, which result in more balanced distributions of covariates across treatment groups, make subsequent model-based adjustments (e.g., covariance adjustments, logistic regression relative risk adjustments, instrumental variables models) more reliable. This second point arises because such model adjustments can be extremely unreliable when the treatment groups are far apart on covariates. This unreliability of model-based adjustments in such cases, although critical, seems to have been frequently ignored in many applied fields, even though documented for at least a half-century as we shall see in the next section. Recent work confirming this unreliability in practice appears in Lalonde (1986); also see Dehija and Wahba (1999), where this unreliability of these models with Lalonde's data is rectified using propensity score methods.

5. Traditional Benchmarks for the Use of Regression Adjustment—When is it Reliable?

The statistical literature has, for many years, warned that regression analysis cannot reliably adjust for differences in observed covariates when there are substantial differences in the distribution of these covariates in the two groups.

For example, William G. Cochran wrote extensively on methods for the analysis of observational studies, as summarized in Rubin (1984). In Cochran (1957), he wrote:

> ... when the *x*-variables [i.e., covariates] show real differences among groups—the case in which adjustment is needed most—covariance adjustments [i.e., regression adjustments] involve a greater or less degree of extrapolation. To illustrate by an extreme case, suppose that we were adjusting for differences in parents' income in a comparison of private and public school children, and that the private-school incomes ranged from $10,000–$12,000, while the public-school incomes ranged from $4,000–$6,000. The covariance would adjust results so that they allegedly applied to a mean income of $8,000 in each group, although neither group has any observations in which incomes are at or near this level.

And later, in Cochran (1965), he wrote:

> If the original *x*-distributions diverge widely, none of the methods [e.g., regression adjustment] can be trusted to remove all, or nearly all, the bias. This discussion brings out the importance of finding comparison groups in which the initial differences among the distributions of the disturbing variables are small.

And in the same article:

> With several *x*-variables, the common practice is to compare the marginal distributions in the two groups for each *x*-variable separately. The above argument makes it clear, however, that if the form of the regression of *y* on the *x*'s is unknown, identity of the whole multi-variate distribution is required for freedom from bias.

In particular, there are three basic distributional conditions that in general practice must simultaneously obtain for regression adjustment (whether by ordinary linear regression, linear logistic regression, or linear-log regression) to be trustworthy. If any of these conditions is not satisfied, the differences between the distributions of covariates in the two groups must be regarded as substantial, and regression adjustment will be unreliable and cannot be trusted. These conditions are:

1. The difference in the means of the propensity scores in the two groups being compared must be small (e.g., the means must be less than half a standard deviation apart), unless the situation is benign in the sense that: (a) the distributions of the covariates in both groups are nearly symmetric, (b) the distributions of the covariates in both groups have nearly the same variances, and (c) the sample sizes are approximately the same.
2. The ratio of the variances of the propensity score in the two groups must be close to one (e.g., 1/2 or 2 are far too extreme).
3. The ratio of the variances of the residuals of the covariates after adjusting for the propensity score must be close to one (e.g., 1/2 or 2 are far too extreme); "residuals" precisely defined shortly.

These three guidelines also address regression adjustments on the logit or linear log scale because they too rely on linear additive effects in the covariates (for discussion of this point, see e.g., Anderson et al., 1980).

Specific tabulations and calculations relevant to these guidelines can be found, for example, in Cochran and Rubin (1973); Rubin (1973b); Rubin (1979); and Rubin and Thomas (2000). In particular, Cochran and Rubin (1973) state at page 426 that "linear regression on random samples gives wildly erratic results... , sometimes markedly overcorrecting [percentage bias reduction $\gg 100\%$], or even ... greatly increasing the original bias [percentage bias reduction $\ll 0\%$]." Tables in that article, summarized here in Table 1, imply that, when the ratio of the variances of any covariate is two or one-half, linear regression can grossly overcorrect for bias or grossly undercorrect for bias; B is the number of standard deviations between the means of the groups and R is the ratio of treatment variance to control variance. When there is a large initial bias, the remaining bias, even if most of it has been removed, can still be substantial (e.g., the $B = 1$, $R = 1$, marked nonlinearity condition of Table 1).

The reasons why conditions, 1,2,3, are relevant in the general situation with many covariates are the following. All mean bias is, by definition, along the propensity score. Thus, both the bias along the propensity score and its variance ratio are relevant to assessing the degree of extrapolation involved in regression adjustment, especially when the propensity score is transformed to approximate normality. Orthogonal to the propensity score, the variables have the same mean, and thus the variance condition is relevant to assessing the degree of extrapolation involved with regression adjustment. Operationally, regress each of the original variables on the estimated linear propensity score (i.e., project

Table 1. Percent reduction in bias using regression adjustment

	$B = 1/4$				$B = 1/2$				$B = 3/4$				$B = 1$			
R	Moderate		Marked		Moderate		Marked		Moderate		Marked		Moderate		Marked	
$y =$	exp $(x/2)$	exp $-(x/2)$	exp (x)	exp $(-x)$	exp $(x/2)$	exp $(-x/2)$	exp (x)	exp $(-x)$	exp $(x/2)$	exp $(-x/2)$	exp (x)	exp $(-x)$	exp $(x/2)$	exp $(-x/2)$	exp (x)	exp $(-x)$
2	62	298	48	−304	80	146	72	292	90	129	88	170	96	113	102	139
1	100	100	101	101	101	101	102	102	101	101	104	104	102	102	108	108
1/2	298	62	−304	48	146	80	292	72	123	90	170	88	113	96	139	102

x is normally distributed and outcome y is related to x by one of the moderately non-linear or markedly nonlinear relationships; B is the number of standard deviations between the means of the x distributions in the two groups, and R is the ratio of the variances of x in the two groups.

Note: If all bias were removed by regression adjustment, then all tabled values would be 100%. A negative number means that the adjustment, instead of removing bias, creates more bias in the same direction as the original bias; 0% means that the adjustment does not accomplish any bias reduction; a value larger than 200% indicates that the adjustment increases bias beyond the original amount but in the opposite direction.

Sources of values: Cochran and Rubin (1973). "Controlling Bias in Observational Studies: A Review". Sankhya, Series A, Vol. 35, 4, Tables 3.2.1, 3.2.2, and 3.2.3, pp. 427–429.

each original covariate on the linear combination of the covariates that defines the estimated propensity score), and then take the residual of this regression (i.e., the part of the original variable orthogonal to—uncorrelated with—the propensity score). The variance ratios of these residuals are what are referenced in the third condition.

These conditions implicitly assume normally distributed covariates, or at least variables whose distributions can be adequately summarized by means and variances. With markedly nonnormal covariates, analogous conditions for reliability of regression adjustment can be more complex. An obvious condition with nonnormally distributed propensity scores is the overlap of distributions of the propensity scores in the two groups. This point was made and illustrated a quarter-century ago in Rubin (1977) and is critical in applications such as the Lalonde (1986) data. When there are some treated subjects with propensity scores outside the range of the control subjects, no inference can be drawn concerning the effect of treatment exposure for these treated subjects from the data set without invoking heroic modeling assumptions based on extrapolation.

6. Comparing Smokers and Never-Smokers in NMES

We illustrate these ideas for observational study design with NMES, where the outcome variables, which we never see in initial design, are health-care expenditures of various types and the occurrence of various smoking-related diseases. The covariates in NMES are numerous and include age, sex, race, marital status, education, etc. Also available are detailed smoking information that allows us to classify people first, as "never smokers", "former smokers", and "current smokers", and then to classify all smokers further by length and density of smoking behaviors, and former smokers also by years since quitting.

The objects of inference are "smoking attributable fractions", "conduct attributable fractions", "relative expenditure risks", etc., all of which are based on a comparison of specific health-care expenditures (or disease rates) for a particular type of smoker with never smokers with the same values of covariates, as a function of dosage and covariates (see Rubin (2000a) for the definition of these quantities). The comparisons are typically based on linear regressions or part linear regression/logistic regression models (e.g., see Harrison, 1998; Zeger et al., 2000) with no initial design effort. Here we focus solely on initial design and try to create samples of smokers and never smokers in NMES with the same multivariate distribution of covariates. We evaluate the success of these efforts using the benchmarks of Section 5.

Males and females were considered separately. Two treatment groups were defined as current smokers—3,510 males and 3,434 females, and former smokers—3,384 males and 2,657 females. The control group for both current and former smokers consisted of never smokers—4,297 males and 7,691 females. All four propensity scores (male/female $\times$ current/former) were estimated by logistic regression using the 146 covariates defined by main effects, quadratic effects, and interaction effects; see Display 1 for a listing of these covariates. These analyses used the NMES sampling weights.

All propensity scores (which are estimated probabilities) were then transformed to the logit scale so that they were linear in the original covariates and their squares and products. This transformation was done for three reasons. First, relative to the raw propensity (probability) scale, the linear propensity (logit probability) is more relevant for assessing the efficacy of linear modeling adjustments (including those based on linear regression, logistic regression, and linear-log models). Second, the linear propensity scores tend to produce more benign distributions with more similar variances and more symmetry, because they are weighted averages of the original covariate values. And third, the linear propensity scores are more directly related to the benchmarks in the literature on adjustments for covariates based on linearity assumptions.

To assess the degree of overlap in distributions using the guidelines of Section 5, the following quantities were then calculated for all estimated linear propensity scores: (1) the standardized difference in the means of the propensity scores between smokers and never smokers, B; (2) the ratio of the variances of the propensity scores for smokers and never smokers, R; and (3) for each of the covariates, the ratio of the variance of the residuals orthogonal to the propensity scores for smokers to the variance of these residuals for never smokers (i.e., the residuals after adjusting for the propensity scores). The results of those calculations are found in Table 2 for the four comparisons (smokers vs. never smokers for male/female $\times$ current/former). The left two-thirds of Figure 1 displays the corresponding histograms of the propensity scores in the initial (unmatched) samples.

Comparing the results in Table 2 to the benchmarks in Section 5, it is clear that any linear (or part linear) regression model cannot be said to adjust reliably for these covariates, even if they were perfectly normally distributed. All values of B are greater than 1/2, and many of the values of R for the residuals of the covariates are outside the range (4/5, 5/4).

7. Mahalanobis Metric Matching of Smokers and Never-Smokers Within Propensity Score Calipers

For males and females, and current and former smokers, a one-one matched sample of never smokers was then selected. That is, for example, for the 3,510 male current smokers, 3,510 "matching" male never smokers were chosen from the pool of 4,297 male never smokers. The technique used was Mahalanobis metric matching (Rubin, 1976a) within propensity score calipers as defined in Rosenbaum and Rubin (1985). The caliper width was 0.2 of a linear propensity score standard deviation, and the variables included in the metric were: age, education, body mass index, and sampling weight. The use of the metric matching after propensity score matching is the observational study equivalent of blocking in a randomized experiment; see Rosenbaum and Rubin (1985) for an example and Rubin and Thomas (2000) for further explanation.

More specifically, for each "smoker", a "donor" pool of available matches was defined to include all never smokers who were within ± 0.2 standard deviations on the estimated propensity score; in our case, all such donor pools had never smokers. Starting with the hardest to match smoker (i.e., the one with the largest propensity score), individual

Display 1.

Variables Used in Propensity Model	Description
Seatbelt	5 levels of reported seat belt use
Arthritis	Whether reported suffering from arthritis
Census Division	9 census regions
Champ Insurance	Whether have military insurance
Diabetes	Doctor ever told having diabetes
Down time	6 levels of reported emotional down time
Dump time	6 levels of reported in the dumps time
Employment	Indicating employment status each quarter
English	English is a primary language
Retirement	Indicator for retirement status
Number of Friends	7 levels measuring the number of friends
Membership in Clubs	6 levels measuring memberships in clubs
Education	Completed years of education
HMO coverage	Indicating HMO coverage each quarter
High blood pressure	Doctor ever told having high blood pressure
Industry Code	14 Industry codes
Age	Age of the respondent
Labor Union	Indicator for a member of labor union
Log Height	Natural Logarithm of height
Log Weight	Natural Logarithm of weight
Marital Status	Marital status in each quarter
Medicaid	On medicaid (each quarter)
Medicare	On medicare (each quarter)
Occupation	Occupation code (13 levels)
Public Assistance	Other public assistance program (each quarter)
Friends over	Frequency of having friends over (7 levels)
Physical Activity	Indicator variable for physically active
Population density	3 levels
Poverty Status	6 levels
Pregnant 1987	Pregnancy status in 1987 (women)
Private Insurance	Other private insurance (each quarter)
Race	4 levels
Rated Health	5-point self rating of health status
Home ownership	Indicator for owning home
Rheumatism	Indicator for suffering from rheumatism
Share Life	Indicator variable for having somebody to share their life
Region	4 levels of region of the country
MSA	4 levels indicating types of metropolitan statistical area
Risk	General risk taking attitude (5 levels)
Uninsured	Indicator for lack insurance (each quarter)
Veteran	Indicator for veteran status
Incapler	Survey weight in NMES database
Agesq	Age*Age
Educat.sq	Education*Education
Age_wt	Age*Logwt
Age_educt	Age*Education
Age_ht	Age*Loght
Educat_wt	Education*Logwt

Display 1 (*continued*)

Variables Used in Propensity Model	Description
Educat_ht	Education*Loght
Loght_logwt	Loght*Logwt
Loghtsq	Loght*Loght
Logwtsq	Logwt*Logwt

matches were chosen. Specifically, the match chosen for the smoker was the never-smoker with the smallest Mahalanobis distance with respect to the few key continuous covariates mentioned in the previous paragraph. The smoker and his match were then set aside. This process continued until each smoker had been assigned a match. The unchosen never smokers were discarded (e.g., $4,297 - 3,510 = 787$ never smokers were discarded in the male current-smoker matching). Other versions of matching can certainly be defined, such as with-replacement of donors, or several donors to each treatment.

Table 3 displays the resultant matched samples with respect to the original estimated propensity score and the orthogonal components. Notice the dramatic reduction in bias along this estimated propensity score and the improved variance ratios in the orthogonal components. The right third of Figure 1 displays histograms in the matched control samples. Clearly, discarding the unmatched controls has created similar distributions of the propensity score in the matched samples. More refinement is still desirable, but the initial matching has done well at bringing the smokers and their control groups into balance.

The issue of what we would have done if some donor pools had been empty now arises. That is, what if there is a treated unit who has a propensity score not close to any control

Table 2. Estimated propensity scores on the logit scale for "smokers" versus never smokers in full NMES

			Percent of covariates with specified variance ratio orthogonal to the propensity score				
Treated Group	*B*	*R*	$\leq 1/2$	$>1/2$ and $\leq 4/5$	$>4/5$ and $\leq 5/4$	$>5/4$ and ≤ 2	>2
Male Current $N=3,510$	1.09	1.00	3	9	57	26	5
Male Former $N=3,384$	1.06	0.82	2	15	61	15	7
Female Current $N=3,434$	1.03	0.85	1	15	59	23	2
Female Former $N=2,657$	0.65	1.02	5	7	85	7	5

B = Bias, R = Ratio of "smoker" to never-smoker variances; also displayed is the distribution of the ratio of variances in the covariates orthogonal to the propensity score.

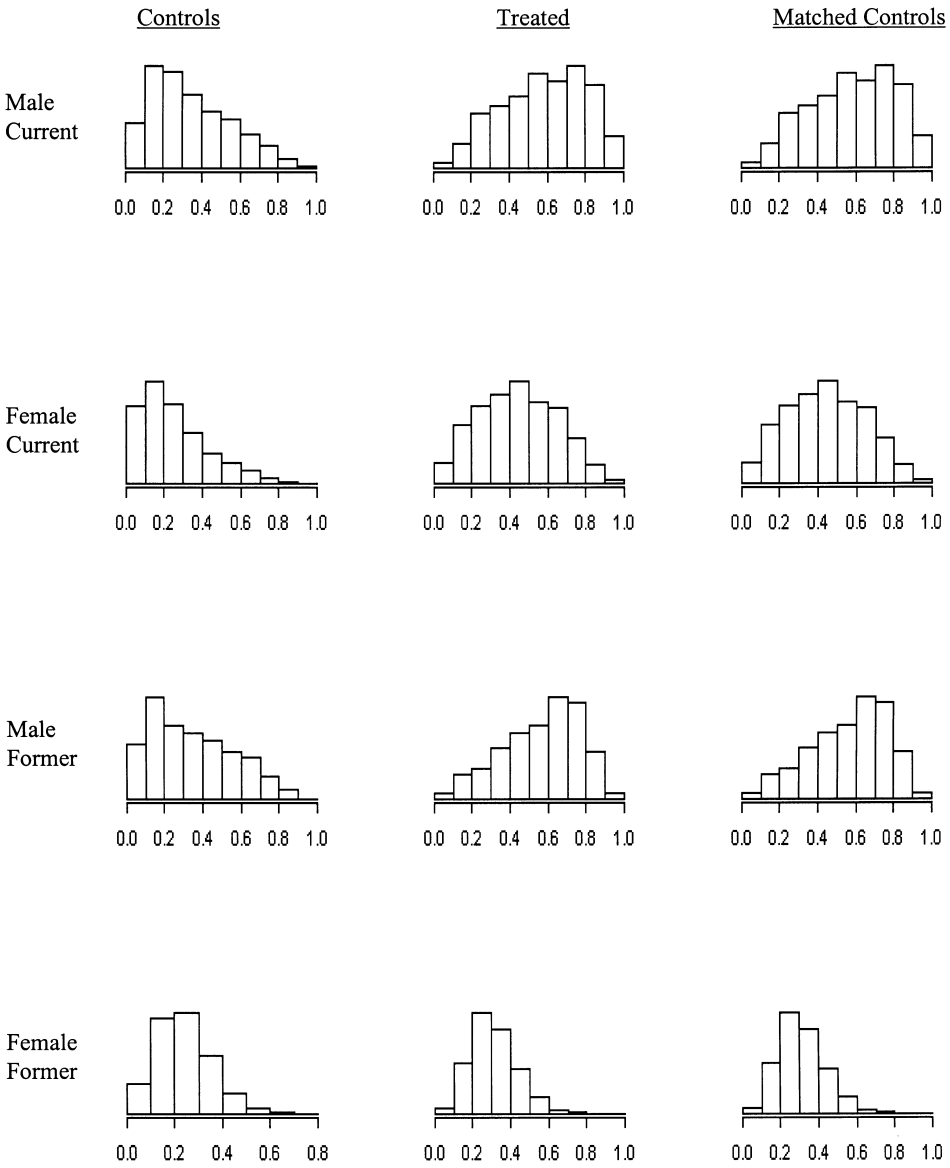

Figure 1. Histograms of propensity scores in: Full NMES controls, NMES treated, Matched NMES controls.

units' propensity scores? The correct answer is that inferences for the causal effects of treatment on such a unit cannot be drawn without making relatively heroic modeling assumptions involving extrapolations. Usually, such a unit should be explicitly excluded from the analysis. In the tobacco litigation, however, such a unit legally cannot be

Table 3. Estimated propensity scores from full NMES on the logit scale for "smokers" versus never smokers in matched NMES

			Percent of covariates with specified variance ratio orthogonal to the propensity score				
Treated Group	B	R	$\leq 1/2$	$>1/2$ and $\leq 4/5$	$>4/5$ and $\leq 5/4$	$>5/4$ and ≤ 2	>2
Male Current $N=3{,}510$	0.08	1.16	1	3	90	6	0
Male Former $N=3{,}384$	0.04	0.99	1	1	94	3	1
Female Current $N=3{,}434$	0.04	0.94	1	1	93	5	0
Female Former $N=2{,}657$	0.06	1.02	0	2	91	7	0

B = Bias, R = Ratio of "smoker" to never-smoker variances; also displayed is the distribution of the ratio of variances in the covariates orthogonal to the propensity score.

excluded, and so here would have been included with an attendant acknowledgement of the inferential difficulty of drawing causal inferences for such a unit.

8. Further Subclassification of Matched Samples of Smokers and Never-Smokers

Although Table 3 demonstrates dramatic improvement in balance relative to Table 2, more improvements are still desirable for the following reason. The estimated propensity score being summarized in Tables 2 and 3 (and Fig. 1) is the estimated propensity score in the full samples of Table 2, which does not equal the estimated propensity score in the matched samples themselves. That is, with respect to the linear combination of covariates that is the full sample estimated propensity score, there is now excellent balance in the matched samples. But we could now look at the linear combination of the covariates that is the estimated propensity score in the matched samples, and we might see some imbalance unless all individual covariates have extremely similar distributions in the matched samples; Table 3 reveals that there still are some differences. The difference between these two estimated propensity scores is due theoretically to small sample variation, just like random imbalance in covariates in a completely randomized experiment.

Table 4 gives the same information as Table 3 but now for the re-estimated propensity score, that is, estimated in the matched samples (again, using the NMES sampling weights). Although the differences between the treatment and control groups in Table 4 are minor relative to the differences displayed in Table 2 (e.g., all values of B are now less than 0.4), and arguably minor enough to satisfy the criteria set forth in Section 5, we seek even better balance within the matched samples on all covariates through the use of subclassification on the estimated propensity score within the matched samples.

Table 4. Estimated propensity scores on the logit scale for "smokers" versus never smokers in matched NMES

			Percent of covariates with specified variance ratio orthogonal to the propensity score				
Treated Group	*B*	*R*	≤1/2	>1/2 and ≤4/5	>4/5 and≤5/4	>5/4 and≤2	>2
Male Current $N=3{,}510$	0.39	1.33	0	4	88	8	0
Male Former $N=3{,}384$	0.32	1.33	0	1	95	3	1
Female Current $N=3{,}434$	0.35	1.18	1	1	92	6	0
Female Former $N=2{,}657$	0.31	1.09	0	2	91	7	0

B = Bias, R = Ratio of "smoker" to never-smoker variances; also displayed is the distribution of the ratio of variances in the covariates orthogonal to the propensity score.

Specifically, we rank the smokers and their matched sample of never smokers on the estimated propensity score in the matched samples. First, we create two equal-size (weighted) subclasses, low and high on the propensity score. The treated and control units with low propensity scores are to be compared against each other, and those with high propensity scores are to compared against each other, and the two estimates of treatment effect averaged. When doing this comparison for any covariate (including treating the propensity score as a covariate), the answer, ideally, should be zero since there is no effect of treatment on the covariates. Effectively, this subclassification creates new weights within each subclass: for the treated, the new weights are equal to the total number (weighted) of treated and controls in that subclass divided by the number (weighted) of treated in that subclass, and for the controls, the total number (weighted) of units in that subclass divided by the number (weighted) of controls in that subclass. Because the eventual goal is to draw inferences about the smokers if they had been never smokers, it can be argued that these high and low subclasses should be equal size within smokers. For the points in this article, however, this refinement makes little difference.

The same idea for weighting works no matter how many subclasses we have. Now the subclass specific weights can be attached to each unit, treated or control, and we can then do a new weighted propensity score analysis that reflects the extra balance created by the subclassification. The results of these weighted subclass propensity score analyses (for $k = 2$, 4, 6, 8, 10 subclasses) are displayed in Table 5A–5D. Table 4 serves as a comparison for these results because it reflects only one subclass—no subclassification adjustment.

The results in Table 5 show the reduction in initial bias that occurs with further subclassification. It is rather dramatic, both along the propensity score (re-estimated in matched samples) and orthogonal to it. The choice of the number of subclasses can be made by intense examination of specific covariates without fear of "biasing" any result involving the outcomes because there is no outcome variable being used. Once the outcome variables are available, modeling adjustments within the subclasses should be performed and then combined across subclasses. This approach allows for nonlinear

Table 5A. Propensity subclassification analyses on the logit scale for current vs. never-smoker males in matched NMES

			Percent of covariates with specified variance ratio orthogonal to the propensity score				
Number of Subclasses	B	R	$\leq 1/2$	$>1/2$ and $\leq 4/5$	$>4/5$ and $\leq 5/4$	$>5/4$ and ≤ 2	>2
1*	0.39	1.33	0	4	88	8	0
2	0.18	1.36	0	2	98	0	0
4	0.10	1.25	0	1	99	0	0
6	0.09	1.30	0	0	100	0	0
8	0.08	1.16	0	0	100	0	0
10	0.07	1.12	0	0	100	0	0

B = Bias, R = Ratio of "smoker" to never-smoker variances; also displayed is the distribution of the ratio of variances in the covariates orthogonal to the propensity score.
*Results from Table 4, row 1.

relationships between outcomes and covariates through a separate adjustment model within each subclass; see Benjamin (1999) for a specific example.

9. Other Approaches to Designing an Observational Study in NMES

We could let the number of subclasses continue to grow until each subclass had at most one treated or control unit—maximal subclassification (Rosenbaum, 1989). This, however, would preclude further standard modeling adjustments, although they could be

Table 5B. Propensity subclassification analyses on the logit scale for former vs. never-smoker males in matched NMES

			Percent of covariates with specified variance ratio orthogonal to the propensity score				
Numbers of Subclasses	B	R	$\leq 1/2$	$>1/2$ and $\leq 4/5$	$>4/5$ and $\leq 5/4$	$>5/4$ and ≤ 2	>2
1*	0.32	1.33	0	1	95	3	1
2	0.16	1.38	0	1	98	1	0
4	0.09	1.32	0	1	98	1	0
6	0.07	1.30	0	0	99	1	0
8	0.07	1.37	0	0	99	1	0
10	0.07	1.31	0	0	99	1	0

B = Bias, R = Ratio of "smoker" to never-smoker variances; also displayed is the distribution of the ratio of variances in the covariates orthogonal to the propensity score.
*Results from Table 4, row 2.

Table 5C. Propensity subclassification analyses on the logit scale for current vs. never-smoker females in matched NMES

			Percent of covariates with specified variance ratio orthogonal to the propensity score				
Number of Subclasses	B	R	$\leq 1/2$	$>1/2$ and $\leq 4/5$	$>4/5$ and $\leq 5/4$	$>5/4$ and ≤ 2	>2
1*	0.35	1.18	1	1	92	6	0
2	0.14	1.26	0	1	98	1	0
4	0.08	1.44	0	1	99	0	0
6	0.06	1.69	0	1	99	0	0
8	0.05	1.69	0	0	100	0	0
10	0.05	1.70	0	0	100	0	0

B = Bias, R = Ratio of "smoker" to never-smoker variances; also displayed is the distribution of the ratio of variances in the covariates orthogonal to the propensity score.
*Results from Table 4, row 3.

Table 5D. Propensity subclassification analyses on the logit scale for former vs. never-smoker females in matched NMES

			Percent of covariates with specified variance ratio orthogonal to the propensity score				
Number of Subclasses	B	R	$\leq 1/2$	$>1/2$ and $\leq 4/5$	$>4/5$ and $\leq 5/4$	$>5/4$ and ≤ 2	>2
1*	0.31	1.09	0	2	91	7	0
2	0.13	1.09	0	0	97	3	0
4	0.08	0.85	0	0	99	1	0
6	0.07	0.85	0	0	100	0	0
8	0.06	0.77	0	0	100	0	0
10	0.06	0.92	0	0	100	0	0

B = Bias, R = Ratio of "smoker" to never-smoker variances; also displayed is the distribution of the ratio of variances in the covariates orthogonal to the propensity score.
*Results from Table 4, row 4.

performed on "matched" pair differences formed within each maximal subclass—like a classical matched pair analysis. This is an unstudied approach to the best of my knowledge.

Another approach is to form weights directly from the estimated propensity score without subclassification. Thus, a treated unit's weight is the inverse of its propensity score (times its NMES weight) and a control person's weight is the inverse of one minus its propensity score (times its NMES weight). This process can generate unrealistically extreme weights when an estimated propensity score is near zero or one, something that is avoided in the subclassification approach.

Table 6 displays results from a weighted propensity score analysis using these inverse probabilities as multipliers of the NMES weights. The odd result for the variance ratio of the propensity score for the male former-smoker group is due to an extreme weight—a propensity score near zero. Another approach, unstudied to the best of my knowledge, would be to use subclasses for the highest and lowest estimated propensity score ranges (e.g., the upper and lower 2.5%) and the weights for the remaining interior 95% of the propensity scores.

A summary of all analyses can be displayed in one table for each treatment group comparison, as illustrated for male-current smokers in Table 7. This table documents the nice progression of increasing multivariate balance that can be obtained using the combined techniques of matching and finer and finer subclassification.

10. Discussion

The analyses we have done here are really just an indication of the types of analyses that can be done to help design an observational study using propensity scores, all without any fear of opportunistically biasing estimates of treatment effects. Care must be taken when estimating standard errors, as for the purpose of obtaining confidence intervals. Typically the standard errors and confidence intervals calculated assuming estimated propensity score are conservative (see Rubin and Thomas, 1992b). Refinements in our example could include defining finer categories of smoking (e.g., heavy or light) and selecting matches from the never smokers for each such treatment group.

The most principled strategy requires, in addition to the type of initial study design described here, the full specification of all analyses to be performed and assessment of their calibration. That is, ideally we may also wish to specify particular model-based analyses we intend to apply, again before observing any outcomes, to encourage complete objectivity in the analyses. For example, the sequence of "primary" and "secondary"

Table 6. Weighted propensity score analyses based on inverse probabilities for weights in matched NMES

			Percent of covariates with specified variance ratio orthogonal to the propensity score				
Treated Group	B	R	$\leq 1/2$	$>1/2$ and $\leq 4/5$	$>4/5$ and $\leq 5/4$	$>5/4$ and ≤ 2	>2
Male Current $N=3{,}510$	0.03	1.19	0	0	100	0	0
Male Former $N=3{,}384$	0.08	0.22	0	0	100	0	0
Female Current $N=3{,}434$	0.03	1.70	0	0	100	0	0
Female Former $N=2{,}657$	0.03	0.66	0	0	100	0	0

$B=$ Bias, $R=$ Ratio of "smoker" to never-smoker variances; also displayed is the distribution of the ratio of variances in the covariates orthogonal to the propensity score.

Table 7. Estimated propensity scores on the logit scale for male current smokers ($N = 3510$) versus male never smokers in NMES

			Percent of covariates with specified variance ratio orthogonal to the propensity score				
Analysis	B	R	$\leq 1/2$	$>1/2$ and $\leq 4/5$	$>4/5$ and $\leq 5/4$	$>5/4$ and ≤ 2	>2
Full $N = 4{,}297$	1.09	1.00	3	9	57	26	5
Matched $N = 3{,}510$	0.08	1.16	1	3	90	6	0
Matched							
$K = 1$	0.39	1.33	0	4	88	8	0
$K = 2$	0.18	1.36	0	2	98	0	0
$K = 4$	0.10	1.25	0	1	99	0	0
$K = 6$	0.09	1.30	0	0	100	0	0
$K = 8$	0.08	1.16	0	0	100	0	0
$K = 10$	0.07	1.12	0	0	100	0	0
$K = \infty$	0.03	1.19	0	0	100	0	0

B = Bias, R = Ratio of "smoker" to never-smoker variance; also displayed is the distribution of the ratio of variances in the covariates orthogonal to the propensity score.

analyses could be specified a priori, as in an FDA submitted protocol for a pharmaceutical company's randomized experiment. As mentioned earlier, if substantial balance in covariates has been obtained at the initial design stage, the exact form of the modeling adjustment is not critical because the similar treated and control covariate distributions implies only limited model-based sensitivity. Of course, repeated subgroup analyses, looking for "action" somewhere, are subject to more complicated interpretations, just as they are in a randomized experiment. Also, analyses involving adjustments for unobserved covariates are nearly always quite subjective, although recent progress on understanding adjustments for intermediate outcomes using principle strata (Frangakis and Rubin, 2002), shows promise for clarifying underlying assumptions.

It is hoped that this article will encourage the development of designed observational studies.

Acknowledgments

I thank T.E. Raghunathan for carrying out the computations described in this paper, and Guido Imbens and three reviewers for extremely helpful comments on an earlier draft.

References

AHCPR, "National medical expenditure survey, calendar year 1987," *Center for General Health Services Research, Agency for Health Care Policy and Research*, Public Health Service: Rockville, MD, 1992.

S. Anderson, A. Auquier, W. W. Hauck, D. Oakes, W. Vandaele and H. I. Weisberg, *Statistical methods for comparative studies*, John Wiley, New York, 1980.

D. J. Benjamin, "Does 401(k) eligibility increase net national savings?: reducing bias in the eligibility effect estimate," A. B. Honors Thesis in Economics, Harvard University, Cambridge, MA, 1999.

D. Card and A. Kreuger, "Minimum wages and employment: a case study of the fast food industry in New Jersey and Pennsylvania," *American Economic Review*, 84, pp. 772–793, 1994.

R. G. Carpenter, "Matching when covariates are normally distributed," *Biometrika*, 64, pp. 299–307, 1977.

W. G. Cochran, "Analysis of covariance: its nature and uses," *Biometrics*, 13, pp. 261–281, 1957.

W. G. Cochran, "The planning of observational studies of human populations," *Journal of the Royal Statistical Society-A*, 128, pp. 234–265, 1965.

W. G. Cochran and D. B. Rubin, "Controlling bias in observational studies: a review," *Sankhya-A*, 35, pp. 417–446, 1973.

J. C. Czajka, S. M. Hirabayashi, R. J. A. Little and D. B. Rubin, "Projecting from advance data using propensity modeling," *Journal of Business and Economics Statistics*, 10, pp. 117–131, 1992.

R. D'Agostino and D. B. Rubin, "Estimation and use of propensity scores with incomplete data," *Journal of the American Statistical Association*, 95, pp. 749–759, 2000.

R. Dehejia and S. Wahba, "Causal effects in non-experimental studies: re-evaluating the evaluation of training programs," *Journal of the American Statistical Association*, 94, pp. 1053–1062, 1999.

C. Frangakis and D. B. Rubin, "principal stratification in Casual Inference" Vol. 58(1), pp. 21–29, *Biometrics*, 2002.

GAO (U.S. General Accounting Office), "Breast conservation versus mastectomy: patient survival in day-to-day medical practice and randomized studies," *Report #GAO-PEMD-95-9*, U.S. General Accounting Office: Washington, D.C., 1995.

X. Gu and P. Rosenbaum, "Comparison of multivariate matching methods: structures, distances, and algorithms," *Journal of Computational and Graphical Statistics*, 2, pp. 405–420, 1993.

G. W. Harrison, 'Expert Report, April 27, 1998: "Health care expenditures attributable to smoking in Oklahoma,"' *The State of Oklahoma, ex rel., et al., Plaintiffs, vs. Reynolds Tobacco Co., et al., Defendants*, Case No. CJ-96-1499-L, District Court of Cleveland County, Oklahoma, 1998.

J. J. Heckman, "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models," *Annals of Economic and Social Measurement*, 5, pp. 475–492, 1976.

J. J. Heckman and V. J. Hotz, "Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training," *Journal of the American Statistical Association*, 84, pp. 862–880, 1989.

J. Hill, D. B. Rubin and N. Thomas, "The design of the New York school choice scholarship program evaluation," in *Research Designs: Inspired by the Work of Donald Campbell*, (L. Bickman, ed.), Sage Publications, Thousand Oaks, CA, 155–180, 1999.

D. G. Horvitz and D. J. Thompson, "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, 47, pp. 663–685, 1952.

G. W. Imbens, "The role of the propensity score in estimating dose-response functions," *Biometrika*, 87, pp. 706–710, 2000.

R. Lalonde, "Evaluating the econometric evaluations of training programs with experimental data," *American Economic Review*, 76, pp. 604–620, 1986.

O. Miettinen, "Stratification by a multivariate confounder score," *American Journal of Epidemiology*, 104, pp. 609–620, 1976.

C. C. Peters, "A method of matching groups with no loss of population," *Journal of Educational Research*, 34, pp. 606–612, 1941.

J. Reinisch, S. Sanders, E. Mortensen and D. B. Rubin, "In utero exposure to phenobarbital and intelligence deficits in adult men," *Journal of the American Medical Association*, 274, pp. 1518–1525, 1995.

L. Roseman, "Reducing bias in the estimate of the difference in survival in observational studies using subclassification on the propensity score," Ph.D. Thesis, Department of Statistics, Harvard University, Cambridge, MA, 1998.

P. R. Rosenbaum, "Optimal matching for observational studies," *Journal of the American Statistical Association*, 84, pp. 1024–1032, 1989.
P. R. Rosenbaum, "A characterization of optimal designs for observational studies," *Journal of the Royal Statistical Society-B*, 53, pp. 597–610, 1991.
P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, pp. 41–55, 1983a.
P. R. Rosenbaum and D. B. Rubin, "Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome," *Journal of the Royal Statistical Society-B*, 45, pp. 212–218.
P. R. Rosenbaum and D. B. Rubin, "Reducing bias in observational studies using subclassification on the propensity score," *Journal of the American Statistical Association*, 79, pp. 516–524, 1984.
P. R. Rosenbaum and D. B. Rubin, "Constructing a control group using multivariate matched sampling incorporating the propensity score," *The American Statistician*, 39, pp. 33–38, 1985.
D. B. Rubin, "The use of matched sampling and regression adjustment in observational studies," Ph.D. Thesis, Department of Statistics, Harvard University: Cambridge, MA, 1970.
D. B. Rubin, "Matching to remove bias in observational studies," *Biometrics*, 29, pp. 159–183, 1973a. Printer's correction note 30, p. 728.
D. B. Rubin, "The use of matched sampling and regression adjustment to remove bias in observational studies," *Biometrics*, 29, pp. 184–203, 1973b.
D. B. Rubin, "Multivariate matching methods that are equal percent bias reducing, I: some examples," *Biometrics*, 32, pp. 109–120, 1976a. Printer's correction note p. 955.
D. B. Rubin, "Multivariate matching methods that are equal percent bias reducing, II: maximums on bias reduction for fixed sample sizes," *Biometrics*, 32, pp. 121–132, 1976b. Printer's correction note p. 955.
D. B. Rubin, "Assignment to treatment group on the basis of a covariate," *Journal of Educational Statistics*, 2, pp. 1–26, 1977.
D. B. Rubin, "Using multivariate matched sampling and regression adjustment to control bias in observational studies," *Journal of the American Statistical Association*, 74, pp. 318–328, 1979.
D. B. Rubin, "Bias reduction using Mahalanobis' metric matching," *Biometrics*, 36, pp. 295–298, 1980. Printer's Correction p. 296 ((5,10) = 75%).
D. B. Rubin, "William, G. Cochran's contributions to the design, analysis, and evaluation of observational studies," in *W. G. Cochran's Impact on Statistics* (Rao and Sedransk, eds.), John Wiley, New York, pp. 37–69, 1984.
D. B. Rubin, "Statistical issues in the estimation of the causal effects of smoking due to the conduct of the tobacco industry," in *Statistical Science in the Courtroom* (J. Gastwirth, ed.), Springer-Verlag: New York, Chapter 16, pp. 321–351, 2000a.
D. B. Rubin, "Statistical assumptions in the estimation of the causal effects of smoking due to the conduct of the tobacco industry," in *Social Science Methodology in the New Millennium*. Proceedings of the Fifth International Conference on Logic and Methodology (J. Blasius, J. Hox, E. de Leeuw and P. Schmidt, eds.), October 6, 2000, Cologne, Germany, 1–22, 2000b.
D. B. Rubin, "Estimating the causal effects of smoking," *Statistics in Medicine*, 20, pp. 1395–1414, 2001.
D. B. Rubin and N. Thomas, "Affinely invariant matching methods with ellipsoidal distributions," *Annals of Statistics*, 20, pp. 1079–93, 1992a.
D. B. Rubin and N. Thomas, "Characterizing the effect of matching using linear propensity score methods with normal covariates," *Biometrika*, 79, pp. 797–809, 1992b.
D. B. Rubin and N. Thomas, "Matching using estimated propensity scores: relating theory to practice," *Biometrics*, 52, pp. 249–264, 1996.
D. B. Rubin and N. Thomas, "Combining propensity score matching with additional adjustments for prognostic covariates," *Journal of the American Statistical Association*, 95, pp. 573–585, 2000.
Smith and Todd, *Health Services and Outcomes Research Methodology*, 2002.
S. L. Zeger, T. Wyant, L. Miller and J. Samet, "Statistical testimony on damages in *Minnesota v. Tobacco Industry*," in *Statistical Science in the Courtroom* (J. Gastwirth, ed.), Springer-Verlag, New York, Chapter 15, 303–320, 2000.