Reducing Bias in Observational Studies Using Subclassification on the Propensity Score

Author(s): Paul R. Rosenbaum and Donald B. Rubin

Source: *Journal of the American Statistical Association*, Sep., 1984, Vol. 79, No. 387 (Sep., 1984), pp. 516-524

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/2288398

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms

Taylor & Francis, Ltd. and *American Statistical Association* are collaborating with JSTOR to digitize, preserve and extend access to *Journal of the American Statistical Association*

This content downloaded from 188.151.232.152 on Wed, 05 May 2021 12:42:11 76 12:34:56 UTC
All use subject to https://about.jstor.org/terms

Reducing Bias in Observational Studies Using Subclassification on the Propensity Score

PAUL R. ROSENBAUM and DONALD B. RUBIN*

The propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates. Previous theoretical arguments have shown that subclassification on the propensity score will balance all observed covariates. Subclassification on an estimated propensity score is illustrated, using observational data on treatments for coronary artery disease. Five subclasses defined by the estimated propensity score are constructed that balance 74 covariates, and thereby provide estimates of treatment effects using direct adjustment. These subclasses are applied within subpopulations, and model-based adjustments are then used to provide estimates of treatment effects within these subpopulations. Two appendixes address theoretical issues related to the application: the effectiveness of subclassification on the propensity score in removing bias, and balancing properties of propensity scores with incomplete data.

KEY WORDS: Bias reduction; Stratification; Logistic models; Log-linear models; Direct adjustment; Balancing scores.

1. INTRODUCTION: SUBCLASSIFICATION AND THE PROPENSITY SCORE

1.1 Adjustment by Subclassification in Observational Studies

In observational studies for causal effects, treatments are assigned to experimental units without the benefits of randomization. As a result, treatment groups may differ systematically with respect to relevant characteristics and, therefore, may not be directly comparable. One commonly used method of controlling for systematic differences involves grouping units into subclasses based on observed characteristics, and then directly comparing only treated and control units who fall in the same subclass. Obviously such a procedure can only control the bias due to imbalances in *observed* covariates.

Cochran (1968) presents an example in which the mortality rates of cigarette smokers, cigar/pipe smokers, and nonsmokers are compared after subclassification on the covariate age. The age-adjusted estimates of the average mortality for each type of smoking were found by direct adjustment—that is, by combining the subclass-specific mortality rates, using weights equal to the proportions of the population within the subclasses. Cochran (1968) shows that five subclasses are often sufficient to remove over 90% of the bias due to the subclassifying variable or covariate. However, as noted in Cochran (1965), as the number of covariates increases, the number of subclasses grows exponentially; so even with only two categories per covariate, there are 2^p subclasses for p covariates. If p is moderately large, some subclasses will contain no units, and many subclasses will contain either treated or control units but not both, making it impossible to form directly adjusted estimates for the entire population.

Fortunately, however, there exists a scalar function of the covariates, namely the propensity score, that summarizes the information required to balance the distribution of the covariates. Specifically, subclasses formed from the scalar propensity score will balance all p covariates. In fact, often five subclasses constructed from the propensity score will suffice to remove over 90% of the bias due to each of the covariates.

1.2 The Propensity Score in Observational Studies

Consider a study comparing two treatments, labeled 1 and 0, where z indicates the treatment assignment. The propensity score is the conditional probability that a unit with vector $\mathbf{x}$ of *observed* covariates will be assigned to treatment 1, $e(\mathbf{x}) = \Pr(z = 1 \mid \mathbf{x})$. Rosenbaum and Rubin (1983a, Theorem 1) show that subclassification on the population propensity score will balance $\mathbf{x}$, in the sense that within subclasses that are homogeneous in $e(\mathbf{x})$, the distribution of $\mathbf{x}$ is the same for treated and control units; formally, $\mathbf{x}$ and z are conditionally independent given $e = e(\mathbf{x})$,

$$\Pr(\mathbf{x}, z \mid e) = \Pr(\mathbf{x} \mid e) \Pr(z \mid e). \quad (1)$$

The proof is straightforward. Generally, $\Pr(\mathbf{x}, z \mid e) = \Pr(\mathbf{x} \mid e) \Pr(z \mid \mathbf{x}, e)$. But since e is a function of $\mathbf{x}$, $\Pr(z \mid \mathbf{x}, e) = \Pr(z \mid \mathbf{x})$. To prove (1), it is thus sufficient to show that $\Pr(z = 1 \mid \mathbf{x}) = \Pr(z = 1 \mid e)$. Now $\Pr(z = 1 \mid \mathbf{x}) = e$ by definition, and $\Pr(z = 1 \mid e) = E(z \mid e) = E\{E(z \mid \mathbf{x}) \mid e\} = E(e \mid e) = e$, proving (1).

* Paul R. Rosenbaum is Research Statistician, Research Statistics Group, Educational Testing Service, Princeton, NJ 08541. Donald B. Rubin is Professor, Departments of Statistics and Education, University of Chicago, Chicago, IL 60637. This research was sponsored in part by U.S. Army Contract DAAG29-80-C-0041, U.S. National Cancer Institute Grant P30-CA-14520 to the the Wisconsin Clinical Cancer Center, the Wisconsin Alumni Research Foundation, the Educational Testing Service, and the U.S. Health Resources Administration. The authors acknowledge Arthur Dempster for valuable conversations on the subject of this paper and Bruce Kaplan for assistance with Figures 1 and 2.

© Journal of the American Statistical Association
September 1984, Volume 79, Number 387
Applications Section

This content downloaded from
188.151.232.152 on Wed, 05 May 2021 12:42:11 UTC
All use subject to https://about.jstor.org/terms

Expression (1) suggests that to produce subclasses in which $\mathbf{x}$ has the same distribution for treated and control units, distinct subclasses should be created for each distinct value of the known propensity score. In common practice, $e(x)$ is not known, and it is not feasible to form subclasses that are exactly homogeneous in $e(x)$ and contain both a treated and control unit. In Section 2 we examine the balance obtained from subclassification on an estimated propensity score. Appendix A considers the consequences of coarse or inexact subclassification on $e(\mathbf{x})$. Of course, although we expect subclassification on an estimated $e(\mathbf{x})$ to produce balanced distributions of $\mathbf{x}$, it cannot, like randomization, balance unobserved covariates, except to the extent that they are correlated with $\mathbf{x}$.

Cochran and Rubin (1973) and Rubin (1970,1976a,b) proposed and studied discriminant matching as a method for controlling bias in observational studies. As noted by Rosenbaum and Rubin (1983a, Sec. 2.3 (i)), with multivariate normal $\mathbf{x}$ distributions having common covariance in both treatment groups, the propensity score is a monotone function of the discriminant score. Consequently, subclassification on the propensity score is a strict generalization of this work to cases with arbitrary distributions of $\mathbf{x}$.

Subclassification on the propensity score is not, however, the same as any of the several methods proposed later by Miettinen (1976); as Rosenbaum and Rubin (1983a, Sec. 3.3) state formally, the propensity score is not generally a "confounder" score. First, the propensity score depends only on the joint distribution of $\mathbf{x}$ and z, whereas a confounder score depends additionally on the conditional distribution of a discrete outcome variable given $\mathbf{x}$ and z, and is not defined for continuous outcome variables. Second, by Theorem 2 of Rosenbaum and Rubin (1983a), the propensity score is the coarsest function of $\mathbf{x}$ that has balancing property (1), so unless a confounder score is finer than the propensity score, it will not have this balancing property.

2. FITTING THE PROPENSITY SCORE AND ASSESSING THE BALANCE WITHIN SUBCLASSES

2.1 The First Fit and Subclassification

We illustrate subclassification based on the propensity score with observational data on two treatments for coronary artery disease: 590 patients with coronary artery bypass surgery ($z = 1$), and 925 patients with medical therapy ($z = 0$). The vector of covariates, $\mathbf{x}$, contains 74 hemodynamic, angiographic, laboratory, and exercise test results.* The propensity score was estimated using a logit model (Cox 1970) for z,

$$\log [e(\mathbf{x})/(1 - e(\mathbf{x})] = \alpha + \boldsymbol{\beta}^T \mathbf{f}(\mathbf{x}),$$

where α and $\boldsymbol{\beta}$ are parameters and $f(\cdot)$ is a specified function.

Not all of the 74 covariates and their interactions were included in the logit model for the 1,515 patients in the study. Main effects of variables were selected for inclusion in the first logit model using an inexpensive stepwise discriminant analysis. A second stepwise discriminant analysis added cross-products or interactions of those variables whose main effects were selected by the first stepwise procedure. Using these selected variables and interactions, the propensity score was then estimated by maximum likelihood logistic regression (using the SAS system). The result was the first logit model. (Alternatively, stepwise logit regression could have been used to select variables, e.g., Dixon et al. 1981.)

Based on Cochran's (1968) results and a new result in Appendix A of this article, we may expect approximately a 90% reduction in bias for each of the 74 variables when we subclassify at the quintiles of the distribution of the *population* propensity score. Consequently we subclassified at the quintiles of the distribution of the *estimated* propensity score based on this initial analysis, which we term the first model.

We now examine the balance achieved by this first subclassification. Each of the 74 covariates was subjected to a two-way (2 (treatments) $\times$ 5 (subclasses)) analysis of variance. Above the word *none*, Figures 1 and 2 display a five-number summary (i.e., minimum, lower quartile, median, upper quartile, maximum) of the 74 F ratios prior to subclassification, that is, the squares of the usual two-sample t statistics for comparing the medical and surgical group means for each covariate prior to subclassification. Above the word *one*, F ratios are displayed for the main effect of the treatment (Figure 1) and the treatment $\times$ subclass interaction (Figure 2) in the two-way analysis of variance. Although there has been a substantial reduction in most F ratios, several are still quite large, possibly indicating that the propensity score is poorly estimated by the first model. Indeed, as a consequence of Theorem 1 of Rosenbaum and Rubin (1983a), each such F test is an approximate test of the adequacy of the model for the propensity score; the test is only approximate primarily because the subclasses are not exactly homogeneous in the fitted propensity score.

2.2 Refinement of the Fitted Propensity Score and the Balance Obtained in the Final Subclassification

Figures 1 and 2 display summaries of F ratios from a sequence of models constructed by a gradual refinement of the first model. At each step, variables with large F ratios that had previously been excluded from the model were added. All logistic models were fitted by maximum likelihood. If a variable produced a large F ratio even

* The data analysis that follows is intended to illustrate statistical techniques, and does not by itself constitute a study of coronary bypass surgery. The literature on the efficency of coronary bypass surgery is quite extensive with many subtleties addressed and controversies exhibited. Furthermore, the data being used were considered "preliminary and unverified" for this application.

This content downloaded from 188.151.232.152 on Wed, 05 May 2021 12:42:11 UTC
All use subject to https://about.jstor.org/terms

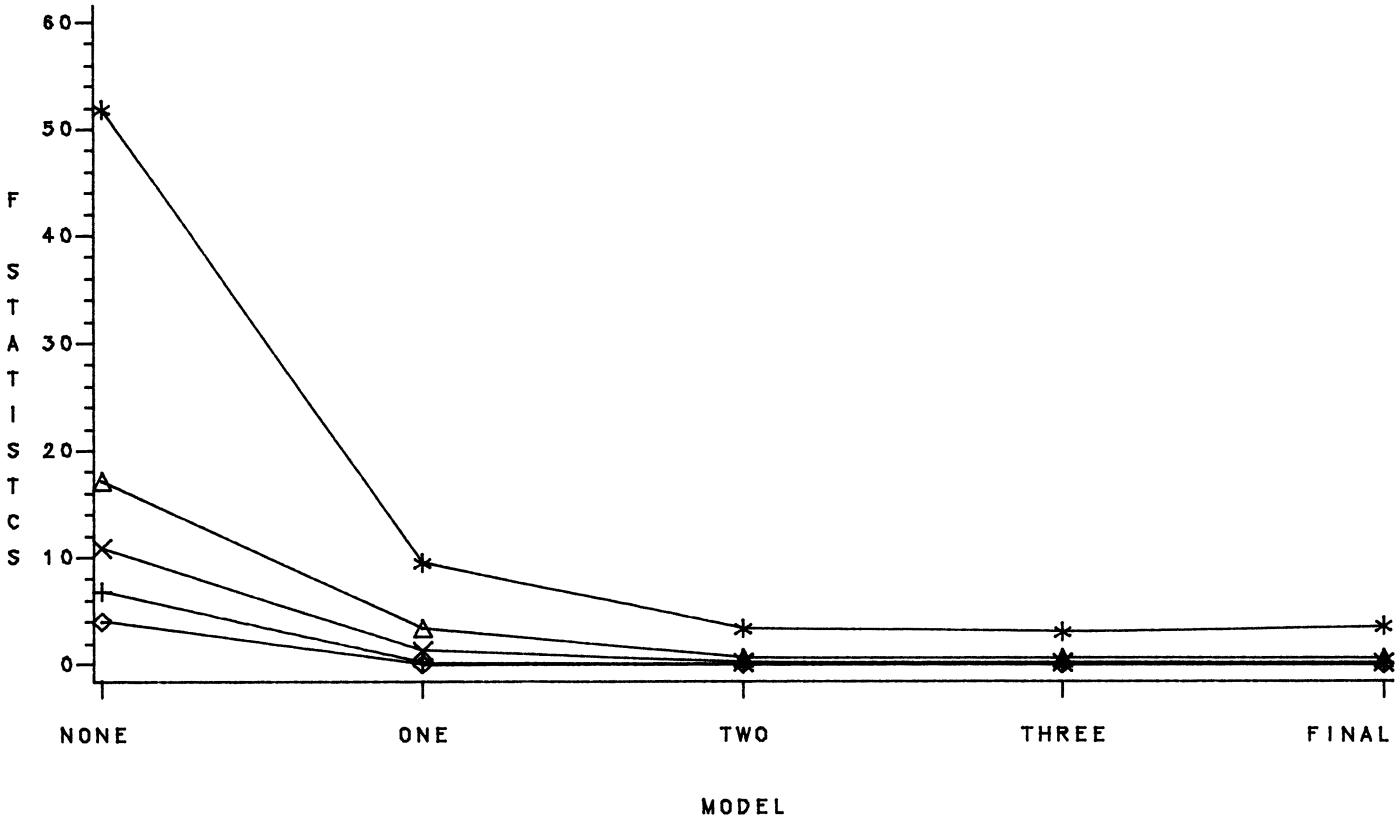

Figure 1. F Tests of Balance Before and After Subclassifications: Main Effects (5-point summary). (Minimum ◇; lower quartile +; median ×; upper quartile △; maximum ∗.)

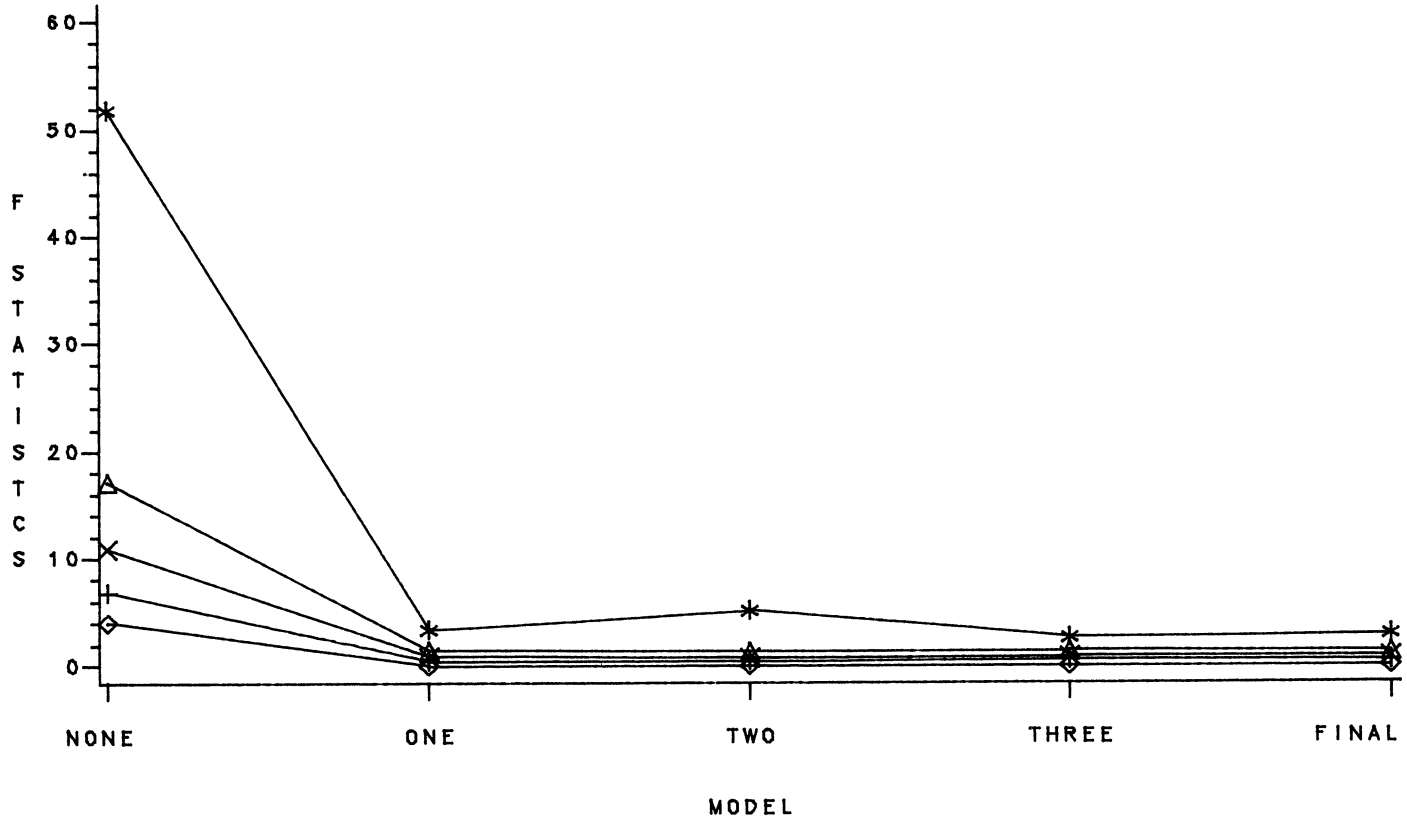

Figure 2. F Tests of Balance Before and After Subclassifications: Interactions (5-point summary). (Minimum ◇; lower quartile +; median ×; upper quartile △; maximum ∗.)

This content downloaded from
188.151.232.152 on Wed, 05 May 2021 12:42:11 UTC
All use subject to https://about.jstor.org/terms

after inclusion in the model, then the square of the variable and cross-products with other clinically important variables were tried. In the final model, β and $\mathbf{f}(\mathbf{x})$ in (2) were of dimension 45, including 7 interaction degrees of freedom and 1 quadratic term. There is considerably greater balance on the observed covariates $\mathbf{x}$ within these final subclasses than would have been expected from randomized assignment to treatment within subclasses.

Figures 3–5 display the balance within subclasses for three important covariates. Although the procedure used to form the subclasses may not be accessible to some nonstatisticians, the comparability of patients within subclasses can be examined with the simplest methods, such as the bar charts used here. For example, Figure 5 indicates some residual imbalance on the percentage of patients with poor left ventrical (LV) contraction, at least for patients in subclass 1—that is, in the subclass with the lowest estimated probabilities of surgery. This imbalance is less than would be expected from randomization within subclasses; the main-effect F ratio is .4 and the interaction F ratio is .9. Nonetheless, we would possibly want to adjust for this residual imbalance, perhaps using methods described in Section 3.3.

2.3 The Fitted Propensity Score: Overlap of Treated and Control Groups

Figure 6 contains boxplots (Tukey 1977) of the final fitted propensity scores. By construction, most surgical patients have higher propensity scores—that is, higher estimated probabilities of surgery—than most medical patients. There are a few surgical patients with higher estimated probabilities of surgery than any medical patient, indicating a combination of covariate values not appearing in the medical group. For almost every medical patient, however, there is a surgical patient who is comparable in the sense of having a similar estimated probability of surgery.

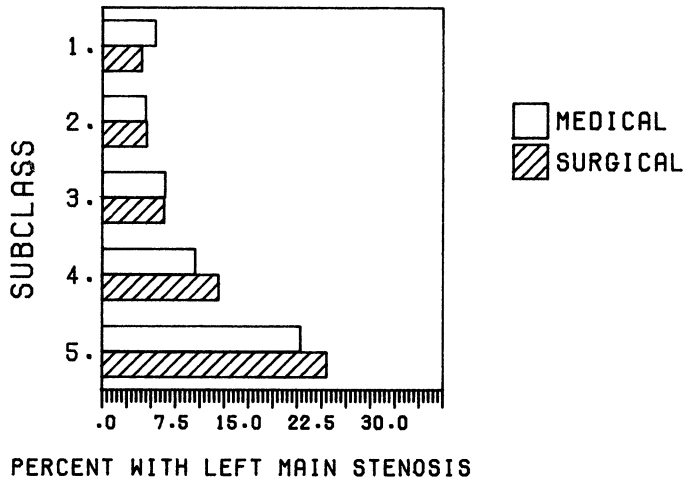

Figure 4. Balance Within Subclasses: Left Ventricular Contraction.

2.4 Incomplete Covariate Information

Five variables—four related to exercise tests and one quantitative measure of left ventrical function—were not measured during the early years of the study, so many patients are missing these covariate values. If the propensity score is defined as the conditional probability of assignment to treatment 1 given the observed covariate information *and* the pattern of missing data, then Appendix B shows that subclassification on the propensity score will balance both the observed data and the pattern of missing data. Essentially, we estimated the probabilities of surgical treatment separately for early and late patients, and then used these estimated probabilities as propensity scores. Subclassification on the corresponding population propensity scores can be expected to balance, within subclasses, each of the following: (a) the distribution of those covariates that are measured for both early and late patients, (b) the proportions of early and late patients, and (c) the distribution of all covariates for

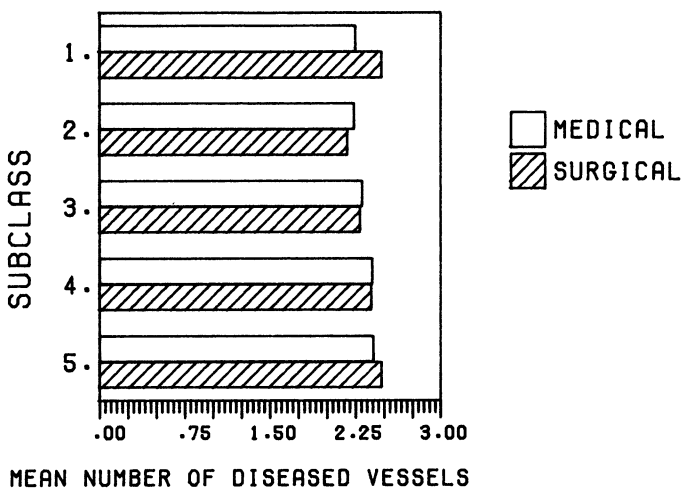

Figure 3. Balance Within Subclasses: Number of Diseased Vessels.

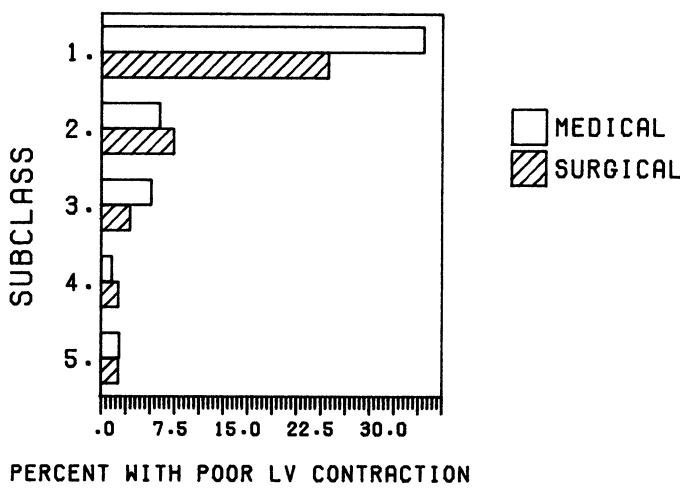

Figure 5. Balance Within Subclasses: Left Main Stenosis.

This content downloaded from
188.151.232.152 on Wed, 05 May 2021 12:42:11 UTC
All use subject to https://about.jstor.org/terms

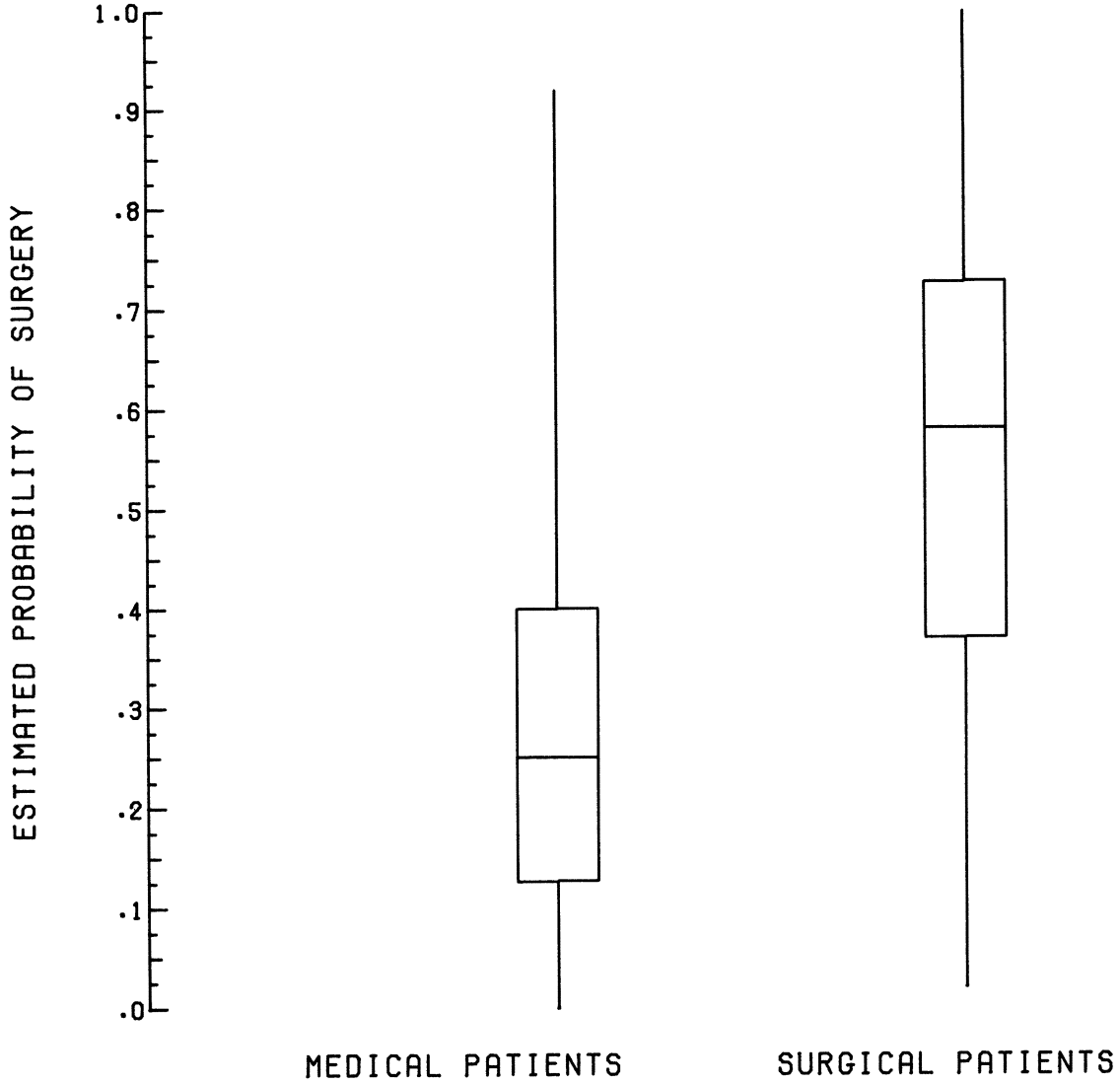

Figure 6. Boxplots of the Estimated Propensity Score.

the late patients. (For proof, see Corollary B.1 of Appendix B.) The observed values of these five covariates were indeed balanced by our procedure: the main-effect F ratios were 2.1, .1, .3, .2, and .0; the interaction F ratios were .4, 1.4, .1, .6, and .3.

3. ESTIMATING THE AVERAGE TREATMENT EFFECT

3.1 Survival; Functional Improvement; Placebo Effects

In this section, we show how balanced subclasses may be used to estimate the average effects of medicine and surgery on survival and functional improvement. Functional capacity is measured by the crude four-category (I = best, II, III, IV = worst) New York Heart Association classification, which measures a patient's ability to perform common tasks without pain. The current study is confined to patients in classes II, III, or IV at the time of cardiac catheterization, that is, patients who could improve. A patient is defined to have *uninterrupted improvement* to t years after cardiac catheterization if he:

1. is alive at t years and
2. has not had a myocardial infarction before t years and
3. is in class I or has improved by two classes (i.e., IV to II) at every follow-up before t years;

otherwise the patient does not have uninterrupted improvement to t years.

It should be noted that there is substantial evidence that patients suffering from coronary artery disease respond to placebos; for a review of this evidence, see Benson and McCallie (1979). Part or all of the difference in functional improvement may reflect differences in the placebo effects of the two treatments.

3.2 Subclass-Specific Estimates; Direct Adjustment

The estimated probabilities of survival and functional improvement at six months in each subclass for medicine and surgery are displayed in Table 1. (These estimates

This content downloaded from
188.151.232.152 on Wed, 05 May 2021 12:42:11 UTC
All use subject to https://about.jstor.org/terms

Table 1. Subclass Specific Results at Six Months

Subclass[a]	Treatment Group	No. of Patients	Survival to 6 Months: Estimate	Survival to 6 Months: Standard Error	Substantial Improvement at 6 Months: Estimate	Substantial Improvement at 6 Months: Standard Error
1	Medical	277	.892	(.019)	.351	(.030)
	Surgical	26	.846	(.071)	.538	(.098)
2	Medical	235	.953	(.014)	.402	(.032)
	Surgical	68	.926	(.032)	.705	(.056)
3	Medical	205	.922	(.019)	.351	(.034)
	Surgical	98	.898	(.031)	.699	(.047)
4	Medical	139	.941	(.020)	.303	(.042)
	Surgical	164	.933	(.020)	.706	(.036)
5	Medical	69	.924	(.033)	.390	(.063)
	Surgical	234	.914	(.018)	.696	(.030)
Directly Adjusted Across Subclasses	Medical	—	.926	(.022[b])	.359	(.042[b])
	Surgical	—	.903	(.039[b])	.669	(.059[b])

[a] Based on estimated propensity score.
[b] Standard errors for the adjusted proportions were calculated following Mosteller and Tukey (1977, Chap. 11c).

take censoring into account by using the Kaplan–Meier (1958) procedure.) In each subclass the proportion improved under surgery exceeds the proportion improved under medical therapy; the proportion surviving to six months is higher following medical treatment, although the standard errors are quite large.

Each subclass contains 303 patients. Therefore, for medical therapy and surgery, the directly adjusted proportions, with subclass total weights, are simply the averages of the five subclass-specific proportions. These adjusted proportions are displayed in Table 2 for $t = 6$ months, 1 year, and 3 years. Note that for $t = 6$ months, 1 year, and 3 years, the medical versus surgical differences in survival are small compared to their standard errors, but consistently higher probabilities of improvement are estimated for surgical treatment. As noted previously, improvement may be affected by differential placebo effects of surgery (Benson and McCallie 1979).

If the subclasses were perfectly homogeneous in the propensity score and the sample sizes were large, the distributions of $\mathbf{x}$ for surgical and medical patients would be identical within each subclass. Consequently, if this were the case, the difference between surgical and medical adjusted proportions would have no bias due to $\mathbf{x}$, or using terminology in Cochran (1968), subclassification on the propensity score (followed by direct adjustment) would remove all of the initial bias due to $\mathbf{x}$. Of course, initial bias due to unmeasured covariates will be removed only to the extent that they are correlated with $\mathbf{x}$.

Table 2. Directly Adjusted Probabilities of Survival and Uninterrupted Improvement (and Standard Errors)*

	6 Months: Pr	6 Months: SE	1 Year: Pr	1 Year: SE	3 Years: Pr	3 Years: SE
Survival						
Medical	.926	(.022)	.902	(.025)	.790	(.040)
Surgical	.903	(.039)	.891	(.040)	.846	(.049)
Uninterrupted Improvement						
Medical	.359	(.042)	.226	(.040)	.126	(.036)
Surgical	.669	(.059)	.452	(.060)	.298	(.057)

NOTE: Standard errors (SE) for the adjusted proportions were calculated following Mosteller and Tukey (1977, Chapter 11c).

In our example, however, the five subclasses are not perfectly homogeneous in $\mathbf{x}$. Results in Cochran (1968) show that for many examples of univariate x, five subclasses will remove approximately 90% of the initial bias in x; Cochran did not consider multivariate $\mathbf{x}$. Nevertheless, his results can be applied, using a theorem presented in Appendix A, to suggest that adjustment with five subclasses based on the propensity score will remove approximately 90% of the initial bias in each coordinate of multivariate $\mathbf{x}$.

3.3 Adjustment and Estimation Within Subpopulations Defined by $\mathbf{x}$

It is often of interest to estimate average treatment effects within subpopulations. This section shows how balanced subclassification may be combined with model-based adjustment to obtain estimates of the average effect of the treatment within subpopulations defined by $\mathbf{x}$. Specifically, we estimate the probabilities of uninterrupted improvement to six months for subpopulations of patients defined by the number of diseased vessels (N) and the New York Heart Association functional class at the time of cardiac catheterization (F). To avoid an excessive number of subpopulations, the small but clinically important subset of patients with significant left main stenosis has been excluded.

Patients were cross-classified according to the number of diseased vessels (N), initial functional class (F), treatment (Z), subclass based on the estimated propensity score (S), and condition at six months (I; improved = substantial treatment as defined in Sec. 3.1). A log-linear model, which fixed the IZN, IZF, ISN, SZ, SF, and FN margins, provided a good fit to this table (likelihood ratio $\chi^2 = 122.5$ on 120 degrees of freedom). (Here IZN denotes the marginal table formed by summing the entries in the table over initial functional class F and subclass S, leaving a three-way table.)

The directly adjusted estimates in Table 3 were calculated from the fitted counts, using the NFS marginal table for weights; in other words, within each subpopulation defined by the number of diseased vessels (N) and the initial functional class (F), estimates of the probabilities of improvement were adjusted using subclass (S) total weights. In all six subpopulations, the estimated probabilities of substantial improvement at six months are higher following surgery than following medical treatment (between 30% and 387% higher). The estimated probabilities differ least for one-vessel disease, functional class IV, and differ most for three-vessel disease, functional class III. The definition of substantial improvement has resulted in lower estimated probabilities of improve-

This content downloaded from 188.151.232.152 on Wed, 05 May 2021 12:42:11 UTC
All use subject to https://about.jstor.org/terms

Table 3. Directly Adjusted Estimated Probabilities of Substantial Improvement

No. of Diseased Vessels	Initial Functional Class		
	II	III	IV
1			
Medical Therapy	.469	.277	.487
Surgery	.708	.629	.635
2			
Medical Therapy	.404	.221	.413
Surgery	.780	.706	.714
3			
Medical Therapy	.248	.133	.278
Surgery	.709	.649	.657

ment for class III patients than for class II and IV patients. The estimated probabilities of improvement under surgery vary less than the estimated probabilities of improvement under medicine.

4. SENSITIVITY OF ESTIMATES TO THE ASSUMPTION OF STRONGLY IGNORABLE TREATMENT ASSIGNMENT

The estimates presented in Section 3 are approximately unbiased under the assumption that all variables related to *both* outcomes and treatment assignment are included in $\mathbf{x}$. This condition is called strongly ignorable treatment assignment by Rosenbaum and Rubin (1983a), and in fact Corollary 4.2 of that paper asserts that if (a) treatment assignment is strongly ignorable, (b) samples are large, and (c) subclasses are perfectly homogeneous in the population propensity score, then direct adjustment will produce unbiased estimates of the average treatment effect. In randomized experiments, $\mathbf{x}$ is constructed to include all covariates used to make treatment assignments (e.g., block indicators) with the consequence that treatment assignment is strongly ignorable.

Of course with most observational data, such as the data presented here, we cannot be sure that treatment assignment is strongly ignorable given the observed covariates because there may remain unmeasured covariates that affect both outcomes and treatment assignment. It is then prudent to investigate the sensitivity of estimates to this critical assumption.

Rosenbaum and Rubin (1983b) develop and apply to the current example a method for assessing the sensitivity of these estimates to a particular violation of strong ignorability. They assume that treatment assignment is not strongly ignorable given the observed covariates $\mathbf{x}$, but is strongly ignorable given $(\mathbf{x}, u)$, where u is an unobserved binary covariate. The estimate of the average treatment effect was recomputed under various assumptions about u. A related Bayesian approach was developed by Rubin (1978).

5. CONCLUSIONS: THE PROPENSITY SCORE AND MULTIVARIATE SUBCLASSIFICATION

With just five subclasses formed from an estimated scalar propensity score, we have substantially reduced the bias in 74 covariates simultaneously. Although the process of estimating the propensity score for use in balanced subclassification does require some care, the comparability of treated and control patients within each of the final subclasses can be verified using the simplest statistical methods, and therefore results based on balanced subclassification can be persuasive even to audiences with limited statistical training. The same subclasses can also be used to estimate treatment effects within subpopulations defined by the covariates $\mathbf{x}$. Moreover, balanced subclassification may be combined with model-based adjustments to provide improved estimates of treatment effects within subpopulations.

APPENDIX A: THE EFFECTIVENESS OF SUBCLASSIFICATION ON THE PROPENSITY SCORE IN REMOVING BIAS

Cochran (1968) studies the effectiveness of univariate subclassification in removing bias in observational studies. In this Appendix, we show how Cochran's results are related to subclassification on the propensity score.

Let $f = f(\mathbf{x})$ be any scalar valued function of $\mathbf{x}$. The initial bias in f is $B_I = E(f \mid z = 1) - E(f \mid z = 0)$. The (asymptotic) bias in f after subclassification on the propensity score and direct adjustment with subclass total weights is

$$B_S = \sum_{j=1}^{J} \{E(f \mid z = 1, e \in I_j) - E(f \mid z = 0, e \in I_j)\}\Pr(e \in I_j),$$

where there are J subclasses, and I_j is the fixed set of values of e that define jth subclass. The percent reduction in bias in f due to subclassification on the propensity scores is $100[1 - B_S/B_I]$.

Cochran's (1968) results do not directly apply to subclassification on the propensity score, since his work is concerned with the percent reduction in bias in f after subclassification on f, rather than the percent reduction in bias in f after subclassification on e. Nonetheless, as the following theorem shows, Cochran's results are applicable providing (a) the conditional expectation of f given e, that is $E(f \mid e) = \bar{f}$, is a monotone function of e, and (b) $\bar{f}$ has one of the distributions studied by Cochran. In particular, under these conditions, subclassification at the quintiles of the distribution of the propensity score, e, will produce approximately a 90% reduction in the bias of f. Note that in the following theorem, Cochran's (1968) results apply directly to the problem of determining the percent reduction in bias in $\bar{f}$ after subclassification on $\bar{f}$.

Theorem A.1. The percent reduction in the bias, $100(1 - B_S/B_I)$, in f following subclassification at specified quantiles of the distribution of the propensity score, e, equals the percent reduction in the bias in $\bar{f}$ after subclassification at the same quantiles of the distribution of $\bar{f}$, providing $\bar{f}$ is a strictly monotone function of e.

Proof. First, we show that within a subclass defined by $e \in S$, the bias in f equals the bias in $\bar{f}$; that is, we

This content downloaded from 188.151.232.152 on Wed, 05 May 2021 12:42:11 UTC
All use subject to https://about.jstor.org/terms

show that

$$E(f \mid e \in S, z = 1) - E(f \mid e \in S, z = 0) = E(\bar{f} \mid e \in S, z = 1) - E(\bar{f} \mid e \in S, z = 0). \quad \text{(A.1)}$$

To show this it is sufficient to observe that for $t = 0, 1$,

$$E(f \mid e \in S, z = t) = E\{E(f \mid e, e \in S, z = t) \mid e \in S, z = t\} = E\{E(f \mid e) \mid e \in S, z = t\} = E(\bar{f} \mid e \in S, z = t),$$

where the second equality follows from the fact that e is the propensity score (i.e., from Equation (1)).

From (A.1) with $S = [0, 1]$, it follows that the initial bias in f equals the initial bias in $\bar{f}$. To complete the proof, we need to show that the bias in f after subclassification on e equals the bias in $\bar{f}$ after subclassification on $\bar{f}$. Since by assumption $\bar{f}$ is a strictly monotone function of e, subclasses defined at specified quantiles of the distribution of e contain exactly the same units as subclasses defined at the same quantiles of the distribution of $\bar{f}$. It follows from this observation and (A.1) that the bias in f within each subclass defined by e equals the bias in $\bar{f}$ within each subclass defined by $\bar{f}$. Since (a) the initial biases in f and $\bar{f}$ are equal, (b) the subclasses formed from e contain the same units as the subclasses formed from $\bar{f}$, and (c) within each subclass, the bias in f equals the bias in $\bar{f}$, it follows that the percent reduction in bias in f after subclassification on e equals the percent reduction in bias in $\bar{f}$ after subclassification on $\bar{f}$.

APPENDIX B: BALANCING PROPERTIES OF THE PROPENSITY SCORE WITH INCOMPLETE DATA

In Section 2.4, we noted that several covariates were missing for a large number of patients. Let $\mathbf{x}^*$ be a p-coordinate vector, where the jth coordinate of $\mathbf{x}^*$ is a covariate value if the jth covariate was observed, and is an asterisk if the jth covariate is missing. (Formally, $\mathbf{x}^*$ is an element of $\{R, *\}^p$.) Then $e^* = \Pr(z = 1 \mid \mathbf{x}^*)$ is a generalized propensity score. The following theorem and corollary show that e^* has balancing properties that are similar to the balancing properties of the propensity score e. The notation $a \perp\!\!\!\perp b \mid c$ means that a is conditionally independent of b given c (see Dawid 1979).

Theorem B.1. $\mathbf{x}^* \perp\!\!\!\perp z \mid e^*$

Proof. The proof of Theorem B.1 is identical to the proof of Theorem 1 of Rosenbaum and Rubin (1983a), with $\mathbf{x}^*$ in place of $\mathbf{x}$ and e^* in place of e.

Theorem B.1 implies that subclassification on the generalized propensity score e^* balances the observed covariate information and the pattern of missing covariates. Note that Theorem B.1 does *not* generally imply that subclassification on e^* balances the unobserved coordinates of $\mathbf{x}$; that is, it does not generally imply

$$\mathbf{x} \perp\!\!\!\perp z \mid e^*.$$

The consequences of Theorem B.1 are clearest when there are only two patterns of missing data, with $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$, where $\mathbf{x}_1$ is always observed and $\mathbf{x}_2$ is sometimes missing. Let $c = 1$ when $\mathbf{x}_2$ is observed, and let $c = 0$ when $\mathbf{x}_2$ is missing. Then $e^* = \Pr(z = 1 \mid \mathbf{x}_1, \mathbf{x}_2, c = 1)$ for units with $\mathbf{x}_2$ observed, and $e^* = \Pr(z = 1 \mid \mathbf{x}_1, c = 0)$ for units with $\mathbf{x}_2$ missing. Subclasses of units may be formed using e^*, ignoring the pattern of missing data.

Corollary B.1. (a) For units with $\mathbf{x}_2$ missing, there is balance on $\mathbf{x}_1$ at each value of e^*; that is,

$$\mathbf{x}_1 \perp\!\!\!\perp z \mid e^*, c = 0.$$

(b) For units with $\mathbf{x}_2$ observed, there is balance on $(\mathbf{x}_1, \mathbf{x}_2)$ at each value of e^*; that is,

$$(\mathbf{x}_1, \mathbf{x}_2) \perp\!\!\!\perp z \mid e^*, c = 1.$$

(c) There is balance on $\mathbf{x}_1$ at each value of e^*; that is,

$$\mathbf{x}_1 \perp\!\!\!\perp z \mid e^*.$$

(d) The frequency of missing data is balanced at each value of e^*; that is,

$$c \perp\!\!\!\perp z \mid e^*.$$

Proof. Parts a and b follow immediately from Theorem 1 of Rosenbaum and Rubin (1983a), and Parts c and d follow immediately from Theorem B.1.

In practice, we may estimate e^* in several ways. In a large study with only a few patterns of missing data, we may use a separate logit model for each pattern of missing data. In general, however, there are 2^p potential patterns of missing data with p covariates. If the covariates are discrete, then we may estimate e^* by treating the * as an additional category for each of the p covariates, and we may apply standard methods for discrete cross-classifications (Bishop, Fienberg, and Holland 1975).

[*Received February 1983. Revised September 1983.*]

REFERENCES

BENSON, H., and McCALLIE, D. (1979), "Angina Pectoris and the Placebo Effect," *New England Journal of Medicine*, 330, 1424–1428.

BISHOP, Y., FIENBERG, S., and HOLLAND, P. (1975), *Discrete Multivariate Analysis*, Cambridge, Mass.: MIT Press.

COCHRAN, W.G. (1965), "The Planning of Observational Studies of Human Populations," *Journal of the Royal Statistical Society*, Ser. A, 128, 234–255.

——— (1968), "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies," *Biometrics*, 24, 205–213.

COCHRAN, W.G., and RUBIN, D.B. (1973), "Controlling Bias in Observational Studies: A Review," *Sankya*, Ser. A, 35, 417–446.

COX, D.R. (1970), *The Analysis of Binary Data*, London: Methuen.

DAWID, A.P. (1979), "Conditional Independence in Statistical Theory" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 41, 1–31.

DIXON, W., BROWN, M.B., ENGELMAN, L., FRANE, J.W., HILL, M.A., JENNRICH, R.I., and TOPOREK, J.D. (1981), *BMD-81: Biomedical Computer Programs*, Berkeley: University of California Press.

KAPLAN, E.L., and MEIER, P. (1958), "Nonparametric Estimation From Incomplete Observations," *Journal of the American Statistical Association*, 53, 457–481.

MIETTINEN, O. (1976), "Stratification by a Multivariate Confounder Score," *American Journal of Epidemiology*, 104, 609–620.

MOSTELLER, C.F., and TUKEY, J.W. (1977), *Data Analysis and Regression*, Reading, Mass.: Addison-Wesley.

This content downloaded from 188.151.232.152 on Wed, 05 May 2021 12:42:11 UTC
All use subject to https://about.jstor.org/terms

ROSENBAUM, P.R., and RUBIN, D.B. (1983a), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

——— (1983b), "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study With Binary Outcome," *Journal of the Royal Statistical Society*, Ser. B, 45, 212–218.

RUBIN, D.B. (1970), "The Use of Matched Sampling and Regression Adjustment in Observational Studies, unpublished Ph.D. thesis, Harvard University.

——— (1976a), "Matching Methods That Are Equal Percent Bias Reducing: Some Examples," *Biometrics*, 32, 109–120.

——— (1976b), "Multivariate Matching Methods That Are Equal Percent Bias Reducing: Maximums on Bias Reduction for Fixed Sample Sizes," *Biometrics*, 32, 121–132.

——— (1978), "Bayesian Inference for Casual Effects: The Role of Randomization," *Annals of Statistics*, 6, 34–58.

TUKEY, J.W. (1977), *Exploratory Data Analysis*, Reading, Mass.: Addison-Wesley.

This content downloaded from
188.151.232.152 on Wed, 05 May 2021 12:42:11 UTC
All use subject to https://about.jstor.org/terms