STATISTICS IN MEDICINE
Statist. Med. **17**, 2265–2281 (1998)

TUTORIAL IN BIOSTATISTICS

PROPENSITY SCORE METHODS FOR BIAS REDUCTION IN THE COMPARISON OF A TREATMENT TO A NON-RANDOMIZED CONTROL GROUP

RALPH B. D'AGOSTINO, Jr.*

Department of Public Health Sciences, Section on Biostatistics, Wake Forest University School of Medicine, Medical Center Boulevard, Winston-Salem, NC 27157-1063, U.S.A.

SUMMARY

In observational studies, investigators have no control over the treatment assignment. The treated and non-treated (that is, control) groups may have large differences on their observed covariates, and these differences can lead to biased estimates of treatment effects. Even traditional covariance analysis adjustments may be inadequate to eliminate this bias. The propensity score, defined as the conditional probability of being treated given the covariates, can be used to balance the covariates in the two groups, and therefore reduce this bias. In order to estimate the propensity score, one must model the distribution of the treatment indicator variable given the observed covariates. Once estimated the propensity score can be used to reduce bias through matching, stratification (subclassification), regression adjustment, or some combination of all three. In this tutorial we discuss the uses of propensity score methods for bias reduction, give references to the literature and illustrate the uses through applied examples. © 1998 John Wiley & Sons, Ltd.

INTRODUCTION

Observational studies occur frequently in medical research. In these studies, investigators have no control over the treatment assignment. Therefore, large differences on observed covariates in the two groups may exist, and these differences could lead to biased estimates of treatment effects. The propensity score for an individual, defined as the conditional probability of being treated given the individual's covariates, can be used to balance the covariates in the two groups, and thus reduce this bias. The propensity score has been used to reduce bias in observational studies in many fields. In particular, there are good recent examples in the literature where propensity scores were discussed in either applied statistical journals[1–7] or in medical journals.[8–21] Topics discussed in these articles come from a variety of fields including epidemiology, health services research, economics and social sciences.

* Correspondence to: Ralph B. D'Agostino, Jr, Department of Public Health Sciences, Section on Biostatistics, Wake Forest University School of Medicine, Medical Center Boulevard, Winston-Salem, NC 27157-1063, U.S.A. E-mail: rdagosti@rc.phs.bgsm.edu

CCC 0277–6715/98/192265–17$17.50

In a randomized experiment, the randomization of units (that is, subjects) to different treatments guarantees that on average there should be no systematic differences in observed or unobserved covariates (that is, bias) between units assigned to the different treatments. However, in a non-randomized observational study, investigators have no control over the treatment assignment, and therefore direct comparisons of outcomes from the treatment groups may be misleading. This difficulty may be partially avoided if information on measured covariates is incorporated into the study design (for example, through matched sampling) or into estimation of the treatment effect (for example, through stratification or covariance adjustment). Traditional methods of adjustment (matching, stratification and covariance adjustment) are often limited since they can only use a limited number of covariates for adjustment. However, propensity scores, which provide a scalar summary of the covariate information, do not have this limitation.

Formally, the propensity score[22] for an individual is the probability of being treated conditional on (or based only on) the individual's covariate values. Intuitively, the propensity score is a measure of the likelihood that a person would have been treated using only their covariate scores. Rosenbaum and Rubin[22] showed that the propensity score is a balancing score and can be used in observational studies to reduce bias through the adjustment methods mentioned above.

The three goals of this tutorial are: to present the formal definition of propensity scores with some theoretical findings; to illustrate common uses of the propensity score; and to present applied examples that illustrate applications of the propensity score. The Appendix includes SAS code used to perform some of the analyses presented. The tutorial will conclude with a discussion about areas of current and future research.

DEFINITION

With complete data, Rosenbaum and Rubin[22] introduced the propensity score for subject $i (i = 1, \ldots, N)$ as the conditional probability of assignment to a particular treatment ($Z_i = 1$) versus control ($Z_i = 0$) given a vector of observed covariates, x_i:

$$e(x_i) = \text{pr}(Z_i = 1 \mid X_i = x_i)$$

where it is assumed that, given the X's, the Z_i are independent:

$$\text{pr}(Z_1 = z_1, \ldots, Z = z_N \mid X_1 = x_1, \ldots, X_N = x_N) = \prod_{i=1}^{N} e(x_i)^{z_i}\{1 - e(x_i)\}^{1 - z_i}.$$

The propensity score is the 'coarsest function' of the covariates that is a balancing score, where a balancing score, $b(X)$, is defined as 'a function of the observed covariates X such that the conditional distribution of X given $b(X)$ is the same for treated ($Z = 1$) and control ($Z = 0$) units'.[22] For a specific value of the propensity score, the difference between the treatment and control means for all units with that value of the propensity score is an unbiased estimate of the average treatment effect at that propensity score, if the treatment assignment is strongly ignorable, given the covariates. Thus, matching, stratification, or regression (covariance) adjustment on the propensity score tends to produce unbiased estimates of the treatment effects when treatment assignment is strongly ignorable. Treatment assignment is considered strongly ignorable if the treatment assignment, Z, and the response, Y, are known to be conditionally independent given the covariates, X (that is, when $\mathbf{Y} \perp \mathbf{Z} \mid \mathbf{X}$).

When covariates contain no missing data, the propensity score can be estimated using discriminant analysis or logistic regression. Both of these techniques lead to estimates of probabilities of treatment assignment conditional on observed covariates. Formally, the observed covariates are assumed to have a multivariate normal distribution (conditional on Z) when discriminant analysis is used, whereas this assumption is not needed for logistic regression.

A question that may arise from investigators who have not used propensity scores before is: 'Why must we estimate the probability that a subject receives a certain treatment since we know for certain which treatment was given?' An answer to this question is that if we use the probability that a subject would have been treated (that is, the propensity score) to adjust our estimate of the treatment effect, we can create a 'quasi-randomized' experiment. That is, if we find two subjects, one in the treated group and one in the control, with the same propensity score, then we could imagine that these two subjects were 'randomly' assigned to each group in the sense of being equally likely to be treated or control. In a controlled experiment, the randomization, which assigns pairs of individuals to the treated and control groups, is better than this because it does not depend on the investigator conditioning on a particular set of covariates; rather it applies to any set of observed or unobserved covariates. Although the results of using the propensity scores are conditional only on the observed covariates, if one has the ability to measure many of the covariates that are believed to be related to the treatment assignment, then one can be fairly confident that approximately unbiased estimates for the treatment effect can be obtained.

USES OF PROPENSITY SCORES

Currently in observational studies, propensity scores are used primarily to reduce bias and increase precision. The three most common techniques that use the propensity score are matching, stratification (also called subclassification) and regression adjustment. Each of these techniques is a way to make an adjustment for covariates prior to (matching and stratification) or while (stratification and regression adjustment) calculating the treatment effect. With all three techniques, the propensity score is calculated the same way, but once it is estimated it is applied differently. Propensity scores are useful for these techniques because by definition the propensity score is the conditional probability of treatment given the observed covariates $e(X) = \mathrm{pr}(Z = 1 \mid X)$, which implies that Z and X are conditionally independent given $e(X)$. Thus, subjects in treatment and control groups with equal (or nearly equal) propensity scores will tend to have the same (or nearly the same) distributions on their background covariates.[6] Exact adjustments made using the propensity score will, on average, remove all of the bias in the background covariates. Therefore bias-removing adjustments can be made using the propensity scores rather than all of the background covariates individually.

MATCHING

Often investigators are confronted with studies where there are a limited number of treated patients and a larger (usually much larger) number of control patients. An example is a March of Dimes funded study examining the effects of post-term birth on neuropsychiatric, social and academic achievements among school-aged children (that is, 5–10 year old children). At the onset of the study, the investigators had a collection of over 9000 birth records (749 treated (post-term) babies and over 9000 potential control (term) babies), with prenatal and birth history

information. It was financially unfeasible for the investigators to collect outcome measurements on all potential subjects, so some form of sampling had to be performed.

Matching is a common technique used to select control subjects who are 'matched' with the treated subjects on background covariates that the investigator believes need to be controlled. Although the idea of finding matches seems straightforward, it is often difficult to find subjects who are similar (that is, can be matched) on all important covariates, even when there are only a few background covariates of interest. The investigators for the March of Dimes study had to confront this problem as they had more than ten variables on which they desired to match subjects.

Propensity score matching solves this problem by allowing an investigator to control for many background covariates simultaneously by matching on a single scalar variable. Prior to propensity score matching, a common matching technique was Mahalanobis metric matching using several background covariates.[23–27] Mahalanobis metric matching is employed by randomly ordering subjects, and then calculating the distance between the first treated subject and all controls, where the distance, $d(i, j)$, between a treated subject i and a control subject j is defined by the Mahalanobis distance:

$$d(i, j) = (u - v)^{\mathrm{T}} C^{-1} (u - v)$$

where u and v are values of the matching variables for treated subject i and control subject j, and C is the sample covariance matrix of the matching variables from the full set of control subjects. The control subject, j, with the minimum distance $d(i, j)$ is chosen as the match for treated subject i, and both subjects are removed from the pool. This process is repeated until matches are found for all treated subjects. One of the drawbacks of this technique is that it is difficult to find close matches when there are many covariates included in the model. As the number of dimensions on which the Mahalanobis distance is calculated increases, the average distance between observations increases as well. Propensity scores, on the other hand, can be calculated using many covariates, yet its score is still a scalar summary of the variables, and therefore matching is usually easy.

Rosenbaum and Rubin[6] outline three techniques for constructing a matched sample which use the propensity score: (i) nearest available matching on the estimated propensity score; (ii) Mahalanobis metric matching including the propensity score; and (iii) nearest available Mahalanobis metric matching within calipers defined by the propensity score.

Nearest available matching on the estimated propensity score. This method consists of randomly ordering the treated and control subjects, then selecting the first treated subject and finding the control subject with closest propensity score. Both subjects are then removed from consideration for matching and the next treated subject is selected. Rosenbaum and Rubin[6] suggest using the logit of the estimated propensity score to match (that is, $\hat{q}(X) = \log[(1 - \hat{e}(X))/\hat{e}(X)]$) because the distribution of $\hat{q}(X)$ is often approximately normal. Here, and in the following matching methods, recall the propensity score model may include many more covariates than employed in the Mahalanobis distance calculations.

Mahalanobis metric matching including the propensity score. This procedure is performed exactly as described above for Mahalanobis metric matching, with an additional covariate, the logit of the estimated propensity score ($\hat{q}(X)$) included with the other covariates in the calculation of the Mahalanobis distance. Rubin[24] showed that when the covariates have multivariate normal distributions and the treated and control groups have a common

covariance matrix, Mahalanobis metric matching is an equal per cent bias reducing (EPBR) technique, where the bias is the mean for the treated minus the mean for the control. In other words, the per cent bias reduced on all covariates is equal, and there are no covariates (or linear combinations of covariates) whose bias will increase due to matching.

Nearest available Mahalanobis metric matching within calipers defined by the propensity score. This method combines the previous two methods into one. The treated subjects are randomly ordered, and the first treated subject is selected. All control subjects within a preset amount (or caliper) of the treated subject's estimated propensity score ($\hat{e}(X)$) or estimated logit of the propensity score ($\hat{q}(X)$) are then selected, and Mahalanobis distances, based on a smaller number of covariates, are calculated between these subjects and the treated subject. The closest control subject and the treated subject are then removed from the pool, and the process is repeated. All remaining control subjects are available for the next matching with a treated subject. The size of the caliper is determined by the investigator. Cochran and Rubin[23] give advice on how large a caliper should be chosen based on the average of the variances of the covariates in the treated and control groups. Rosenbaum and Rubin[6] suggest that caliper size of a quarter of a standard deviation of the logit of the propensity score be used.

All three methods are useful techniques for reducing bias. Rosenbaum and Rubin[6] concluded that nearest available matching on the estimated propensity score was the easiest technique in terms of computational considerations. The second method, Mahalanobis metric matching including the propensity score, 'produced smaller standardized differences for individual variables but left a substantial difference along the propensity score'. They found the third method, nearest available Mahalanobis metric matching within calipers defined by the propensity score, to be the best technique among the three. It produced the best balance between the covariates in the treated and control groups, as well as the best balance of the covariates' squares and cross-products between the two groups. This third technique can be considered in the following way. By defining calipers based on the propensity score, the investigator is trying to create the 'quasi-randomized' experiment discussed above. Then, the use of Mahalanobis metric matching within calipers on a subset of the important covariates to choose subjects can be likened to blocking on important background variables in randomized controlled experiments. Rubin[28] also discusses this interpretation of Mahalanobis metric matching within calipers based on the propensity score.

APPLIED EXAMPLE: MARCH OF DIMES MATCHING

We now describe the steps taken in the March of Dimes study where the Mahalanobis metric matching within calipers based on the propensity score method was employed. First, we examined the distribution of 13 background covariates on which the investigators desired to match the term and post-term subjects. Table I contains descriptive statistics for these 13 covariates and the logit of the estimated propensity score, separately for the term and post-term groups. The first four columns contain the mean and standard deviation for each covariate and the logit of the estimated propensity score and the last two columns contain two statistics that are used to compare the groups. These statistics are the two-sample t-statistic and the standardized percentage difference. Based on these statistics, we see that there is moderate to large differences between the term and post-term groups on several covariates.

Table I. Group comparisons prior to matching

Variable	Post-term		Term		Comparisons	
	Mean	SD	Mean	SD	Two-sample t-statistic	Standardized difference in %†
	$N = 749$		$N = 9241$			
Sex of child	0·527	0·500	0·500	0·500	1·42	5·4
Parity	0·697	1·12	0·790	1·01	− 2·40*	− 8·7
Mother's age (years)	28·2	5·20	28·8	5·1	− 3·38**	− 12·7
Deliverey mode	1·28	0·455	1·23	0·431	2·75**	10·2
Hobel prenatal score	8·20	7·09	9·05	7·50	− 2·99**	− 11·6
Hobel Intrapartum score	10·09	8·62	7·41	7·46	9·37**	33·3
Child's age (months)	23·01	11·58	22·19	13·34	1·62	6·5
Child's birthweight (log grams)	8·20	0·143	8·11	0·149	15·58**	60·3
Mother's race (white = 1, non-white = 2)	1·19	0·488	1·22	0·539	− 1·77	− 6·7
Class (high = 3, low = 1)	1·628	0·778	1·650	0·759	− 0·79	− 3·0
Antepartum complications (yes/no)	0·729	0·445	0·699	0·459	1·71	6·5
Vaginal bleeding (yes/no)	0·128	0·335	0·124	0·329	0·36	1·4
Abnormal labour (yes/no)	0·453	0·498	0·354	0·478	5·42**	20·6
Logit of the propensity score	2·15	0·798	2·83	0·797	− 22·34**	− 60·0

* $0{\cdot}05 > p > 0{\cdot}01$

** $p < 0{\cdot}01$

† The standardized difference in % is the mean difference as a percentage of the average standard deviation: $100(\bar{x}_p - \bar{x}_t)/\sqrt{\{(s_p^2 + s_t^2)/2\}}$, where for each covariate $\bar{x}_p$ and $\bar{x}_t$ are the sample means in the post-term and term groups, respectively, and s_p^2 and s_t^2 are the corresponding sample variances

Two covariates with large initial differences between the term and post-term groups are the Hobel prenatal risk score and the Hobel intrapartum risk score.[29] Both of these have large two-sample t-statistics and standardized percentage differences. The Hobel risk scores were prenatal and intrapartum complications scales, respectively. These scores were calculated by determining whether or not certain risks were present for the women during the pregnancy and labour and then assigning specified weights to the presence of each risk factor. They were then calculated as the sum of these weights. For instance, for the Hobel prenatal risk score, a weight of 10 was given if the pregnant woman had a previous stillbirth, a weight of 5 was given if the pregnant woman was $\geqslant 35$ or $\leqslant 15$ years of age, and a weight of 5 was given if the pregnant woman had an abnormal glucose tolerance test. Thus, a woman who had these three risks, and no others, had a score of 20. In addition to covariates measured on the mother, there were two variables measured on the newborn infant, gender and date of birth (which is used to determine the subject's age at the time of the study). The goal of the matching was to reduce the differences between the term and post-term subjects on each of the covariates.

A propensity score model was estimated using discriminant analysis. In addition to the 13 covariates in Table I, 15 additional variables were included in this model. These 15 variables consisted of seven interactions terms and eight quadratic terms, which were calculated from the original 13 covariates, giving a total of 28 terms in the model. These interaction and quadratic terms were included based upon both statistical and scientific criteria (that is, certain interactions

Table II. Group comparisons after matching for variables used in Mahalanobis metric matching

Variable	Post-term		Term		Comparisons	
	Mean	SD	Mean	SD	Two-sample t-statistic*	Standardized difference in %
	$N = 749$		$N = 749$			
Sex†	0·527	0·500	0·527	0·500	0·00	0·0
Parity	0·697	1·12	0·629	0·997	1·24	6·4
Mother's age (years)	28·2	5·20	28·1	4·68	0·40	2·1
Delivery mode	1·28	0·455	1·28	0·452	0·01	0·0
Hobel prenatal score	8·20	7·09	7·63	6·53	1·62	8·4
Hobel intrapartum score	10·09	8·62	9·72	8·13	0·87	4·5
Child's age (months)	23·01	11·58	23·0	11·25	0·01	0·07
Child's birthweight (log grams)	8·20	0·143	8·20	0·129	0·82	4·4
Mother's race (white = 1, non-white = 2)	1·19	0·488	1·19	0·460	− 0·03	0·2
Class (high = 3, low = 1)	1·628	0·778	1·676	− 0·738	− 1·23	− 6·3
Antepartum complications (yes/no)	0·729	0·445	0·716	0·451	0·57	2·9
Vaginal bleeding (yes/no)	0·128	0·335	0·097	0·295	1·94	10
Abnormal labour (yes/no)	0·453	0·498	0·428	0·495	0·97	5
Logit of the propensity score	2·15	0·798	2·18	0·773	− 0·68	− 2·5

* p-values for all t-tests larger than 0·05
† Sex was exactly matched by design

and quadratic terms were included even if they were not found to be statistically significant). Recall, we do not use the propensity score model to make inferential statements concerning the term and post-term groups, rather, we use it to find propensity scores which are used to match term and post-term subjects and therefore create balance between the term and post-term groups. Thus, estimating a propensity score model with many terms does not create a problem.

Once the propensity scores were estimated, Mahalanobis metric matching was preformed as follows. The post-term subjects were randomly ordered, and the first post-term subject was selected. All term subjects within a determined caliper of the post-term subject's estimated logit of the propensity score ($\hat{q}(X)$) were then selected. The caliper chosen was equal to one-quarter of a standard deviation of the logit of the propensity score (0·20 from Table I). For example, if the first post-term subject had an estimated propensity score equal to 2·3 on the logit scale, then all term subjects with logit propensity scores between 2·1 and 2·5 would be selected as potential matches. The next step is to calculate Mahalanobis distances between these term subjects and the post-term subject. The closest term subject and the post-term subject are then removed from the pool, and the process is repeated. From the 28 terms in the propensity score model, the investigators chose eight covariates (the first eight variables of Table I) and the propensity score to be used in the Mahalanobis metric matching. These eight covariates were chosen because they were determined to be the most important variables for the final matching.

Propensity score matching using this method succeeded in removing most of the bias between the term and post-term groups. Table II contains descriptive statistics (means and standard deviations), two-sample t-statistics and standardized percentage differences for the original

Table III. Per cent reduction in bias for variables with initial standardized bias greater than 20 per cent

Variable	Initial bias	Bias after matching	Per cent reduction*
Hobel intrapartum risk score	2·687	0·377	85·9
Child's birthweight (log grams)	0·088	0·006	93·2
Abnormal labour (yes/no)	0·099	0·025	74·7
Logit of the propensity score	− 0·677	− 0·028	95·9

* Per cent reduction equals $100(1 - b_m/b_i)$ where b_m and b_i are post-term minus term differences in covariate means after matching and initially, respectively

post-term group ($N = 749$) and the matched term group ($N = 749$ matched term subjects out of the original $N = 9241$ term subjects). As can be seen, the matched sample had similar means for each of the 13 covariates included in the model. Table III shows the bias reduction for the four covariates with the largest initial bias: the Hobel Intrapartum Risk Score; child's birthweight; abnormal labour indicator, and the logit of the propensity score. As can be seen, each of these covariates had over 74 per cent bias reduction after matching.

The tables describe the results based on choosing the best available term match for each post-term subject. In addition to this, we generated a list of potential matches for each of the 749 post-term babies. For each post-term baby we provided the investigators with a list of 15 potential term matches. Based on this information, the investigators were able to identify matches for a subset of the post-term babies and then gather data on the matched pairs. Currently, the data on the matched pairs has been collected and analyses have begun to examine the hypothesis of interest, that is whether post-term birth is associated with neuropsychiatric, social and academic achievements among school-aged children (that is, 5–10 year old children).

In this example, the use of propensity scores proved useful. In particular, we were able to assess the fit of the propensity score model and compare the balance of background covariates prior to committing any resources (time or money) to collecting outcome data on the matched controls. Also, it is important to realize that, since these comparisons involve only covariates and not outcome variables, there is no chance of biasing results in favour of one treatment condition versus the other through the selection of matched controls.

STRATIFICATION

Stratification (sometimes referred to as subclassification) is also commonly used in observational studies to control for systematic differences between the control and treated groups. This technique consists of grouping subjects into strata determined by observed background characteristics. Once the strata are defined, treated and control subjects who are in the same stratum are compared directly. Many of the same problems occur in stratification as with matching when the number of covariates increases. Cochran[30] notes that as the number of covariates increases, the number of strata grows exponentially. For instance, if all covariates were dichotomous categorical variables, then there would be 2^k subclasses for k covariates. If k is large, then some strata might contain subjects from only the treated group, which would make it impossible to estimate a treatment effect in that stratum. Here again the propensity score is very useful. Because the propensity score is a scalar summary of all the observed background covariates, stratification on it alone can balance the distributions of the covariates in the treated and control groups without the exponential increase in number of strata.

Rosenbaum and Rubin[22] present theoretical results showing that perfect stratification based on the propensity score will produce strata where the average treatment effect within strata is an unbiased estimate of the true treatment effect. Again they assume that the treatment assignment is strongly ignorable. Rosenbaum and Rubin[5] state that Cochran's[31] result, which indicated that creating five strata removes 90 per cent of the bias due to the stratifying variable or covariate, holds for stratification based on the propensity score. They state that, in fact, stratification on the propensity score balances all k covariates that are used to estimate the propensity score, and often five strata based on the propensity score will remove over 90 per cent of the bias in each of these covariates.

The technique used for determining strata is straightforward. First, the propensity score is estimated by logistic regression or discriminant analysis. The investigator then must decide whether the stratum boundaries should be based on the values of the propensity score for both groups combined or in the treated or control group alone. Typically, in our work, we use the quintiles of the estimated propensity score from the combined group to determine the cut-offs for the different strata.

There are many examples in the recent literature of studies that have used propensity scores for stratification.[5,9,16,17,19–21] We now describe briefly some of these studies.

In Stone[20] investigators wished to compare outcomes on 747 patients with community-acquired pneumonia (CAP) who were either hospitalized ($n = 265$) or ambulatory ($n = 482$). Since patients were not randomized to be either hospitalized or ambulatory, propensity scores were estimated using classification tree techniques. Patients were then assigned to one of seven strata based on their estimated propensity score. The investigators found that there were imbalances between the two groups on 29 out of 44 baseline variables, and that after stratification on the propensity score only 13 of these remained significant at $p = 0{\cdot}05$. The investigators then estimated treatment effects using direct standardization methods of the stratum-specific means.

Fiebach *et al.*[9] used propensity scores to stratify patients who had received one of two possible treatments when they came to a hospital with uncomplicated chest pain. The two treatments were either admittance to a stepdown unit or admittance to a coronary care unit. Covariates used to estimate the propensity score included variables for the actual triage location and independent clinical predictors for an adverse event. These clinical predictors consisted of more than 50 clinical characteristics. A stepwise procedure was used to estimate the propensity score where covariates were entered into the model if they were significant at the 0·50 level in a stepwise discriminant analysis.

In Rosenbaum and Rubin[5] the authors wished to study the properties of the propensity score when used to stratify subjects in different treatment groups. In their example, the propensity score was the probability of receiving either coronary artery bypass surgery or medical therapy given 74 different covariates. These covariates consisted of haemodynamic, angiographic, laboratory and exercise test results. The investigators used a multi-stage procedure to find the best model for the propensity score. They found that using five strata based on the estimated propensity score was able to substantially reduce the bias in all 74 covariates simultaneously.

APPLIED EXAMPLE: FROM THE ACT STUDY

To illustrate further how to estimate and use the propensity score for stratification, we now present an applied example using data from the Active Management of Labor Trial (ACT).[32] The ACT trial is a randomized experiment to study the effects of active management of labour on the

Table IV. Comparison of covariates for subjects with and without epidural before and after propensity score stratification

	No epidural ($N = 775$) mean (sd)	Epidural ($N = 1003$) mean (sd)	F-statistics before stratification†	F-statistics after stratification‡
Pregnancy and labour characteristics				
Treated with active management of labour protocol	0·337 (0·47)	0·279 (0·45)	6·87**	0·20
Centimetres dilated at admission	3·95 (1·96)	2·79 (1·42)	208·01***	0·65
Artificially ruptured membranes (yes/no)	0·556 (0·50)	0·594 (0·49)	2·60	0·03
Gestational age (weeks)	39·9 (1·24)	40·2 (1·24)	22·28***	0·17
Infant birthweight (grams)	3374 (401)	3463 (416)	20·65***	0·20
Infant's gender (male = 1)	0·529 (0·50)	0·510 (0·50)	0·60	0·28
Initial rate of cervical dilation	58·3 (28·3)	42·9 (27·1)	135·20***	0·70
Maternal chronic hypertension	0·026 (0·16)	0·021 (0·14)	0·46	0·03
Maternal pregnancy induced hypertension (yes/no)	0·023 (0·15)	0·028 (0·16)	0·38	0·17
Maternal demographic/physical characteristics				
Maternal height (inches)	64·9 (2·8)	64·5 (2·6)	11·14**	0·10
Maternal pre-pregnant weight (pounds)	131·3 (21·6)	133·9 (22·9)	5·58*	0·07
Mother's age (years)	29·3 (5·1)	29·4 (5·3)	0·19	0·43
Insurance: private	0·857 (0·35)	0·882 (0·32)	2·55	2·75
public	0·101 (0·30)	0·084 (0·28)	1·51	0·54
Maternal race: white	0·677 (0·47)	0·735 (0·44)	7·01**	0·07
black	0·134 (0·34)	0·127 (0·33)	0·17	0·12
Hispanic	0·080 (0·27)	0·071 (0·26)	0·54	0·03

*$0{\cdot}05 > p > 0{\cdot}01$ **$0{\cdot}01 > p > 0{\cdot}001$ ***$0{\cdot}001 > p$

† F-statistic = square of two-sample t-statistic

‡ F-statistic for main effect of epidural use after adjusting for propensity score quintile

probability of having a Caesarean section. There were two components to this trial: a baseline component and a randomized component. In addition to the original study questions, the investigators were interested in determining whether the use of epidural anaesthesia was associated with Caesarean section in nulliparous women. To study this question, they wished to examine all eligible women from the baseline and randomized components of the trial. Propensity scores were used in these analyses since women were not randomly assigned to receive the treatment (an epidural). In this report we include 1778 women in the analyses, of these 1003 (56·4 per cent) had received an epidural. The investigators identified 15 variables (Table IV) which they felt may be imbalanced between the women who received an epidural and those who did not. Table IV shows the covariate imbalance before and after stratification based on the quintiles of the propensity score. Ten of the covariates were included in the final propensity score model used for stratification. The initial imbalance was measured by calculating F-statistics (squares of two-sample t-statistics) comparing the epidural and no epidural groups.

Propensity scores were estimated for each woman using logistic regression. Two of the covariates in this model, mother's insurance and mother's race, were transformed into dummy variables for the logistic regression model. Women were then separated into quintiles defined by

Table V. Comparison of quintile means for variable centimetres at admission

		N	Centimetres dilated at admission Mean (SD)
Overall	No epidural	775	3·95 (1·96)
	Epidural	1003	2·79 (1·42)
After stratification into quintiles based on propensity scores			
Quintile 1	No epidural	55	1·93 (1·02)
	Epidural	263	1·90 (1·03)
Quintile 2	No epidural	83	2·55 (1·00)
	Epidural	236	2·62 (1·11)
Quintile 3	No epidural	126	3·00 (1·28)
	Epidural	193	3·05 (1·19)
Quintile 4	No epidural	157	3·54 (1·32)
	Epidural	162	3·61 (1·42)
Quintile 5	No epidural	268	5·40 (1·78)
	Epidural	50	4·68 (1·19)

their propensity scores. We then compared the epidural/no epidural groups on their 15 covariates, after adjusting for their propensity quintile. This was done using a two-way analysis of variance model which included main effects for propensity score quintile (coded as a class variable with 4 degrees of freedom) and epidural use (coded as yes/no). We compared the F-statistic for epidural use after adjustment for propensity score quintile with the F-statistic for epidural use prior to adjustment for propensity score quintile to determine whether balance was achieved after stratification based on the propensity score. We also examined the two-way interaction of quintile and epidural use. We found that the eight covariates which were significantly different between the two groups prior to stratification, were all non-significantly different after adjustment for propensity score quintile (see Table IV). Among the interaction terms, only one was significant for the variable centimetres dilated at first exam ($F = 3{\cdot}16$, $p = 0{\cdot}013$). We further examined this variable by presenting the quintile means for the epidural and no epidural group (Table V and Figure 1). As can be seen by this table and figure, in the first four quintiles the mean centimetres at admission are very close, whereas in the 5th quintile the two groups are still somewhat separated. As can be seen in Figure 1, the means for the epidural/no epidural groups cross, and this explains the significant interaction between epidural use and propensity score quintile. Nevertheless, as can be seen both in the table and figure, the two groups are more similar within each propensity score quintile than they were before stratification. We have included SAS code which was used to estimate the propensity scores and perform the analyses presented here in the Appendix.

The investigators had several options for how they would estimate the effects of epidural use on the rate of Caesarean section use employing propensity scores. One method would be to estimate the treatment effects separately within each quintile defined by the propensity score and then combine the quintile estimates into an overall estimate of the treatment effect. An alternative method would be to perform a multiple logistic regression with use of Caesarean as the outcome and epidural use as the independent variable. In this model the propensity score could be

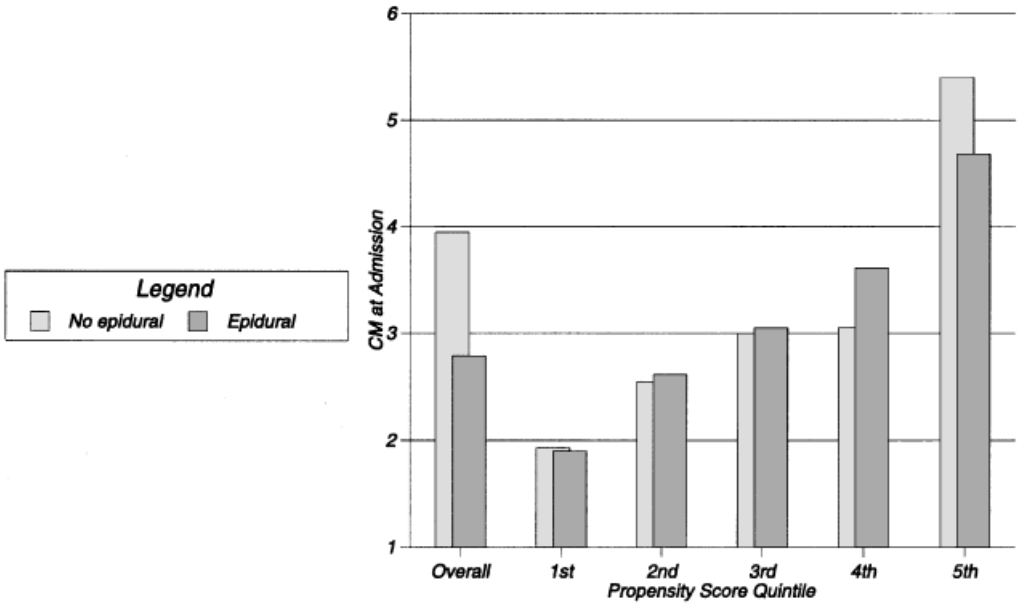

Figure 1. Comparison of quintile means for variable centimetres (cm) dilated at admission

included either as its raw score or the propensity score quintile itself could be included. In addition to the propensity score, a subset of the other covariates could be included (such as centimetres at admission). One advantage of this method is that the final model to estimate the treatment effect contains fewer covariates than if all the covariates in the propensity score model had been used. Therefore traditional diagnostics can be more easily employed to determine the fit of this final model. This was the method employed in this particular application, and the results indicated that after adjustment for propensity score (either the raw score or quintile) and a subset of important covariates, the rate of use of Caesarean sections in the epidural group was still significantly higher than the no epidural group (odds ratio = 3·7 with a 95 per cent confidence interval from 2·4 to 5·7).[21]

REGRESSION (COVARIANCE) ADJUSTMENT

Propensity scores can also be used in regression (covariance) adjustment. In regression adjustment, the treatment effect, τ, is estimated as

$$\hat{\tau} = (\bar{Y}_t - \bar{Y}_c) - \beta(\bar{X}_t - \bar{X}_c)$$

where the t and c indicate treatment and control groups. The effect of the background covariates is adjusted for by subtracting out the second term on the right hand side of the above equation, where β is an estimate of the regression of the responses for the treated and control groups on the background covariates. For the reasons stated above, the propensity score is a useful variable in regression adjustments, since one only has to find the regression of the responses on the propensity scores in the treated and control groups and use this to adjust the final estimate of the

treatment effect. Roseman[33] finds that if the response surfaces in the treatment and control groups are parallel and either linear or non-linear, then the regression adjustment using the propensity scores reduces the bias in the estimate of the treatment effect. In addition, if one stratifies and then uses regression adjustment within the strata, then this estimator of estimated treatment effect appears to be a more efficient estimator than one based on matching alone.

Another approach to regression adjustment is to use a large set of background covariates to estimate the propensity score and then take a subset of these covariates and the propensity score and use them in the regression adjustment. This is the analysis performed above by the ACT investigators. This method is also analogous to performing Mahalanobis metric matching within calipers on a subset of the important covariates including the propensity score as discussed above.

As with matching and subclassification, there are examples in the literature where regression adjustments were used with propensity scores.[10,14,15] Here is a brief description of two of them.

In Berk and Newton,[14] investigators wished to determine whether new spousal violence was influenced by whether men were either arrested or not arrested for wife-battery. Here the propensity score was the probability of being arrested and 14 covariates were used to estimate it. Some of these covariates included the suspect's age, whether the victim was injured, whether the suspect was drinking, and whether the victim had called the police. The investigators then 'regressed whether or not there was new spousal violence during the follow-up (by the same offender against the same victim) on the propensity scores, *separately for the arrested and not arrested group*'. They found that the arrested and not arrested groups had similar intercepts but different slopes. For the arrested group, the slope was near zero, but for the not arrested group the slope was very steep. This seemed to suggest that subjects who had a high propensity to be arrested, but were not arrested, were more likely to commit violence during the follow-up period.

In Muller *et al.*,[10] investigators studied the effects of digoxin in mortality rates in patients after myocardial infarction. Here the non-randomized treatment was digoxin. The investigators calculated an 'imbalance risk score', which appears to be a propensity score, based on 19 covariates. The covariates in this model included the subject's heart rate, age, and whether the subject had beta-blockers in the previous three weeks. Cox proportional-hazards regression was used to determine the association between digoxin therapy and survival 'taking into account the effects of baseline prognostic factors'. It appears that they 'took into account' the baseline differences by including a propensity score as a covariate in their model in order to adjust their final treatment effect.

One question which may arise when using regression adjustment with propensity scores is whether there is any gain in using the propensity score rather than performing a regression adjustment with all of the covariates used to estimate the propensity score included in the model. Rosenbaum and Rubin[25] showed that the 'point estimate of the treatment effect from an analysis of covariance adjustment for multivariate X is … equal to the estimate obtained from a univariate covariance adjustment for the sample linear discriminant based on X, whenever the same sample covariance matrix is used for both the covariance adjustment and the discriminant analysis'. Thus, the results from both methods should lead to the same conclusions. However, one advantage to performing the two-step procedure is that one can fit a very complicated propensity score model with interactions and higher order terms first. Since the goal of this propensity score model is to obtain the best estimated probability of treatment assignment, one is not concerned with over-parameterizing this model. Then when the model for estimating the treatment effect is estimated the investigator can include only a subset of the most important variables and the

propensity score in the model. This smaller model may allow the investigator to perform diagnostic checks on the fit of the model more reliably than if there were many covariates included in the model.

In general, covariance adjustment should be performed with caution. Rubin[25] showed that covariance adjustment may in fact increase the expected squared bias if the covariance matrices in the treated and untreated groups are unequal (that is, if the discriminant is *not* a monotone function of the propensity score). Another difficulty arises when the variance in the treated and untreated groups are very different (that is, the untreated group variance is much larger than the treated groups variance). Under these circumstances, one may consider using propensity score methods for matching or subclassification, rather than using covariance adjustment.

DISCUSSION AND CURRENT RESEARCH

In all the examples stated above, except for Rosenbaum and Rubin,[5] there was no mention of how missing values on covariates were handled when estimating propensity scores. This is an important issue in most real data applications. For instance, in the ACT example presented above, over 100 subjects would have been excluded from the final analyses based on the fact that they were missing covariates included in the propensity score models. The March of Dimes Study also had covariates with missing data into the models. Currently, methods are being developed to handle this problem.[34,35] which allow for different missing-data mechanisms and use the EM[36] or ECM[37] algorithms to estimate propensity scores.

A second area of current research is using propensity scores to estimate treatment effects in clinical trials where subjects drop out prior to the trials completion.[38] Here, propensity scores are estimated as the probability that an individual will complete the trial conditional on their baseline and early outcomes. This work also uses the EM and ECM algorithms to estimate propensity scores with missing data.

Propensity scores are being widely used in statistical analyses, particularly in the area of applied medicine. Their use should only increase as the cost for randomized clinical trials rises and more investigators turn to observational studies as a means of performing less expensive research. The propensity score methodology appears to produce the greatest benefits when it can be incorporated into the design stages of studies (through matching or stratification). These benefits include providing more precise estimates of the true treatment effects as well as saving time and money. This saving results from being able to avoid recruitment of subjects who may not be appropriate for particular studies. Finally, it is important to note that we are not advocating the use of only propensity scores in analyses of observational studies, rather we are encouraging the use of propensity scores in addition to traditional methods of analysis. The propensity score should be thought of as an additional tool available to the investigators as they try to estimate the effects of treatments in studies.

APPENDIX

The following is a series of SAS code that will estimate propensity scores using the logistic regression procedure in SAS. We first perform a series of *t*-tests to determine the initial level of bias between the two groups. Then we calculate propensity scores using stepwise logistic regression. Next, propensity score quintiles are created and then we examine whether the groups are balanced after adjustment for the propensity score quintiles.


```
* This will perform a series of ttests to determine what the initial difference between the
* treated and control groups are. Here the variable epidural is the treatment indicator in this
* model.

proc ttest data = matchset; class epidural;
var amladmit cm l arom gestage birthwt gender rate chyper phyper
    height weight momage insprvt momw momb;

* This performs a stepwise logistic regression to estimate propensity scores for each subject.
* The variable pr is the propensity score The variable epidural is the treatment indicator in
* this model.

proc logistic data = matchset nosimple;
model epidural = amladmit cm l arom gestage birthwt gender rate chyper phyper
    height weight momage insprvt momw momb/selection = stepwise;
output out = preds pred = pr;

* This takes the propensity score and creates quintiles based on the estimated propensity
* score;

proc rank groups = 5 out = r;
ranks rnks;
var pr;
data a; set r; quintile = rnks + l;

* This will show the breakdown of subjects by treatment (here epidural) and propensity
  score quintile;
proc freq; tables quintile*epidural;

* This will perform the 2-way anovas to determine whether the propensity score quintiles
* removed the initial bias found by the t-tests above.
proc glm;
    class quintile;
    model amladmit cm l arom gestage birthwt gender rate chyper phyper
    height weight momage insprvt momw momb = quintile epidural quintile*epidural;
```

ACKNOWLEDGEMENTS

The author wishes to thank Dr. Ellice Lieberman and the investigators from the Active Management of Labor Trial (National Institute of Child Health and Human Development grant no. R01-HD26813) for use of their data. The author also wishes to thank Dr. Curtis Deutch and the March of Dimes Birth Defects Foundation Social and Behavioral Science Research grant for use of their data. The author also wishes to acknowledge the support of his wife Carey and daughters Lucy, Serena and Sophia in this work.

REFERENCES

1. Rubin, D. B. and Thomas, N. 'Matching using estimated propensity scores: Relating theory to practice', *Biometrics*, **52**, 249–264 (1996).
2. Bloch, D. A. and Segal, M. R. 'Empirical comparison of approaches to forming strata: using classification trees to adjust for covariates', *Journal of the American Statistical Association*, **84**, 897–905 (1989).
3. Ciampi, A., Hogg, S. A., McKinney, S. and Thiffault, J. 'RECPAM: a computer program for recursive partition and amalgamation for censored survival data and other situation frequently occurring in biostatistics. I. Methods and program features', *Computer Methods and Programs in Biomedicine*, **26**, 239–256 (1988).
4. Rosenbaum, P. R. 'Conditional permutation tests and the propensity score in observational studies', *Journal of the American Statistical Association*, **79**, 565–574 (1984).

5. Rosenbaum, P. R. and Rubin, D. B. 'Reducing bias in observational studies using subclassification on the propensity score', *Journal of the American Statistical Association*, **79**, 516–524 (1984).
6. Rosenbaum, P. R. and Rubin, D. B. 'Constructing a control group using multivariate matched sampling methods that incorporate the propensity score', *American Statistician*, **39**, 33–38 (1985).
7. Lavori, P. W. and Keller, M. B. 'Improving the aggregate performance of psychiatric diagnostic methods when not all subjects receive the standard test', *Statistics in Medicine*, **7**, 723–737 (1988).
8. Cook, E. F. and Goldman, L. 'Asymmetric stratification: An outline for an efficient method for controlling confounding in cohort studies', *American Journal of Epidemiology*, **127**, 626–639 (1988).
9. Fiebach, N. H., Cook, E. F., Lee, T. H., Brand, D. A., Rouan, G. W., Weisberg, M. and Goldman, L. 'Outcomes in patients with myocardial infarction who are initially admitted to stepdown units: data from the multicenter chest pain study', *American Journal of Medicine*, **89**, 15–20 (1990).
10. Muller, J. E., Turi, Z. G., Stone, P. H., Rude, R. E., Raabe, D. S., Jaffe, A. S., Gold, H. K., Gustafson, N., Poole, W. K., Passamani, E., Smith, T. W., Braunwald, E. and The MILIS Study Group. 'Digoxin therapy and mortality after myocardial infarction: experience in the MILIS Study', *New England Journal of Medicine*, **314**, 265–271 (1986).
11. Myers, W. O., Gersh, B. J., Fisher, L. D., Kosinski, A. S., Mock, M. B., Holmes, D. R., Schaff, H. V., Gillispie, S., Ryan, T. J., Kaiser, G. C. and other CASS Investigators. 'Time to first new myocardial infarction in patients with mild angina and three-vessel disease comparing medicine and early surgery: A Cass registry study of survival', *Annals of Thoracic Surgery*, **43**, 599–612 (1987).
12. Myers, W. O., Gersh, B. J., Fisher, L. D., Kosinski, A. S., Mock, M. B., Holmes, D. R., Schaff, H. V., Gillispie, S., Ryan, T. J., Kaiser, G. C. and other CASS Investigators. 'Multiple versus early surgical therapy in patients with triple-vessel disease and mild angina pectoris: A Cass Registry Study of Survival', *Annals of Thoracic Surgery*, **44**, 471–486 (1987).
13. Myers, W. O., Schaff, H. V., Gersh, B. J., Fisher, L. D., Kosinski, A. S., Mock, M. B., Holmes, D. R., Ryan, T. J., Kaiser, G. C. and CASS Investigators. 'Improved survival of surgically treated patients with triple vessel coronary artery disease and severe angina pectoris', *Journal of Thoracic and Cardiovascular Surgery*, **97**, 487–495 (1989).
14. Berk, R. A. and Newton, P. J. 'Does arrest really deter wife battery? An effort to replicate the findings of the Minneapolis Spouse Abuse Experiment', *American Sociological Review*, **50**, 253–262 (1985).
15. Berk, R. A., Newton, P. J. and Berk, S. F. 'What a difference a day makes: an empirical study of the impact of shelters for battered women', *Journal of Marriage and the Family*, **48**, 481–490 (1986).
16. Czajka, J. L., Hirabayashi, S. M., Little, R. J. A. and Rubin, D. B. 'Projecting from advance data using propensity modeling: an application to income and tax statistics', *Journal of Business and Economic Statistics*, **10**, 117–131 (1992).
17. Hoffer, T., Greeley, A. M. and Coleman, J. S. 'Achievement growth in public and Catholic schools', *Sociology of Education*, **58**, 74–97 (1985).
18. Lavori, P. W. 'Clinical trials in psychiatry: should protocol deviation censor patient data?', *Neuropsychopharmacology*, **6**, (1), 39–48 (1992).
19. Lavori, P. W., Keller, M. B. and Endicott, J. 'Improving the validity of Fh-Rdc diagnosis of major affective disorder in uninterviewed relatives in family studies: a model based approach', *Journal of Psychiatric Research*, **22**, 249–259 (1988).
20. Stone, R. A., Obrosky, S., Singer, D. E., Kapoor, W. N. and Fine, M. J. 'Propensity score adjustment for pretreatment differences between hospitalized and ambulatory patients with community-acquired pneumonia', *Medical Care*, **33**, AS56–AS66 (1995).
21. Lieberman, E., Lang, J. M., Cohen, A., D'Agostino, Jr. R., Datta, S. and Frigoletto, Jr., F. D., 'Association of epidural analgesia with caesareans in nulliparous women', *Obstetrics and Gynecology*, **88**, 993–1000 (1996).
22. Rosenbaum, P. R. and Rubin, D. B. 'The central role of the propensity score in observational studies for causal effects', *Biometrika*, **70**, 41–55 (1983).
23. Cochran, W. G. and Rubin, D. B. 'Controlling bias in observational studies: a review', *Sankya, Series A*, **35**, 417–446 (1973).
24. Rubin, D. B. 'Matching methods that are equal percent bias reducing: some examples', *Biometrics*, **32**, 109–120 (1976).
25. Rubin, D. B. 'Using multivariate matched sampling and regression adjustment to control bias in observational studies', *Journal of the American Statistical Association*, **74**, 318–324 (1979).

26. Rubin, D. B. 'Bias reduction using Mahalanobis metric matching', *Biometrics*, **36**, 293–298 (1980).
27. Carpenter, R. G. 'Matching when covariables are normally distributed', *Biometrika*, **64**, 299–307 (1977).
28. Rubin, D. B. 'Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism', *Biometrics*, **47**, 1213–1234 (1991).
29. Hobel, C. J., Youkeles, L. and Forsythe, A. 'Prenatal and intrapartum high-risk screening. II Risk factors reassessed', *American Journal of Obstetrics and Gynecology*, **135**, 1051–1056 (1979).
30. Cochran, W. G. 'The planning of observational studies of human populations', *Journal of the Royal Statistical Society*, *Series A*, **128**, 234–255 (1965).
31. Cochran, W. G. 'The effectiveness of adjustment by subclassification in removing bias in observational studies', *Biometrics*, **24**, 205–213 (1968).
32. Frigoletto, F. D., Lieberman, E., Lang, J. M., Cohen, A. P., Barss, V., Ringer, S. A. and Datta, S. 'A clinical trial of active management of labor', *New England Journal of Medicine*, **333**, 745–750 (1995).
33. Roseman, L. 'Using regression and subclassification on the propensity score to control bias in observational studies', unpublished report, Harvard University, 1994.
34. D'Agostino, R. B. Jr. 'Estimating propensity scores when covariates have either ignorable or nonignorable missing values', Ph.D. thesis, Harvard University, 1994.
35. D'Agostino, R. B. Jr. and Rubin, D. B. 'Estimating and using propensity scores with partially missing data', submitted to *Journal of the American Statistical Association*, (1977).
36. Dempster A. P., Laird N. M. and Rubin D. B. 'Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion)' *Journal of the Royal Statistical Society*, *Series B*, **39**, 1–38 (1977).
37. Meng, X. L. and Rubin, D. B. 'Maximum likelihood estimation via the ECM algorithm: A general framework', *Biometrika*, **80**, 267–278 (1993).
38. Dawson, R. and D'Agostino Jr., R. B. 'Propensity-based non-ignorable models for drop-out in clinical trials', submitted to *Biometrics* (1996).

