American Journal of Epidemiology

Vol. 163, No. 12
DOI: 10.1093/aje/kwj149
Advance Access publication April 19, 2006

Practice of Epidemiology

Variable Selection for Propensity Score Models

M. Alan Brookhart[1], Sebastian Schneeweiss[1], Kenneth J. Rothman[1,2], Robert J. Glynn[1,3], Jerry Avorn[1], and Til Stürmer[1,3]

[1] Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA.
[2] Departments of Epidemiology and Medicine, Boston University Medical Center, Boston, MA.
[3] Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA.

Received for publication June 30, 2004; accepted for publication January 10, 2006.

Despite the growing popularity of propensity score (PS) methods in epidemiology, relatively little has been written in the epidemiologic literature about the problem of variable selection for PS models. The authors present the results of two simulation studies designed to help epidemiologists gain insight into the variable selection problem in a PS analysis. The simulation studies illustrate how the choice of variables that are included in a PS model can affect the bias, variance, and mean squared error of an estimated exposure effect. The results suggest that variables that are unrelated to the exposure but related to the outcome should always be included in a PS model. The inclusion of these variables will decrease the variance of an estimated exposure effect without increasing bias. In contrast, including variables that are related to the exposure but not to the outcome will increase the variance of the estimated exposure effect without decreasing bias. In very small studies, the inclusion of variables that are strongly related to the exposure but only weakly related to the outcome can be detrimental to an estimate in a mean squared error sense. The addition of these variables removes only a small amount of bias but can increase the variance of the estimated exposure effect. These simulation studies and other analytical results suggest that standard model-building tools designed to create good predictive models of the exposure will not always lead to optimal PS models, particularly in small studies.

confounding factors (epidemiology); effect modifiers (epidemiology); models, statistical; propensity score; regression analysis; simulation; subset selection; variable selection

Abbreviations: MSE, mean squared error; PS, propensity score.

Propensity score (PS) methods, as formalized by Rosenbaum and Rubin (1), are becoming standard techniques for controlling confounding in nonexperimental studies in medicine and epidemiology. Unlike conventional statistical approaches that depend on a model of the outcome under study, PS methods rely on a model of the exposure or treatment (termed ''the PS model''). A central issue facing researchers using PS methods is how to select the variables to be included in the PS model. Ideally, specification of the model will be guided by knowledge of the subject matter—for example, a detailed understanding of how a particular treatment is assigned to patients. Frequently, however, the researcher does not have the benefit of such knowledge and instead is confronted with a large collection of pretreatment covariates and many derived functions of these covariates (e.g., interactions) and must decide which of these terms to enter into a regression model of the exposure. The bias and variance of the estimated exposure effect can depend

Correspondence to Dr. M. Alan Brookhart, Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital and Harvard Medical School, 1620 Tremont Street, Suite 3030, Boston, MA 02120 (e-mail: abrookhart@rics.bwh.harvard.edu).

strongly on which of these candidate variables are included in the PS model.

Despite the growing popularity of PS methods, relatively little has been written about the variable selection problem for PS models. In the context of multivariate normal confounders, Rubin and Thomas (2) derived approximations for the reduction in the bias and variance of an estimated exposure effect from PS-matched analysis. They suggest including in a PS model all variables thought to be related to the outcome, whether or not they are related to exposure (2). In a later paper, Rubin (3) suggests that including variables that are strongly related to exposure but unrelated to the outcome can decrease the efficiency of an estimated exposure effect; but he argues that if such a variable had even a weak effect on the outcome, the bias resulting from its exclusion would dominate any loss of efficiency for a reasonable-sized study. Some of these guidelines are repeated by Perkins et al. (4). Robins et al. (5) derived analytical results showing that the asymptotic variance of an estimator based on an exposure model is not increased and is often decreased as the number of parameters in the exposure model is increased. These results suggest that the size of a PS model should increase with the study size. Hirano and Imbens (6) proposed a variable selection strategy for use with a multivariable outcome model employing PS weighting.

In practice, variables are often selected in data-driven ways—for example, by using stepwise variable selection algorithms to develop good predictive models of the exposure (7). Furthermore, in many PS analyses, investigators report the *c* statistic (the area under the receiver operating characteristic curve) for the final PS model as a means of assessing the model's adequacy (7, 8). Implicit in this practice is the assumption that PS models that are better predictors or discriminators of the exposure status result in superior estimators of exposure effect. According to this criterion, any variable that increases the *c* statistic or predictive ability of the PS model should be selected for inclusion in the model. Despite the widespread use of such variable selection strategies, there has been little discussion of their appropriateness. In a recent editorial, Rubin (9) expressed doubt over the usefulness of such diagnostics in a PS analysis.

We conducted the present work to illuminate this issue and to help researchers gain some practical insight into the variable selection problem in a PS analysis. We present the results of two Monte Carlo simulation experiments designed to evaluate how different specifications of a PS model affect the bias, variance, and resulting mean squared error (MSE) of an estimated exposure effect under a variety of assumptions about the data-generating process.

MATERIALS AND METHODS

Brief overview of PS methods

In many nonexperimental cohort studies in medicine and epidemiology, the relation between an exposure A and an outcome Y may be confounded by a set of measured baseline variables $X = (X_1, X_2, \ldots, X_p)$. As potential confounders, the elements of X can be both predictors of the exposure and independent risk factors for the outcome. As an illustration, consider an observational cohort study in which the exposure of interest is the use of a particular cholesterol-lowering drug at the start of the study and the outcome is having a myocardial infarction during the follow-up period. The potential confounders that are measured at the start of the study include age, gender, lipid levels, comorbid conditions, previous drug exposures, and diet and exercise habits. For such studies, statistical methods based on the PS can be used to estimate exposure effects.

The PS is the conditional probability that a subject receives a treatment or exposure under study given all measured confounders, that is, $\Pr[A = 1|X]$. The PS has been termed a balancing score, meaning that among subjects with the same propensity to be exposed, treatment is conditionally independent of the covariates (1). This balancing property suggests that estimates of the exposure effect that are not confounded by any of the measured covariates can be obtained by estimating the effect of exposure within groups of people with the same PS. Within such a group, any difference in outcome between the exposed and unexposed subjects is not attributable to the measured confounders. If treatment assignment is strongly ignorable and other specific assumptions hold, estimates derived from a PS analysis can be interpreted causally (1).

In most nonexperimental research, the true PS will not be known and therefore will need to be estimated, typically according to an assumed model. The bias and variance of the estimated exposure effect can depend strongly on how the model of $\Pr[A = 1|X]$ is specified. The model specification problem includes selecting variables from X to be included in the model and deciding how the variables are to be transformed, categorized, and interacted with one another.

Given an estimated PS, exposure effects are usually estimated by either 1) matching exposed subjects with unexposed subjects on the PS to create two comparable groups, 2) including the PS and the exposure in a multivariable model of the outcome under study, or 3) conducting an analysis stratified across categories of the PS. It is also possible to use the PS to generate inverse-probability-of-exposure weights that are then used in a weighted regression (10). The weighting approach generalizes naturally to longitudinal data with time-varying treatments and confounders. More detailed discussions of PS methods can be found elsewhere (1, 11, 12).

Monte Carlo simulation study

To explore the variable selection problem in PS models, we performed two Monte Carlo simulation experiments. The first examined how the inclusion of three different types of covariates in a PS model affected the estimated exposure effect (see figure 1): 1) a variable related to both outcome and exposure—a true confounder (X_1); 2) a variable related to the outcome but not the exposure (X_2); and 3) a variable related to the exposure but not the outcome (X_3). In the second experiment, we considered how the addition of a single confounder to a PS model changes the bias and variance of an estimated exposure effect under varying assumptions about the strength of the confounder-outcome and confounder-exposure relations.

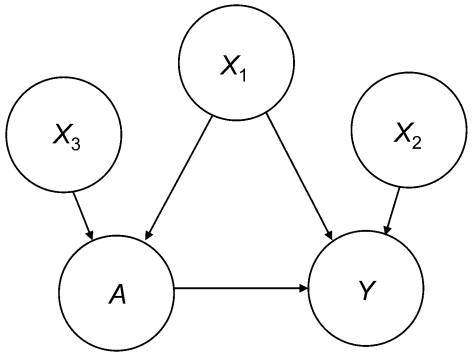

FIGURE 1. Causal diagram for simulation experiment 1.

Both simulation experiments employed the same basic data-generating process. The simulated data consisted of realizations of a dichotomous exposure, an outcome with a Poisson distribution, and continuous confounders. The data were generated in the following order according to the specified probability models:

1. The covariates X_1, X_2, and X_3 are independent standard normal random variables with mean 0 and unit variance.
2. The conditional distribution of the dichotomous exposure A given X_1, X_2, and X_3 follows a Bernoulli distribution with a conditional mean given by the function

$$\Pr[A=1|X_1, X_2, X_3] = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3),$$

   where Φ is the standard normal cumulative distribution function.
3. The conditional distribution of Y given X and A follows a Poisson distribution with two possible specifications of the mean. The first specification (used in the first simulation experiment) is given by

$$\mathbf{E}[Y|A, X_1, X_2, X_3] = \exp\{\alpha_0 + \alpha_1((1+\exp(-3\times X_1))^{-1} - 0.5) + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 A\}.$$

   This specification creates a nonlinear (S-shaped) relation between the confounder X_1 and the log of the expected value of the outcome. The second specification (used in the second simulation experiment) is given by

$$\mathbf{E}[Y|A, X_1, X_2, X_3] = \exp\{\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 A\}.$$

   This model specifies a standard log-linear relation between the covariates and the outcome.

Within both simulation experiments, the effect of exposure is held constant ($\alpha_4 = 0.5$). The simulations differ in how the covariates are related to the exposure and outcome.

We considered two approaches for using the PS to estimate exposure effects. In the first, the exposure effects were estimated by adjusting for the PS in a multivariable Poisson model of the outcome in which the effect of the estimated PS was flexibly modeled through a cubic regression spline with three interior knot points placed at quartiles of the estimated PS. The model is given by

$$\mathbf{E}[Y|PS, A] = \exp\{\lambda + \sum_k \psi_k B_k(PS) + \gamma A\},$$

where λ is the baseline rate, the B_k's are the B-spline basis functions (13), and γ is the treatment effect. The second approach that we employed was based on subclassification. Exposure effects were estimated within strata defined by quintiles of the PS and then averaged across strata to yield an estimate of γ.

Evaluation of PS model performance

The simulation studies presented in this paper compare the performance of various specifications of PS models. To evaluate each PS model, we use the simulation results to determine the variance, bias, and MSE of the corresponding estimator of the exposure effect. Because we have used a log-linear model of the outcome, the parameter estimate $\hat{\gamma}$ from both estimation approaches is consistent for the parameter α_4 from our data-generating distribution at the true PS (14). Therefore, we can estimate the bias of a given estimator with the equation

$$\widehat{\text{Bias}} = \frac{1}{S}\sum_{s=1}^{S}(\hat{\gamma}(s) - \alpha_4)$$

and estimate its MSE with the equation

$$\widehat{\text{MSE}} = \frac{1}{S}\sum_{s=1}^{S}(\hat{\gamma}(s) - \alpha_4)^2,$$

where $\hat{\gamma}(s)$ is the estimated effect of exposure in the sth simulated data set according to a particular PS model and S is the total number of simulations.

Simulation experiment 1

For this experiment, exposure was confounded through X_1, X_3 predicted treatment but was unrelated to the outcome, and X_2 predicted the outcome but was unrelated to treatment ($\alpha_0 = 0.5$, $\alpha_1 = 4$, $\alpha_2 = 1$, $\alpha_3 = 0$, $\beta_0 = 0$, $\beta_1 = 0.5$, $\beta_2 = 0$, $\beta_3 = 0.75$). This scenario is depicted graphically in figure 1.

We simulated 1,000 data sets for both $n = 500$ and $n = 2{,}500$. For each simulated data set, we estimated seven different PS's corresponding to all possible combinations of X_1, X_2, and X_3 in a probit regression model. These models are given by

- PS model 1: $\Pr[A=1|X] = \Phi(\beta_0 + \beta_1 X_1)$.
- PS model 2: $\Pr[A=1|X] = \Phi(\beta_0 + \beta_1 X_2)$.
- PS model 3: $\Pr[A=1|X] = \Phi(\beta_0 + \beta_1 X_3)$.
- PS model 4: $\Pr[A=1|X] = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$.
- PS model 5: $\Pr[A=1|X] = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_3)$.
- PS model 6: $\Pr[A=1|X] = \Phi(\beta_0 + \beta_1 X_2 + \beta_2 X_3)$.
- PS model 7: $\Pr[A=1|X] = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3)$.

For each PS model, we report the estimated bias, variance, and MSE of the corresponding estimator. We also report

TABLE 1. Simulation experiment 1, with results based on an analysis in which the propensity score is entered into an outcome model as a parametric spline term*

	Variable(s) in propensity score model							
	X_1	X_2	X_3	X_1, X_2	X_1, X_3	X_2, X_3	X_1, X_2, X_3	None
$n = 500$								
Bias $\times 10^1$	−0.03	5.97	7.34	−0.03	−0.07	7.36	−0.06	5.94
Var $\times 10^1$	0.32	0.22	0.46	0.22	0.44	0.36	0.31	0.39
MSE $\times 10^1$	0.32	3.79	5.85	0.22	0.44	5.77	0.31	3.92
Average *c* statistic	0.67	0.52	0.76	0.67	0.82	0.76	0.82	
$n = 2{,}500$								
Bias $\times 10^1$	0.00	5.93	7.33	−0.01	−0.04	7.33	−0.03	5.95
Var $\times 10^2$	0.66	0.53	0.96	0.49	0.89	0.79	0.69	0.80
MSE $\times 10^2$	0.66	35.65	54.72	0.49	0.89	54.56	0.70	36.16
Average *c* statistic	0.67	0.51	0.76	0.67	0.81	0.76	0.81	

* The table shows the estimated bias, variance (Var), and mean squared error (MSE) of all possible estimators and the average *c* statistic for the corresponding propensity score model.

these statistics for the simple estimator corresponding to the crude log relative rate. To assess the predictive ability of each PS model, we additionally report the average *c* statistic for the model across simulations. The *c* statistic is computed by forming all discordant pairs of observations (exposed and unexposed combinations) and computing the proportion of these pairs in which the exposed subject had a higher estimated PS than the unexposed subject (15).

We conducted a variety of sensitivity analyses with $n = 500$. These analyses were carried out by holding all parameters at their default values while a single parameter was altered. The following sensitivity analyses were performed: The standard deviation of each covariate was both increased by 50 percent and decreased by 50 percent; the treatment effect was decreased to $\alpha_4 = 0.25$ and increased to $\alpha_4 = 1$; and the baseline prevalence of the exposure was decreased from approximately 50 percent to approximately 20 percent ($\beta_0 = -1$).

Simulation experiment 2

In the second simulation experiment, we examined how the inclusion of a single true confounder in a PS model affected the bias and variance of an estimated exposure effect under varying assumptions about the strength of association between the single confounder and both the outcome and the exposure. For each simulated data set, two estimators were considered: The first was derived from the crude log relative rate, and the second was derived from a PS-adjusted estimate of the effect of A on Y in which the PS model contained only the confounder X_1. In this simulation experiment, the adjustment for the PS used the spline approach. We denote the crude estimator of the log relative rate with $\hat{\gamma}_0$ and the PS-adjusted estimator with $\hat{\gamma}_1$.

The parameter α_1, the strength of association between X_1 and Y, took values in the set $\{0, 0.01, \ldots, 0.20\}$, corresponding to relative rates ranging from 1.00 to 1.28. The parameter β_1, the strength of association between X_1 and A, took values in the set $\{0.00, 0.05, \ldots, 1.25\}$. For all possible combinations of these values of α_1 and β_1, we simulated 1,000 data sets of $n = 500$ and $n = 2{,}500$. In this simulation, the covariates X_2 and X_3 are not used. For each set of 1,000 data sets, we computed the estimated bias, variance, and MSE of each of the two estimators.

Computation

All simulations were performed in R, version 1.9.1 (16, 17), running on a Windows XP platform, using software created by one of the authors (M. A. B.).

RESULTS

Simulation experiment 1

For the simulations controlling for the PS through a spline, we report the estimated bias, variance, and MSE of all estimators in table 1. We also report the average *c* statistic for each candidate PS model. The sole confounder was the covariate X_1; therefore, any estimator that did not contain X_1 in the PS model was biased. For both study sizes, the unbiased estimator with the smallest variance was the one that contained the confounder X_1 and the covariate related to the outcome only, X_2. This estimator had approximately 30 percent ($n = 500$) and 25 percent ($n = 2{,}500$) less variance than the estimator containing just the confounder X_1. Adding X_3, the covariate related only to exposure, increased the variance of the estimated effect for both study sizes. The estimator with all covariates in the PS model had a variance that was approximately 40 percent greater (for both study sizes) than the estimator with just the covariates X_1 and X_2. The *c* statistic of the PS model with X_1 and X_2 was smaller (0.67) than the *c* statistic of the less efficient PS model with all covariates (≈0.8). For both study sizes, the PS models with the highest average *c* statistic contained all variables related to the exposure.

Table 2 shows the results obtained when this simulation experiment was repeated using subclassification instead of

TABLE 2. Simulation experiment 1, with results based on an analysis using subclassification in which strata are defined by quintiles of the estimated propensity score*

	Variable(s) in propensity score model							
	X_1	X_2	X_3	X_1, X_2	X_1, X_3	X_2, X_3	X_1, X_2, X_3	None
$n = 500$								
Bias $\times 10^1$	0.29	6.07	7.96	0.24	0.24	7.93	0.24	5.94
Var $\times 10^1$	0.22	0.14	0.62	0.16	0.71	0.43	0.69	0.39
MSE $\times 10^1$	0.23	3.82	6.95	0.17	0.71	6.71	0.70	3.92
$n = 2{,}500$								
Bias $\times 10^1$	0.28	5.96	7.61	0.29	0.55	7.60	0.56	5.95
Var $\times 10^2$	0.43	0.31	1.02	0.27	1.12	0.87	0.96	0.80
MSE $\times 10^2$	0.51	35.82	58.90	0.35	1.43	58.63	1.27	36.16

* The table shows the estimated bias, variance (Var), and mean squared error (MSE) of all possible estimators.

spline adjustment. The results were qualitatively similar, but there were some notable differences. The effect of adding X_3 was more detrimental in a relative MSE sense. In addition, all unbiased estimators were considerably less variable than the corresponding estimators based on spline adjustment. Finally, all estimators admit some bias due to residual confounding within strata of the PS.

The results of the sensitivity analysis are presented in table 3. In all of the sensitivity analyses, the same essential pattern prevailed: The inclusion of the variable related only to exposure increased the variance of the estimator without altering bias; inclusion of the variable related only to the outcome decreased variance without affecting bias; and failure to include the confounder yielded a biased estimator. However, the perturbation of simulation parameters changed absolute and, in some cases, relative numbers.

Simulation experiment 2

In figure 2, we plot the estimated variance of the PS-adjusted estimator $\hat{\gamma}_1$ and the unadjusted estimator $\hat{\gamma}_0$ across values of β_1 for both $n = 500$ and $n = 2{,}500$. Because the parameter β_1 in the probit model is not directly interpretable, we transform it into a ''relative risk'' (relative exposure prevalence). This is done by computing the probability of treatment at the 75th percentile of X_1 and dividing it by the probability of treatment at the 25th percentile of X_1—in other words, the probability of treatment for someone with a moderately large value of X_1 divided by the probability of treatment for someone with a moderately small value of X_1. For both sample sizes, increasing the value of β_1 (i.e., increasing the strength of association between X_1 and A) increased the variability of the estimated exposure effect $\hat{\gamma}_1$ (the PS-adjusted estimator). The increase in variance did not depend on the strength of association between X_1 and Y (data not presented). The bias of $\hat{\gamma}_0$ increased as the association between either X_1 and Y or X_1 and A increased, unless there was no association between either X_1 and A or X_1 and Y.

In figure 3, we plot contours of the MSE of $\hat{\gamma}_1$ relative to the MSE of $\hat{\gamma}_0$ on a grid of values of α_1 and β_1. The values of β_1 are transformed into relative risks as described previously. This plot indicates values of α_1 and β_1 for which the addition of the confounder X_1 to a PS model is detrimental in an MSE sense; that is, the MSE of $\hat{\gamma}_1$ is greater than the MSE of $\hat{\gamma}_0$. The region between the contour lines at 0.9 and 1.1 represents a zone for which the addition of X_1 to a PS model would have only a moderate effect on the MSE. The region above and to the left of the contour line at 1.1 indicates the region where the analyst might choose to exclude X_1 from the PS, as it would increase the MSE of the estimated exposure effect by more than 10 percent. This region is characterized by large values of β_1 (strong association between X_1 and A) and small values of α_1 (weak association between X_1 and Y). Here the increase in the variance of $\hat{\gamma}_1$ is not offset by a large enough decrease in bias to reduce the MSE of $\hat{\gamma}_1$ relative to $\hat{\gamma}_0$. Similarly, the region below and to the right of the contour line at 0.9 represents the region where the analyst would want to add the confounder to the PS, as it would decrease the MSE by more than 10 percent. Here the bias of an estimator excluding X_1 overwhelms any resulting increase in variance. For $n = 2{,}500$, the same pattern prevailed, but the region for which $\hat{\gamma}_0$ yielded a smaller MSE than $\hat{\gamma}_1$ was greatly reduced.

In figure 4, we plot the contours of the MSE of $\hat{\gamma}_1$ relative to the MSE of $\hat{\gamma}_0$ from the simulation in which exposure effects were estimated using subclassification on the estimated PS. This plot is similar to that seen in figure 3; however, the relative MSE is increased for large values of β_1. This is consistent with the results from simulation experiment 1 and suggests that the variance of a PS estimator-based subclassification may be slightly more sensitive to the strength of the confounder-exposure relation.

DISCUSSION

Our first simulation experiment revealed that the model that best predicted exposure (as measured by the c statistic)

TABLE 3. Sensitivity analysis of simulation study 1*

Parameter change	Variable(s) in propensity score model							
	X_1	X_2	X_3	X_1, X_2	X_1, X_3	X_2, X_3	X_1, X_2, X_3	None
Original								
Bias $\times 10^1$	−0.03	5.97	7.34	−0.03	−0.07	7.36	−0.06	5.94
Var $\times 10^1$	0.32	0.22	0.46	0.22	0.44	0.36	0.31	0.39
MSE $\times 10^1$	0.32	3.79	5.85	0.22	0.44	5.77	0.31	3.92
Decrease in the variance of X_1								
Bias $\times 10^1$	0.13	2.94	3.81	0.12	0.13	3.80	0.12	2.99
Var $\times 10^1$	0.27	0.21	0.47	0.19	0.38	0.38	0.28	0.35
MSE $\times 10^1$	0.27	1.07	1.92	0.19	0.38	1.82	0.28	1.24
Increase in the variance of X_1								
Bias $\times 10^1$	0.06	8.46	10.1	0.02	0.07	10.04	0.02	8.49
Var $\times 10^1$	0.38	0.28	0.50	0.27	0.51	0.41	0.38	0.44
MSE $\times 10^1$	0.38	7.44	10.71	0.27	0.51	10.50	0.38	7.64
Decrease in the variance of X_2								
Bias $\times 10^1$	0.02	5.96	7.38	0.02	0.00	7.38	0.01	5.97
Var $\times 10^1$	0.07	0.13	0.19	0.05	0.12	0.17	0.10	0.16
MSE $\times 10^1$	0.07	3.69	5.64	0.05	0.12	5.62	0.10	3.72
Increase in the variance of X_2								
Bias $\times 10^1$	0.23	6.16	7.59	0.19	0.24	7.53	0.18	6.16
Var $\times 10^1$	1.20	0.60	1.57	1.00	1.62	1.32	1.32	1.25
MSE $\times 10^1$	1.21	4.39	7.33	1.00	1.62	7.00	1.32	5.04
Decrease in the variance of X_3								
Bias $\times 10^1$	0.08	6.89	7.35	0.03	0.04	7.30	0.02	6.94
Var $\times 10^1$	0.34	0.25	0.46	0.23	0.39	0.36	0.28	0.42
MSE $\times 10^1$	0.35	5.00	5.86	0.23	0.39	5.70	0.28	5.24
Increase in the variance of X_3								
Bias $\times 10^1$	0.10	5.07	7.53	0.07	0.05	7.49	0.01	5.1
Var $\times 10^1$	0.29	0.23	0.55	0.21	0.49	0.46	0.39	0.38
MSE $\times 10^1$	0.29	2.80	6.21	0.22	0.49	6.07	0.39	2.98
Decrease in α_4, decrease in the treatment effect								
Bias $\times 10^1$	0.08	5.98	7.46	0.03	0.11	7.41	0.04	6.02
Var $\times 10^1$	0.32	0.24	0.47	0.22	0.43	0.37	0.31	0.39
MSE $\times 10^1$	0.32	3.82	6.03	0.22	0.43	5.86	0.31	4.01
Increase in α_4, increase in the treatment effect								
Bias $\times 10^1$	−0.12	5.68	6.92	−0.08	−0.16	6.99	−0.10	5.62
Var $\times 10^1$	0.34	0.21	0.52	0.23	0.49	0.39	0.33	0.38
MSE $\times 10^1$	0.34	3.43	5.32	0.23	0.49	5.27	0.34	3.54
Decrease in β_0, decrease in exposure prevalence								
Bias $\times 10^1$	0.03	6.02	7.34	0.00	0.01	7.35	0.00	5.99
Var $\times 10^1$	0.32	0.27	0.51	0.23	0.46	0.41	0.34	0.43
MSE $\times 10^1$	0.32	3.90	5.89	0.23	0.46	5.81	0.34	4.01

* We consider nine different perturbations of the simulation parameters. Results are from 1,000 simulations of data ($n = 500$), using a parametric spline to adjust for the estimated propensity score. For each simulation, we report the estimated bias, variance (Var), and mean squared error (MSE) of the estimators corresponding to all possible specifications of the propensity score model.

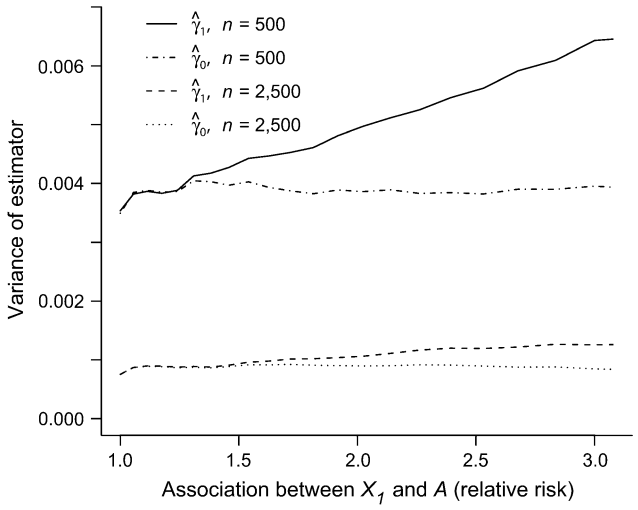

FIGURE 2. Variance of the unadjusted estimator $\hat{\gamma}_0$ and the propensity score-adjusted estimator $\hat{\gamma}_1$ as a function of the strength of association between X_1 and A for $n = 500$ and $n = 2{,}500$.

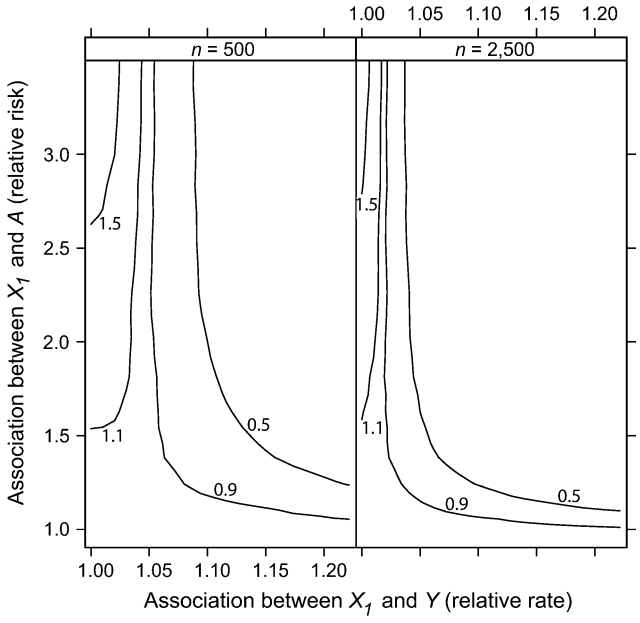

FIGURE 3. Contours of the mean squared error (MSE) of the propensity score (PS)-adjusted estimator relative to the unadjusted estimator, MSE($\hat{\gamma}_1$)/MSE($\hat{\gamma}_0$). For these simulations, the PS is entered into an outcome model as a parametric spline term. This plot shows how the MSE of an estimated exposure effect changes when a variable (X_1) is added to a PS model, according to its association with exposure (A) and outcome (Y).

did not yield the optimal PS model (in terms of MSE). The optimal model was the one that included the confounder and the variable related only to the outcome. This finding is consistent with the advice of Rubin and Thomas (2) that one should include in a PS model all variables thought to be related to the outcome, regardless of whether they are related to the exposure. It might seem unnecessary to include in a PS model a covariate that is known to be unassociated with the exposure. However, for any given realization of a data set, there will usually be some small, statistically insignificant association between such a covariate and the exposure. If that covariate is also related to the outcome, then it is an empirical confounder for that particular study. Including such a covariate in a PS model removes the nonsystematic bias due to the chance association between the covariate and exposure. Across various realizations of a study (or simulated data sets), the removal of this nonsystematic bias tends to bring the estimator closer to its mean, thereby decreasing its variance. This finding is related to the theoretical finding that it is better to use an estimated PS than a known PS (5, 18).

The simulation study also revealed that if a variable unrelated to the outcome but related to exposure is added to a PS model, it will increase the variance of an estimated exposure effect without decreasing bias. Including such a variable in a PS model adds noise to the estimated PS and causes an unnecessary increase in the correlation between the estimated PS and the exposure. In the context of an analysis in which the PS is included as a covariate in an outcome model, increasing the covariance between the exposure and the estimated PS increases the variance of the estimated exposure effect. In the context of a stratified or matched PS analysis, adding noise to the estimated PS may cause subjects to be randomly misclassified or mismatched with respect to important confounders.

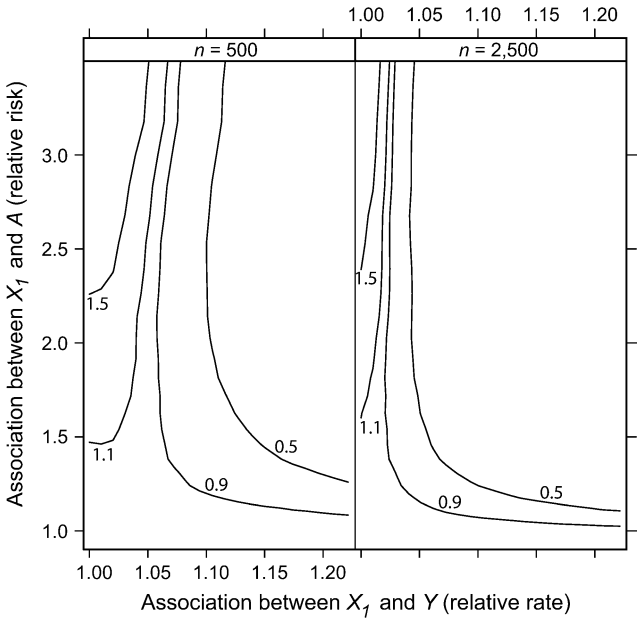

FIGURE 4. Contours of the mean squared error (MSE) of the propensity score (PS)-adjusted estimator relative to the unadjusted estimator, MSE($\hat{\gamma}_1$)/MSE($\hat{\gamma}_0$). For these simulations, the exposure effects are estimated within strata defined by quintiles of the PS. This plot shows how the MSE of an estimated exposure effect changes when a variable (X_1) is added to a PS model, according to its association with exposure (A) and outcome (Y).

The second simulation experiment revealed that if one seeks to minimize the MSE of an estimate, then in very small studies there are situations in which it might be advantageous to exclude true confounders from a PS model. This occurs when a covariate is only weakly related to the outcome but very strongly related to the exposure. The increase in variance due to the inclusion of such a covariate is not offset by a large enough decrease in bias to improve the MSE of the estimator. However, as the study size increases, the variance of the estimator decreases at a rate proportional to $1/n$, yet the bias due to an omitted confounder remains. Therefore, in moderate-sized studies, one would not want to exclude any covariate related to exposure from a PS model unless it was known a priori to be unrelated to the outcome.

Although the results presented in this paper are consistent with theoretical results (e.g., see Rubin and Thomas (2)), the specific numbers are dependent on the specification of the data-generating process and the choice of parameter values considered. Through sensitivity analysis, we varied the parameters that seemed to be the most relevant; however, the probability distributions and other structural elements of the study remained unaltered (e.g., using only three covariates, assuming a constant exposure effect). It is also important to point out that matching and other PS methods can be used in conjunction with standard multivariable outcome models containing additional covariates (19). The variable selection problem in these situations is more complex, as variables can appear in the PS model, the outcome model, or both. The results presented in this paper do not offer insight into the variable selection problem for such hybrid analytical methods.

Our findings and the analytical results presented by Rubin and Thomas (2) and Robins et al. (5) raise questions about the optimality of standard model-building strategies for the construction of PS models, particularly in the setting of small studies. Iterative model-building algorithms (e.g., forward stepwise regression) are designed to create good predictive models of exposure. Similarly, the *c* statistic, which is commonly used to assess the quality of a PS model, is a measure of the predictive ability of the model. The goal of a PS model is to efficiently control confounding, not to predict treatment or exposure. A variable selection criterion based on prediction of the exposure will miss variables related only to the outcome and could miss important confounders that have a weak relation to the exposure but a strong relation to the outcome. Future work in this area should focus on developing and evaluating practical strategies or formal methods (e.g., approaches based on cross-validation (20, 21)) that researchers can use to help them select variables for inclusion in a PS model with an aim of decreasing both the bias and the variance of an estimated exposure effect.

ACKNOWLEDGMENTS

This project was funded by a grant (R01 AG023178) from the National Institute on Aging.

Conflict of interest: none declared.

REFERENCES

1. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;79:516–24.
2. Rubin DB, Thomas N. Matching using estimated propensity score: relating theory to practice. Biometrics 1996;52:249–64.
3. Rubin DB. Estimating causal effects from large data sets using the propensity score. Ann Intern Med 1997;127:757–63.
4. Perkins SM, Tu W, Underhill MG, et al. The use of propensity scores in pharmacoepidemiologic research. Phamacoepidemiol Drug Saf 2000;9:93–101.
5. Robins JM, Mark SD, Newey WK. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. Biometrics 1992;48:479–95.
6. Hirano K, Imbens G. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. Health Serv Outcome Res Methodol 2001;2:259–78.
7. Weitzen S, Lapane KL, Toledano AY, et al. Principles for modeling propensity scores in medical research: a systematic literature review. Phamacoepidemiol Drug Saf 2004;13:841–53.
8. Stürmer T, Joshi M, Glynn RJ, et al. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. J Clin Epidemiol (in press).
9. Rubin DB. On principles for modeling propensity scores in medical research. (Editorial). Phamacoepidemiol Drug Saf 2004;13:855–7.
10. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000;11:550–60.
11. Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. Am J Epidemiol 1999;150:327–33.
12. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998;17:2265–81.
13. Hastie TJ, Tibshirani RJ. Generalized additive models. London, United Kingdom: Chapman and Hall Ltd, 1996.
14. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika 1984;71:431–44.
15. Harrell FE. Regression modelling strategies: with applications to linear models, logistic regression, and survival analysis. New York, NY: Springer-Verlag, 2001.
16. Ihaka R, Gentleman R. R: a language for data analysis and graphics. J Comput Graph Stat 1996;5:299–314.
17. R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2003. (ISBN 3-900051-00-3). (http://www.R-project.org).
18. Rosenbaum PR. Model-based direct adjustment. J Am Stat Assoc 1987;82:387–94.
19. Cochran WG, Rubin DB. Controlling bias in observational studies: a review. Sankhya Ser A 1973;35:417–46.
20. van der Laan MJ, Dudoit S. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite sample oracle inequalities and examples. (U. C. Berkeley Division of Biostatistics Working Paper Series, paper 130). Berkeley, CA: Division of Biostatistics, University of California, Berkeley, 2003. (http://www.bepress.com/ucbbiostat/paper130).
21. Brookhart MA, van der Laan MJ. A semiparametric model selection criterion with applications to the marginal structural model. Comput Stat Data Anal 2006;50:475–98.