PHARMACOEPIDEMIOLOGY AND DRUG SAFETY 2005; **14**: 227–238
Published online 20 July 2004 in Wiley InterScience (www.interscience.wiley.com). **DOI**: 10.1002/pds.986

ORIGINAL REPORT

# Weaknesses of goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder†

Sherry Weitzen PhD[1,2]*, Kate L. Lapane PhD[1,3], Alicia Y. Toledano ScD[1,4], Anne L. Hume PharmD[5,6] and Vincent Mor PhD[1,3]

[1]*Department of Community Health, Brown Medical School, Providence, RI, USA*
[2]*Department of Obstetrics and Gynecology, Division of Research, Women and Infants Hospital, Providence, RI, USA*
[3]*Center for Gerontology and Health Care Research, Brown Medical School, Providence, RI, USA*
[4]*Center for Statistical Sciences, Brown Medical School, Providence, RI, USA*
[5]*Department of Pharmacy Practice, University of Rhode Island, Kingston, RI, USA*
[6]*Department of Family Medicine, Brown Medical School, Providence, RI, USA*

## SUMMARY

**Purpose** Propensity scores are used in observational studies to adjust for confounding, although they do not provide control for confounders omitted from the propensity score model. We sought to determine if tests used to evaluate logistic model fit and discrimination would be helpful in detecting the omission of an important confounder in the propensity score.
**Methods** Using simulated data, we estimated propensity scores under two scenarios: (1) including all confounders and (2) omitting the binary confounder. We compared the propensity score model fit and discrimination under each scenario, using the Hosmer–Lemeshow goodness-of-fit (GOF) test and the c-statistic. We measured residual confounding in treatment effect estimates adjusted by the propensity score omitting the confounder.
**Results** The GOF statistic and discrimination of propensity score models were the same for models excluding an important predictor of treatment compared to the full propensity score model. The GOF test failed to detect poor model fit for the propensity score model omitting the confounder. C-statistics under both scenarios were similar. Residual confounding was observed from using the propensity score excluding the confounder (range: 1–30%).
**Conclusions** Omission of important confounders from the propensity score leads to residual confounding in estimates of treatment effect. However, tests of GOF and discrimination do not provide information to detect missing confounders in propensity score models. Our findings suggest that it may not be necessary to compute GOF statistics or model discrimination when developing propensity score models. 

KEY WORDS — propensity score; logistic regression; goodness-of-fit; c-statistic; residual confounding

## BACKGROUND

The propensity score method is one of a group of analytic tools applied to control for confounding resulting from non-random treatment assignment in observational studies.[1] The validity of results from an observational study is inherently threatened by systematic differences—between exposed and unexposed individuals—in distributions of extraneous factors that could potentially provide an alternative explanation for a relationship observed with the outcome. The propensity score is the conditional probability of being assigned to a treatment, given a constellation of characteristics that are thought to influence treatment selection.[2] Researchers who use this method frequently estimate the propensity score from a multivariable

*Correspondence to: Dr S. Weitzen, Department of Obstetrics and Gynecology, Box G-WIH, Brown University, Providence, RI 02912, USA. E-mail: Sweitzen@wihri.org
†No conflict of interest was declared.

*Received 11 November 2003*
*Revised 25 March 2004*
*Accepted 15 April 2004*

logistic regression model, modeling treatment as a function of these characteristics.[3] Theoretically, applying the propensity score to stratify the sample will create balance between the treated and untreated individuals on these characteristics.[2,4] Therefore, within each stratum the groups are more likely to be directly comparable, similar to a randomized clinical trial.[3]

Rosenbaum and Rubin[2,5,6] clearly state that the propensity score does not control for unobserved confounders that are not included in the model used to estimate the propensity score. In her simulation paper, Drake[7] showed that the resulting bias in the 'causal effect' ranged from 7.2% to 66.4% when the model used to obtain the propensity score estimates excluded a confounder. This bias increased as the strength of the relationship between the confounder and the outcome increased. However, Drake did not demonstrate the imbalance between the treatment groups on the missing confounder after applying the propensity score excluding this confounder. Additionally, she did not evaluate whether statistical measures such as goodness-of-fit (GOF) or model discrimination of the propensity score model might be useful for detecting when an important confounder was omitted from the model.

Achieving balance on covariates between the treatment and comparison groups is the underlying goal of the propensity score method.[2] To evaluate whether the propensity score will provide adequate adjustment for confounding, one must compare the distribution of confounders between treatment groups after applying the propensity score.[2,3,8] However, these analyzes are often not reported in published studies.[9]

From previous research we found that researchers who estimated propensity scores from logistic regression models sometimes reported both the GOF and c-statistic from the propensity score model.[9] The role of these measures in evaluating the ability of the propensity score method to adequately control for confounding when unmeasured confounders exist has not been well explored.

### *GOF and discrimination for logistic regression modeling*

We briefly review the definitions of GOF and discrimination for logistic regression models. When we evaluate the GOF of a logistic regression model, we are interested in whether the differences between the observed values from the data and the predicted values from the model are small and *random*. Lack of good model fit may be due to violations in one or more of the following important assumptions about the logistic model: (1) the logistic model represents the correct function to describe the treatment in terms of the potential confounders; (2) the model includes all important predictors of treatment, continuous variables are in the correct functional form, and the model includes important interactions between variables and (3) the variance of the predicted probabilities are Bernoulli.[10–12]

Tests of model GOF provide information about how well the model describes the data. This is done by summarizing the amount of deviation that exists between the individual observed outcomes and their predicted probabilities from the model into one measure.[10,11,13] In this study, we chose to evaluate model GOF by the Hosmer–Lemeshow test statistic since it was the most frequently used test of model fit in propensity score studies,[9] and the test statistic is available in most statistical programs.[10]

The Hosmer–Lemeshow GOF test statistic is computed after the final model has been selected and the individual patient probabilities are computed from the model. The sample is then ordered by predicted probabilities and *usually* divided into 10 groups of equal size. The sum of the predicted probabilities in each group is compared to the observed number of individuals who experienced the event. A chi-square test statistic is computed and compared to a chi-squared distribution with 8 degrees of freedom (df). Test statistics greater than approximately 16 indicate poor model fit when computed in this way.[10]

Model discrimination is defined as the ability of the model to provide probabilities that accurately classify patients into one treatment group or the other treatment group (in the propensity score case).[10,14,15] In this study, we evaluated model discrimination by computing the area under the receiver operator characteristic (ROC) curve (AUC). We chose to use the AUC as our measure of discrimination because it is easy to interpret and easy to compute using statistical software. Additionally, in studies using the propensity score that assessed model discrimination, the AUC of the propensity score model was most frequently reported.[9]

To compute the AUC, all possible pairs of treated and untreated patients are considered. If the treated patient has a higher predicted probability than the *comparison individual*, then the pair is considered concordant. If the treated patient has a lower probability than the *comparison individual*, then the pair is considered discordant. The AUC is the proportion of concordant pairs. A value of 0.5 indicates that using probabilities predicted from the model to classify patients into treatment groups is as useful as flipping a coin, while a value of 1.0 indicates that the model perfectly predicts

 

each individual's treatment assignment.[10,14] When the outcome being modeled is binary, the AUC is also referred to as the c-statistic.[15]

## STUDY PURPOSE

When using logistic regression methods, we can evaluate how well the model describes the data by assessing the GOF of the model.[10,16] Evaluation of model GOF and discrimination is generally recommended when using logistic regression models for outcome prediction, but not specifically in the context of propensity score methodology.[10,16–20]

GOF and model discrimination could be important in the development of a propensity score to obtain the best possible estimates of the probability of treatment that will, in turn, be most effective in the adjustment of confounding when estimating the treatment effect. The purpose of this study was to examine if model GOF and discrimination are useful in helping us determine how valid our treatment effect estimates will be when using the propensity score method. We focused on examining the effect of an omitted confounder from the propensity score on the propensity score model GOF and model discrimination. In addition, we quantified the effect of using this omitted confounder propensity score on the validity of the treatment effect estimate.

## METHODS

### *Simulated data*

The data for this study were simulated to represent realistic treatment, outcome and confounding variables inspired by a study conducted by Ioannidis *et al.*[21] Patients received either bilateral (BITA) or single (SITA) internal thoracic artery revascularization during coronary artery bypass graft (CABG) procedures. The hypothesis was that BITA is protective against hospital death following surgery.[21]

We created datasets of 2000 observations each to reflect the average sample size among studies employing propensity score methods.[9] In all datasets, treatment (BITA vs. SITA) and outcome (hospital death—yes/no) were binary variables. The prevalence of treatment was approximately 50% and the incidence of hospital death was approximately 10% in these samples. The true relationship between treatment and hospital death was fixed, with a odds ratio of 0.75, indicating a moderate protective effect of BITA compared to SITA. In addition, three confounders were generated in each dataset to represent the following patient characteristics: History of hypertension, yes/no (X1); patient age in years (X2); and percent heart rate variability (HRV) (X3). Table 1 provides the distributional assumptions for each of these confounders. In addition, the confounders were generated to be independent of each other. We focused on three confounders rather than the full complement presented in the paper for ease of interpretation.

The data for the treatment variable (BITA vs. SITA) were generated from the following model:

$$\Pr(T = 1|X_1,\ X_2,\ X_3) = \frac{\exp(\alpha_0 + \alpha_1 \times X_1 + \alpha_2 \times X_2 + \alpha_3 \times X_3)}{1 + \exp(\alpha_0 + \alpha_1 \times X_1 + \alpha_2 \times X_2 + \alpha_3 \times X_3)} \quad (1)$$

Where $\alpha_0$ *is a constant term,* $\alpha_1$ represents the log-odds of BITA ($T = 1$) given a history of hypertension ($X_1 = 1$ vs. $X_1 = 0$); $\alpha_2$ represents the log-odds of BITA for each year increase in age; and $\alpha_3$ represents the log-odds of BITA for each percentage point increase in HRV.

Table 1. Description of simulation parameters for propensity score and outcome models

| Variable | Distribution | Sample prevalence/ mean (SD) | Relationship w/treatment ($\alpha$ coefficient)* | Relationship w/outcome ($\beta$ coefficient)* |
|---|---|---|---|---|
| BITA vs. SITA treatment (T) | Binary | 50% | — | 0.75 (−0.69) |
| Died in hospital outcome (Y) | Binary | 10% | — | — |
| Hypertension confounder (X1) | Binary | 50% | 1.25 (0.22) | 1.25 (0.22) |
| | | | 1.5 (0.41) | 1.5 (0.41) |
| | | | 2.0 (0.69) | 2.0 (0.69) |
| | | | 3.0 (1.10) | 3.0 (1.10) |
| Patient age (years) confounder (X2) | Normal | 65 (10) | 0.96 (−0.04) | 1.04 (0.04) |
| %HRV** confounder (X3) | Normal | 44 (13) | 1.04 (0.04) | 0.96 (−0.04) |

*For explanation of $\alpha$, $\beta$ refer to Equations 1 and 2.
**HRV, heart rate variability.
Treatment, BITA vs. SITA; outcome, death in hospital; confounders, presence of hypertension, age, %HRV.

 

The data for incidence of hospital death, were generated from the following model:

$$\Pr(Y = 1|T,\ X_1, X_2,\ X_3) = \frac{\exp(\beta_0 + \beta_{treat} \times T + \beta_1 \times X_1 + \beta_2 \times X_2 + \beta_3 \times X_3)}{1 + \exp(\beta_0 + \beta_{treat} \times T + \beta_1 \times X_1 + \beta_2 \times X_2 + \beta_3 \times X_3)} \quad (2)$$

In this equation, $\beta_0$ is a constant term. $\beta_{treat}$ represents the log-odds of hospital death ($Y = 1$) given the patient received BITA ($T = 1$ vs. $T = 0$). As stated previously, this relationship was fixed at OR = 0.75 ($\beta_{treat} = -0.69$). The relationship between hospital death and hypertension is represented by $\beta_1$. The log-odds of hospital death for each year increase in age and for each percentage point increase of HRV are represented by $\beta_2$ and $\beta_3$, respectively.

In order to test the degree of residual confounding that resulted from using a propensity score excluding the hypertension variable, we created datasets to examine 16 different combinations of relationships between history of hypertension (X1) and treatment (T), and hypertension (X1) and hospital mortality (Y). The different relationships between these factors are represented as odds ratios and regression coefficients, listed in Table 1.

As seen in Table 1, the relationship between age (X2) and treatment was the same for all datasets generated. As age increased, the probability of receiving BITA over SITA decreased with an OR = 0.96 per year increase ($\alpha_2 = -0.04$ in Eq. 1). Likewise, in all datasets, the relationship between hospital death and age was fixed, with an OR = 1.04 ($\beta_2 = 0.04$ in Eq. 2). We assumed the relationships of HRV and treatment and HRV and hospital death were the same in all datasets. As HRV increased, indicating better cardiac conduction, the probability of receiving BITA increased with an OR = 1.04 per 1 percentage point increase ($\alpha_3 = 0.04$ in Eq. 1). The relationship between hospital death and HRV was fixed, with an OR = 0.96 ($\beta_3 = -0.04$ in Eq. 2).

### *Propensity scores*

To study our questions, we estimated and compared the propensity model GOF and AUC under two scenarios: (1) a full model including history of hypertension (X1), age (X2) and HRV (X3) to develop the propensity score and (2) a propensity score developed with only age (X2) and HRV (X3). To verify that hypertension was indeed an important confounder, we computed the likelihood ratio test comparing the full model propensity score model to the propensity score model without hypertension. Test statistics greater than 3.9[10] indicate that the model including hypertension is better than the one without hypertension. In addition, to evaluate balance under both scenarios, we divided the sample into five strata based on the propensity score and examined if balance on hypertension, age and HRV was achieved between the treatment groups in each stratum. To evaluate the extent of residual confounding in the treatment effect estimate due to omitting hypertension from the propensity score model, we computed treatment effect estimates adjusted by the full model propensity score and by the propensity score from the model that omitted hypertension. We calculated the *absolute value of* percentage bias between the two treatment effect estimates by the following formula:

$$\left| \frac{\left(\exp\left(\hat{\beta}_{\text{treat-pscore w/o } X_1}\right) - \exp\left(\hat{\beta}_{\text{treat-pscore full model}}\right)\right)}{\exp\left(\hat{\beta}_{\text{treat-pscore full model}}\right)} \right| \times 100\% \quad (3)$$

These analyzes were performed for each combination. The results shown represent the average GOF, AUC, treatment effects and residual confounding from 1000 simulations per combination. We computed 95% confidence intervals (CI) around the mean treatment effect estimates. Details of the simulation program are provided in the Appendix A. All simulations and computations were performed using STATA version 7.0 (STATA Corporation, College Station, TX).

## RESULTS

### *Residual confounding after propensity score adjustment*

Table 2 provides the resulting treatment effect estimates from the simulations. Treatment effect estimates obtained by using the full model propensity score were compared to the estimates obtained from using the propensity score when hypertension was omitted from the model. Based on the parameters we selected for simulation, the residual confounding in the treatment effect estimates ranged from a little over 1–31% and was dependent on the strength of the treatment–hypertension relationship as well as the strength of the hospital death–hypertension relationship. We also considered the case where the treatment–hypertension and hospital death–hypertension relationships went in opposite directions. The ranges

 

Table 2. Treatment effect estimates adjusted by full propensity score and propensity score missing hypertension (X1)

| True odds ratio of treatment (hypertension vs. no hypertension) | True odds ratio of hospital death (hypertension vs. no hypertension) | True treatment effect on hospital death | Estimated treatment effect adjusted by full propensity score (95% CI)* | Estimated treatment effect adjusted by propensity score missing hypertension (95% CI)* | % residual confounding (95% CI)* |
|---|---|---|---|---|---|
| 1.25 | 1.25 | 0.75 | 0.75 (0.74–0.76) | 0.76 (0.75–0.77) | 1.1 (1.0–1.2) |
| | 1.5 | 0.75 | 0.76 (0.75–0.77) | 0.77 (0.77–0.78) | 2.0 (1.9–2.1) |
| | 2.0 | 0.75 | 0.76 (0.75–0.77) | 0.79 (0.78–0.80) | 3.7 (3.6–3.8) |
| | 3.0 | 0.75 | 0.76 (0.75–0.77) | 0.81 (0.80–0.81) | 5.5 (5.4–5.7) |
| 1.5 | 1.25 | 0.75 | 0.76 (0.75–0.77) | 0.78 (0.77–0.79) | 2.2 (2.1–2.3) |
| | 1.5 | 0.75 | 0.76 (0.75–0.77) | 0.79 (0.78–0.80) | 3.8 (3.7–4.0) |
| | 2.0 | 0.75 | 0.76 (0.75–0.77) | 0.81 (0.80–0.82) | 6.6 (6.4–6.7) |
| | 3.0 | 0.75 | 0.77 (0.76–0.77) | 0.85 (0.84–0.86) | 10.5 (10.3–10.6) |
| 2.0 | 1.25 | 0.75 | 0.76 (0.75–0.77) | 0.78 (0.77–0.79) | 3.3 (3.1–3.5) |
| | 1.5 | 0.75 | 0.75 (0.75–0.76) | 0.80 (0.80–0.81) | 6.5 (6.3–6.7) |
| | 2.0 | 0.75 | 0.77 (0.76–0.78) | 0.86 (0.85–0.87) | 11.6 (11.4–11.8) |
| | 3.0 | 0.75 | 0.77 (0.76–0.78) | 0.92 (0.91–0.93) | 18.7 (18.4–18.9) |
| 3.0 | 1.25 | 0.75 | 0.76 (0.75–0.77) | 0.80 (0.79–0.81) | 5.4 (5.2–5.7) |
| | 1.5 | 0.75 | 0.77 (0.76–0.78) | 0.85 (0.84–0.86) | 10.5 (10.2–10.7) |
| | 2.0 | 0.75 | 0.77 (0.76–0.78) | 0.91 (0.90–0.92) | 19.0 (18.7–19.3) |
| | 3.0 | 0.75 | 0.78 (0.77–0.78) | 1.0 (1.0–1.0) | 30.5 (30.2–30.9) |

*Average over 1000 simulations.

(and magnitude) of absolute residual confounding resulting from these simulations were similar regardless of the direction of the relationships (<1–28%, results not shown).

Figure 1 displays the observed three-way relationship between the absolute value of residual confounding, the odds of treatment given hypertension, and the relative risk of hospital death given hypertension. The graph indicates that for each treatment effect on the outcome, residual confounding increases moderately as the odds of treatment increases. However, the increase in residual confounding appears to increase more dramatically as the odds of the outcome increases, regardless of the relation between treatment and hypertension.

### *Relationship of GOF, AUC and percent residual confounding in treatment effect estimates*

Table 3 contains the results of the propensity model likelihood ratio test, Hosmer–Lemeshow GOF test statistic and the area under the ROC curve under the different relationships for hypertension and treatment. The likelihood ratio test indicates that in each case, the propensity score model including hypertension is a better model than the one that omits hypertension.

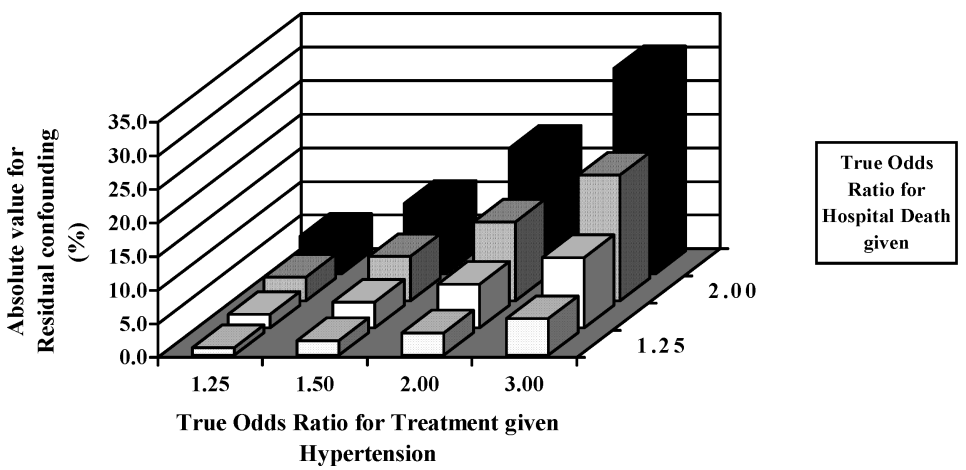


Figure 1. Residual confounding due to omitting hypertension from propensity score used to adjust treatment effect estimates

 

Table 3. Goodness-of-fit (GOF) test statistic and area under receiver operator characteristic (ROC) curve (AUC) for propensity score models

| True odds ratio of treatment (hypertension vs. no hypertension) | LR-test (1 degree of freedom (df))* | GOF full propensity score model | GOF propensity score model excluding hypertension | AUC full propensity score model | AUC propensity score model excluding hypertension |
|---|---|---|---|---|---|
| 1.25 | 6.71 | 7.9 | 8.0 | 0.67 | 0.67 |
| 1.5 | 19.6 | 8.0 | 7.9 | 0.68 | 0.67 |
| 2.0 | 54.7 | 7.9 | 7.9 | 0.69 | 0.67 |
| 3.0 | 134.4 | 8.1 | 8.2 | 0.72 | 0.66 |

*LR test, likelihood ratio test statistic with 1 df $> 3.9$ indicates the model with hypertension is better than the model excluding hypertension.

Whether or not the propensity score model contained the variable hypertension, the GOF of the model was within acceptable bounds ($\chi^2$ with 8 df $< 16$). There was little variation in these results regardless of the strength of the relationship between treatment and hypertension. Likewise, there was no relationship between the GOF test statistic from the propensity score models and the residual confounding in the treatment effect estimates.

Each of the propensity score models had AUCs less than what is considered acceptable model discrimination, which is between 0.7 and 0.8,[10,19,20] regardless of whether they were estimated with all three confounders, or with just age and HRV. However, in each case, the AUCs were slightly smaller for the propensity score models containing only age and HRV compared to the full propensity score model for each variation of the treatment–hypertension relationship. Similar to the GOF test statistic, the AUC had no relationship with residual confounding in the treatment effect estimates.

### *Balance on confounders*

Figure 2 shows the balance across treatment groups for the full model propensity score (Panel A) and for the propensity score missing hypertension (Panel B). We used results from simulations where the odds ratio of treatment given hypertension and the odds ratio of hospital death given hypertension both equaled 3.0, as an example. However, we found similar trends from the other combinations of hypertension–treatment/hypertension–outcome relationships. Overall, the proportion of patients with hypertension was higher in the BITA group compared to the SITA group (62% vs. 37%). The mean age of patients in the BITA group was younger than the SITA group (63 years vs. 67 years). HRV was on average higher among patients in the BITA group compared to those in the SITA group (47% vs. 41%). Figure 2, Panel A illustrates the balance between treatment groups on each of the three confounders, hypertension (X1), age (X2) and HRV (X3) across quintiles based on the full model propensity score.

The right side of Figure 2 (Panel B) displays the results of sample stratification using the propensity score that included only age and HRV in the model.

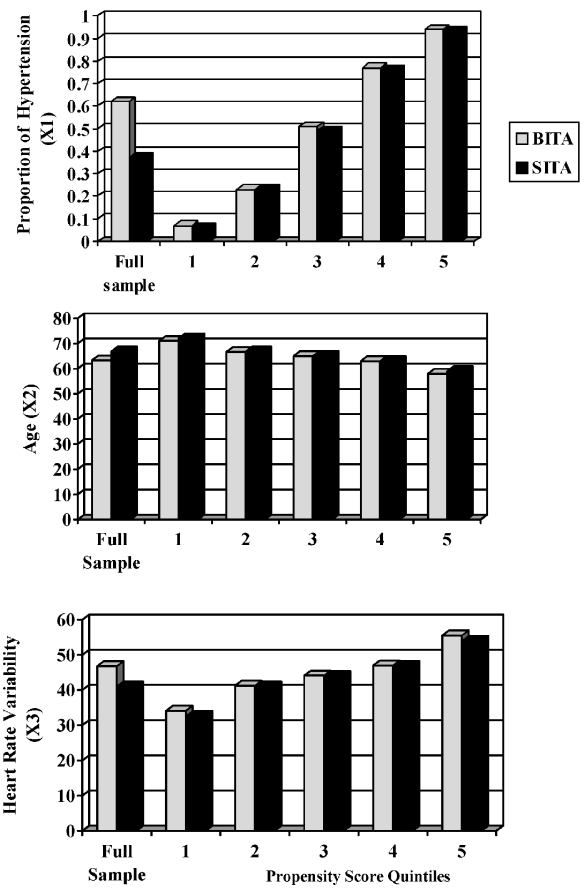


Figure 2. Balance between BITA (T = 1) and SITA (T = 0) treatment groups across propensity score strata on hypertension (X1), age (X2) and heart rate variability (HRV) (X3). Panel A, full model propensity score. Panel B, propensity score missing hypertension

 

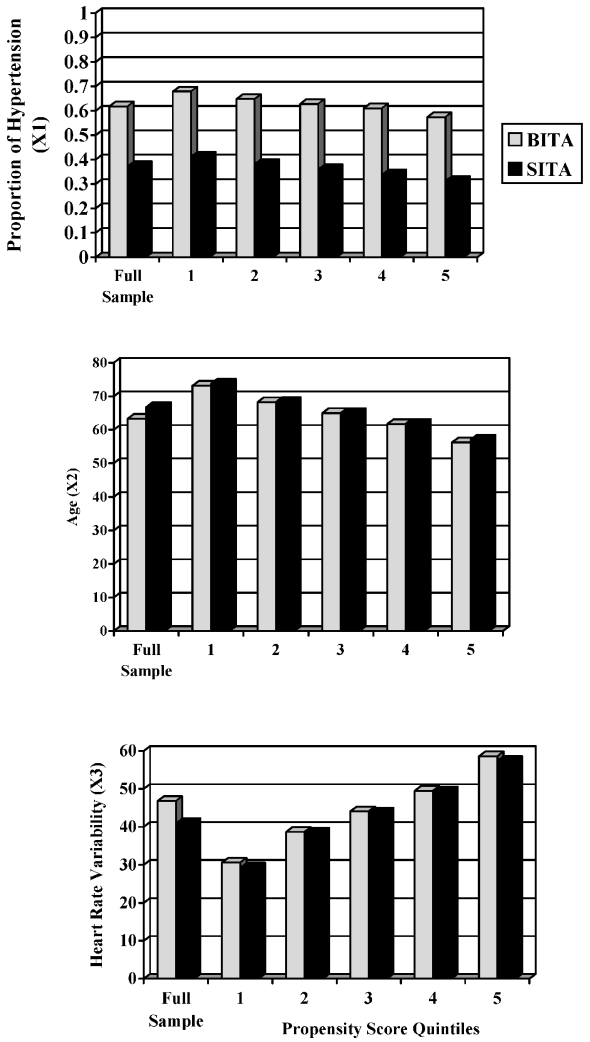


Figure 2. Continued

These results clearly show that while balance between treatment groups is achieved for age and HRV, the proportion of hypertension remains higher in the BITA group across each stratum.

## DISCUSSION

In our simulations, we found that checking the GOF of the propensity score model did not provide any information regarding important confounders that might have been missing from the model. Propensity score models excluding an important confounder had virtually equal Hosmer–Lemeshow GOF test statistics as the full model propensity score. This confirms the recommendations that confounding factors should be evaluated based on their relationships to the exposure and outcome under study rather than by statistical tests.[2,3,8,22]

Additionally, model discrimination as measured by the area under the ROC curve decreased only slightly for propensity score models that excluded an important confounder. As expected, however, using propensity scores developed omitting a confounder resulted in biased estimates of treatment effect compared to those that were adjusted using the full model propensity score. In this example, the residual confounding ranged from 1.1% to over 30%. We also demonstrated that balance between treatment groups is only achieved on confounders that are included in the propensity score. Evaluating balance between groups is more critical than statistical tests of model fit and model discrimination with respect to the effectiveness of the propensity score to adjust for confounding.[2,3,8]

Similar to the study done by Drake,[7] we found that the amount of residual confounding in treatment effect estimates due to omitting a confounder from the propensity score model increased as the relationship between the confounder and treatment and the relationship between confounder and outcome was strengthened. However, in her simulations, the amount of bias was much larger (7.2–66.4%).[7] There are several possible explanations for the differences in results between the two studies. Drake simulated samples of 100, whereas the samples we simulated included 2000 observations. Just as randomization is thought to achieve better balance on covariates between treatment groups when there are large numbers of study subjects, Rubin states that propensity scores are also more effective in larger samples.[4]

After omitting one confounder from her propensity score model, Drake included only one variable in the propensity scores; our propensity scores, after omitting hypertension, included two variables that were moderately predictive of treatment. Additionally, Drake omitted a continuous confounder, while our omitted confounder was binary. Another possible explanation for the differences in magnitude of bias between our studies is that Drake varied the incidence of the outcome and the prevalence of treatment for each combination of parameters she tested, while in our samples, we fixed the incidence of the outcome to be 10% and the prevalence of the treatment to be 50%.

It seems counter-intuitive that the propensity score model that excludes an important confounder would fit equally as well as the full model propensity score. However, there are known limitations of the summary

 

GOF tests available to evaluate logistic regression models.[11,12,20] In particular, for the Hosmer–Lemeshow test, the sample is usually collapsed into 10 groups with equal number of individuals, based on the propensity scores. These groups may contain patients that have very different covariate patterns, despite similar probabilities of treatment. 'The original design structure of the data . . . only contributes to the final classification via the magnitude of fitted probabilities'.[12] In addition, by grouping the sample based on propensity scores, we may mis-identify the lack of fit in one particular individual or group of individuals, while the model appears to fit overall.[10–12,20] Therefore the power of the test to detect overall lack of model fit is quite low. A recent simulation study demonstrated this by testing the GOF of models developed omitting an important confounder. The GOF tests indicated that the model was poorly fit in less than 5% of the replications.[12] The deviance statistic and other summary measures of GOF also have limited power to detect poor fit.[11,12]

Another limitation of the Hosmer–Lemeshow GOF test is that it is sensitive to how many groups the sample is divided into, how the cut points are determined and how ties are handled.[11,23] While we only evaluated the GOF fit test based on 10 groups of equal size, using other methods may have resulted in large GOF test statistics indicating poor model fit. However, the criterion we used to create the groups based on predicted probabilities is the most commonly used in the literature.[10] Indicators of model discrimination, such as the area under the ROC curve, appear to have limited utility in detecting the omission of a confounder from the propensity score model. A model that does not adequately describe the data (i.e. when there are large differences between the propensity scores and the observed treatment assignments) can still have good discrimination because the patients are appropriately classified into treatment groups.[10] Classification into treatment groups may still be adequate despite the omission of a confounder, due to the inclusion of other important predictors of treatment in the propensity score model.

Ideally, balance between the treatment groups is the ultimate goal of using the propensity score method. If balance is achieved, then the treatment groups are thought to be comparable in a similar way as if the study was a randomized trial.[3] Propensity scores not only fail to provide control for unobserved confounders, they will not balance on confounders that are not included in the propensity score model.[4,24,25] Therefore careful decisions must be made about which variables need to be included in the study prior to data collection. Additionally, after the data are collected, thoughtful variable selection into the propensity score model using the available information must be done.

Our results as well as the results from Drake's study[7] indicate that residual confounding increases more rapidly as the strength of the confounder–outcome relationship increases as compared to the strength of the confounder–treatment relationship. However, the maximum residual confounding occurs when both treatment and outcome are strongly related to the confounding variable implying a synergistic relationship. Therefore, when using the propensity score method, we should not only check the balance between treatment groups on a variable but also assess whether this variable is related to the outcome before including or excluding it from the propensity score.

The magnitude of residual confounding in the treatment effect estimate in our study was not very large when using a propensity score that omitted a confounder. This could imply that the propensity score is fairly robust to unobserved confounders if at least some of the key variables that explain treatment assignment are included in the score. Sensitivity analyzes in individual studies have been done to examine how strong the relationships need to be between a hypothetical omitted confounder and treatment, as well as the confounder and outcome, before the estimated treatment effect is substantially changed.[24,26] Rosenbaum and Rubin[2] also found that an omitted binary confounder would have to triple the odds of the treatment, and more than triple the odds of the outcome before their conclusions would change regarding the beneficial effect of the treatment they were studying. Heckman et al.[27] assert that not controlling for unobserved confounders contributes a relatively small amount of bias in non-experimental studies compared to other sources. Nevertheless, the bias due to unobserved confounders continues to be a concern when weighing the evidence between randomized and observational studies.[28]

Our study was limited in several ways. There are several factors that could be related to both the fit of the propensity score and its ability to adequately control confounding in the estimates of treatment effect. We only examined the effect of an omitted confounder. However, other violations of good model fit of the propensity score model may be better detected by GOF tests.[11] These violations include: incorrect functional form of continuous variables, exclusion of important interactions between confounders, or use of the incorrect link function to model treatment. Using a propensity score that was incorrectly modeled due to these problems may cause additional bias in treatment

 

effect estimates. However, the most important criteria for evaluating the effectiveness of the propensity score to control for confounding is to assess whether it creates balance between treatment groups on these confounders.[8]

We attempted to generate variables—treatment, outcome and confounders—that resembled data that might be seen in practice. However, we included only three confounders in the propensity score model. In practice, propensity scores are estimated using many more variables in an effort to adjust for all possible confounders. Additionally, our simulations represent only one scenario. Other combinations of treatment and confounder prevalence, outcome incidence and the treatment effect could yield differing magnitudes of residual confounding. However, our results were consistent with Drake's study, demonstrating that omitting an important confounder from a propensity score model will result in varying degrees of bias in treatment effect estimates adjusted by this propensity score.

There are several benefits to using the propensity score method to adjust the effect of a treatment on an outcome. Because the propensity is a summary of all the confounders included in the model used to estimate it, it can be used as a single variable in a matched, stratified or multivariable analysis while still providing adjustment for all included confounders.[4,24,25] When an outcome is rare, this method is particularly efficient.[24,29] Standard multivariable methods modeling the outcome provide limited control on the confounders because so few variables can be included in the model. If the treatment is fairly common, the propensity score model can accommodate many variables, and then the score can be used to adjust the estimated treatment effect as the only covariate with treatment, in the second stage multivariable model of the outcome, or with limited additional covariates.[4,24] However, Hahn[30] argues that using the propensity score to control confounding may actually decrease the efficiency of the treatment effect estimates, and he suggests that there are better statistical techniques to improve efficiency when controlling for confounding in observational studies.

Despite its attractive efficiency to provide adjusted estimates of treatment effect, propensity score methods do not completely address the limitations of observational studies due to non-random treatment assignment. Still, many researchers believe that 'The problem of unmeasured confounders can be addressed by using epidemiologic data to mimic as closely as possible the design of an RCT. Propensity scores, marginal structural models and structured nested models provide the statistical tools to estimate causal effects from observational data'.[31] However, Rosenbaum and Rubin were cautious about making such assertions,[2] in part, as demonstrated in this study, because propensity score methods do not provide control for factors related to treatment that are neither included in their estimation, nor unobserved or unknown.

## APPENDIX: STATA CODE FOR SIMULATIONS

*Note: 'number' given by the programmer when running simulation (see Table 1 for the assumptions used in the study).

```
**Step 1—creating data
    drop _all
    set obs 2000
**Assumptions
*X1-binary variable with prevalence = 0.5
    gen x1 = uniform() < 0.5

*X2-normally distributed variable with mean = 65,
 SD = 10
    gen x2 = invnorm(uniform())*10 + 65

*X3-normally distributed variable with mean = 44,
 SD = 13
    gen x3 = invnorm(uniform())*13 + 44

**Treatment
**Constant term will be set so that prevalence of
  treatment = 0.5
    local truea0 = '1'
    local truea1 = '2'
    local truea2 = '3'
    local truea3 = '4'
    gen ztreat = '1' + '2'*x1 + '3'*x2 +
    '4'*x3
    gen ptreat = exp(ztreat)/(1 + exp
    (ztreat))
    gen treat = uniform() < (ptreat)
    sum treat
    local prev_treat = r(mean)

**Table 1—distribution of confounders by treat-
  ment
    sum x1 if treat == 1
    local tr1x1 = r(mean)
    sum x1 if treat == 0
    local tr0x1 = r(mean)
    sum x2 if treat == 1
    local tr1x2 = r(mean)
    sum x2 if treat == 0
    local tr0x2 = r(mean)
```

 

```
sum x3 if treat == 1
local tr1x3 = r(mean)
sum x3 if treat == 0
local tr0x3 = r(mean)
```

**Outcome
**Constant term will be adjusted so that prevalence of outcome ~0.1

```
local trueb0 = '5'
local trueb1 = '6'
local trueb2 = '7'
local trueb3 = '8'
local truetreat = '9'
local trueor = exp('9')
gen zout = '5' + '9'*treat + '6'*x1 +
'7'*x2 + '8'*x3
gen pout = exp(zout)/(1 + exp (zout))
gen out = uniform() < (pout)
sum out
local prev_out = r(mean)
```

**Step 2—estimate the propensity score

**Estimated propensity score

```
logit treat x1 x2 x3
predict eprop
matrix b = get(_b)
local propa0 = _b[_cons]
local propa1 = _b[x1]
local propa2 = _b[x2]
local propa3 = _b[x3]
lfit, group(10)
local hltrue = $S_3
lroc, nograph
local auctrue = $S_2
xtile estpsst = eprop, nq(5)
```

**Checking balance of x1

```
sum x1 if treat == 1 & estpsst == 1
local etr1x1st1 = r(mean)
sum x1 if treat == 0 & estpsst == 1
local etr0x1st1 = r(mean)
sum x1 if treat == 1 & estpsst == 2
local etr1x1st2 = r(mean)
sum x1 if treat == 0 & estpsst == 2
local etr0x1st2 = r(mean)
sum x1 if treat == 1 & estpsst == 3
local etr1x1st3 = r(mean)
sum x1 if treat == 0 & estpsst == 3
local etr0x1st3 = r(mean)
sum x1 if treat == 1 & estpsst == 4
local etr1x1st4 = r(mean)
sum x1 if treat == 0 & estpsst == 4
local etr0x1st4 = r(mean)
sum x1 if treat == 1 & estpsst == 5
local etr1x1st5 = r(mean)
sum x1 if treat == 0 & estpsst == 5
local etr0x1st5 = r(mean)
```

**Check balance for x2 and x3 in same way as above (code not shown)

**Step 3—estimating treatment effect

**crude OR

```
cs out treat
local crude = r(rr)
```

**Usual multivariable logistic regression

```
logit out treat x1 x2 x3
matrix b = get(_b)
local mvmodelb = _b[treat]
local mvmodelrr = exp(_b[treat])
```

**Stratification by propensity score and combining stratum specific estimates

```
cs out treat, by(estpsst)
local propmhrr = r(rr)
```

**Propensity score in multivariable model

```
logit out treat eprop
matrix b = get(_b)
local propmvb = _b[treat]
local propmvrr = exp(_b[treat])
```

**Step 4—estimating propensity score missing x1

```
logit treat x2 x3
predict ocprop
matrix b = get(_b)
local ocpropa0 = _b[_cons]
local ocpropa2 = _b[x2]
local ocpropa3 = _b[x3]
lfit, group(10)
local ochl = $S_3
lroc, nograph
local ocauc = $S_2
xtile ocpsst = ocprop, nq(5)
```

**Step 5—checking balance of x1 by stratifying on propensity score

```
sum x1 if treat == 1 & ocpsst == 1
local octr1x1st1 = r(mean)
sum x1 if treat == 0 & ocpsst == 1
local octr0x1st1 = r(mean)
sum x1 if treat == 1 & ocpsst == 2
local octr1x1st2 = r(mean)
sum x1 if treat == 0 & ocpsst == 2
local octr0x1st2 = r(mean)
sum x1 if treat == 1 & ocpsst == 3
local octr1x1st3 = r(mean)
```

```
sum x1 if treat == 0 & ocpsst == 3
local octr0x1st3 = r(mean)
sum x1 if treat == 1 & ocpsst == 4
local octr1x1st4 = r(mean)
sum x1 if treat == 0 & ocpsst == 4
local octr0x1st4 = r(mean)
sum x1 if treat == 1 & ocpsst == 5
local octr1x1st5 = r(mean)
sum x1 if treat == 0 & ocpsst == 5
local octr0x1st5 = r(mean)
```

**Check balance for x2 and x3 in same way as above (code not shown)**

**Step 6—obtaining treatment effect estimates from propensity score without x1**

**Usual multivariable logistic regression**

```
logit out treat x2 x3
matrix b = get(_b)
local ocmvmodelb = _b[treat]
local ocmvmodelrr = exp(_b[treat])
```

**% residual confounding compared to true treatment effect**

```
local resconf = (exp(_b[treat] -
exp('9'))/(exp('9'))
```

**Stratification by propensity score and combining stratum specific estimates**

```
cs out treat, by(ocpsst)
local ocpropmhrr = r(rr)
```

**Propensity score in multivariable model**

```
logit out treat ocprop
matrix b = get(_b)
local ocpropmvb = _b[treat]
local ocpropmvrr = exp(_b[treat])
```

## REFERENCES


1. Robins JM, Mark SD, Newey WK. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* 1992; **48**: 479–495.
2. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**: 41–55.
3. D'Agostino RB, Jr. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med* 1998; **17**: 2265–2281.
4. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med* 1997; **127**: 757–763.
5. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc* 1984; **79**: 516–524.
6. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Statistician* 1985; **39**: 33–38.
7. Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics* 1993; **49**: 1231–1236.
8. Wang J, Donnan PT. Propensity score methods in drug safety studies: practice, strengths and limitations. *Pharmacoepidemiol Drug Safe* 2001; **10**: 341–344.
9. Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiol Drug Safe* 2004; (in press).
10. Hosmer DL, Lemeshow S. *Applied Logistic Regression*. John Wiley and Sons: New York, 2000.
11. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness of fit tests for the logistic regression model. *Stat Med* 1997; **16**: 965–980.
12. Pulkstenis E, Robinson TJ. Two goodness-of-fit tests for logistic regression models with continuous covariates. *Stat Med* 2002; **21**: 79–93.
13. Miller M, Hui SL, Tierney WM. Validation techniques for logistic regression models. *Stat Med* 1991; **10**: 1213–1226.
14. Margolis D, Bilder W, Boston R, Localio R, Berlin JA. Statistical characteristics of area under the receiver operating characteristic curve for a simple prognostic model using traditional and bootstrapped approaches. *J Clin Epidemiol* 2002; **55**: 518–524.
15. Harrell FE, Jr, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. *Stat Med* 1984; **3**: 143–152.
16. Hosmer DW, Taber S, Lemeshow S. The importance of assessing the fit of logistic regression models: a case study. *Am J Public Health* 1991; **81**: 1630–1635.
17. Feinstein A. *Multivariable Analysis: An Introduction*. Yale University Press: New Haven, 1996.
18. Bagley SC, White H, Golomb BA. Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. *J Clin Epidemiol* 2001; **54**: 979–985.
19. Harrell FE, Jr, Lee KL, Mark DB. Tutorial in biostatistics: multivariable prognostic models. Issues in developing models evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; **15**: 361–387.
20. Harrell FE, Jr. *Regression Modeling Strategies*. Springer-Verlag: New York, 2001.
21. Ioannidis JP, Galanos O, Katritsis D, *et al*. Early mortality and morbidity of bilateral versus single internal thoracic artery revascularization: propensity and risk modeling. *J Am Coll Cardiol* 2001; **37**: 521–528.
22. Greenland S. Model and variable selection in epidemiologic analysis. *Am J Public Health* 1989; **79**: 340–349.
23. Bertolini G, D'Amico R, Nardi D, Tinazzi A, Apolone G. One model, several results: the paradox of the Hosmer–Lemeshow goodness-of-fit test for the logistc regression model. *J Epidemiol Biostat* 2000; **5**: 251–253.
24. Braitman L, Rosenbaum PR. Rare outcomes, common treatments: analytic strategies using propensity scores (editorial). *Ann Intern Med* 2002; **137**: 693–696.
25. Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. *Am J Epidemiol* 1999; **150**: 327–333.
26. Connors A, Speroff T, Dawson NV, *et al.* The effectiveness of right heart catheterization in the initial care of critically ill patients. SUPPORT Investigators. *JAMA* 1996; **276**: 889–897.
27. Heckman J, Ichimura H, Todd PE. Matching as an economic evaluation estimator: evidence from evaluating a job training programme. *Rev Econ Stud* 1997; **64**: 605–654.


 

28. Radford MJ, Foody JM. How do observational studies expand the evidence base for therapy? *JAMA* 2001; **286**: 1228–1230.
29. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol* 2003; **158**: 280–287.
30. Hahn J. The role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 1998; **66**: 315–331.
31. Michels K. Hormone replacement therapy in epidemiologic studies and randomized clinical trials—are we checkmate? *Epidemiology* 2003; **14**: 3–5.

 