5-15-2003

A Bootstrap Confidence Interval Procedure for the Treatment Effect Using Propensity Score Subclassification

Wanzhu Tu
Indiana University, wtu1@iupui.edu

Xiao-Hua Zhou
University of Washington, azhou@u.washington.edu

1 Introduction

Causal inferences on the average treatment effect in observational studies are difficult because the effect could be confounded with the covariates whose distributions differ systematically in the two treatment groups, and a direct estimation of the treatment effect is often biased (Rubin 1979, Rosenbaum and Rubin 1983, Greenland, Robins and Pearl 1999). Propensity score subclassification has been shown to be an effective way of reducing this bias in the point estimation of the average treatment effect (Rosenbaum and Rubin 1984, Rubin 1997). However, the inference procedures concerning this average treatment effect have not been well developed. A commonly used approach is to stratify the data based on the estimated propensity scores and carry out the desired inferences as in a stratified random sample. Examples of such analyses can be found in Stone, Obrosky and Singer (1995), D'Agostino (1998), and Perkins, Tu, Underhill, Zhou and Murray (2000).

The validity of such procedures, however, is rather questionable. Because the subclassification is based on the propensity scores estimated from a common logistic model, the responses within each subclass and between the subclasses are not likely to be independent (Du 1998). At the meantime, the estimation of the unknown propensity scores also presents an addition source of variation, which could affect the variance estimate used in the inference (Tu, Perkins, Zhou and Murray 1999).

In this paper, we introduce a bootstrap confidence interval that takes into account the dependent structure of the propensity score stratified data, and the extra variation arisen from the propensity score estimation, under an assumption that the measured covariates can be balanced within all the subclasses based on estimated propensity scores. Unlike the currently used methods, this procedure does not require an estimation of the variance quantity for the purpose of inference. Nor does it assume any specific distribution for the pivotal statistic used in the traditional confidence interval construction.

This paper is organized as follows: In section 2, we briefly review the practice of the propensity score method in causal inferences. We also discuss the deficiencies and limitations of the currently used methods. In Section 3 we propose a bootstrap confidence interval for the treatment effect using propensity score subclassification. In Section 4 we report some preliminary simulation results on the performance of the proposed method. In Section 5, we illustrate the proposed bootstrap method through a clinical example. We conclude the paper in Section 6 with a brief discussion on the potential use

of the new method in practice.

2 Propensity score method in causal inferences

To expedite the discussion, we first introduce the notation. For an individual subject, we denote the treatment assignment as Z ($Z = 1$ if the subject is in the treatment group, $Z = 0$ if the subject is in the control group) and the covariate vector as $\mathbf{X}$. The propensity score is defined as $e(\mathbf{X}) = Pr(Z = 1|\mathbf{X})$. For each subject, the observed outcome variable $Y(Z)$ takes one of the two possible values ($Y(1)$ or $Y(0)$) depending on the treatment assignment that the subject receives. Under this notation, the expected causal effect of the treatment is defined by

$$\delta = E(Y(1) - Y(0)). \tag{1}$$

A fundamental problem in estimating δ is that only one of the two potential outcomes $Y(0)$ and $Y(1)$ is observed so that one cannot directly estimate δ from the observed data (Rosenbaum and Rubin 1984, Rubin 1997).

One way to overcome this difficulty is to use the propensity score $e(\mathbf{X})$. Under the strongly ignorable treatment assignment assumption, Rosenbaum and Rubin (1983) showed that

$$\delta = E_{e(\mathbf{X})}\Big[E\{Y(1) \mid e(\mathbf{X}), Z = 1\} - E\{Y(0) \mid e(\mathbf{X}), Z = 0\}\Big],$$

where $E_{e(\mathbf{X})}$ denotes expectation with respect to the distribution of $e(\mathbf{X})$ in the population of subjects. If we can stratify the subjects into K homogeneous subclasses $I_1, \ldots, I_K$ based on their propensity scores, so that these scores remain constant within each subclass, and suppose that at least one subject in each subclass receives each of the treatment, we then have

$$\begin{aligned} \delta &= \sum_{k=1}^{K} P(e(\mathbf{X}) \in I_k)\Big(E\{Y(1) \mid e(\mathbf{X}) \in I_k, Z = 1\} \\ &\quad - E\{Y(0) \mid e(\mathbf{X}) \in I_k, Z = 0\}\Big). \end{aligned} \tag{2}$$

See Corollary 4.2 of Rosenbaum and Rubin (1983) for a detailed discussion on the result (2).

Since the propensity score $e(\mathbf{X})$ is rarely known, we usually estimate the unknown propensity scores via a logistic model,

$$\log \frac{Pr(Z = 1|\mathbf{X} = \boldsymbol{x})}{1 - Pr(Z = 1|\mathbf{X} = \boldsymbol{x})} = \boldsymbol{x}^t\boldsymbol{\beta}, \tag{3}$$

and then estimate the propensity score $e(\mathbf{X})$ as

$$\hat{e}(\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}^t\hat{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{x}^t\hat{\boldsymbol{\beta}})}. \tag{4}$$

Selecting a propensity score model is a key component in the propensity score methodology. In practice, the model selection process may need to be carried out in an iterative fashion before satisfactory balance can be achieved in all (or most) of the important covariates. In the meantime, substantial input from subject scientists can help to facilitate the model selection process.

In our method, the model selection is done using the original data. We assume that the propensity score model (3) estimated from the original sample is a proxy of the true propensity model. Under this assumption, the ensuing bootstrap iterations simply refit (3) using the resampled data without altering the model. For discussions on the issue of model selection in bootstrap settings, see Sauerbrei and Schumacher (1992), and Sauerbrei (1999).

Using the estimated propensity scores we stratify all the subjects into K subclasses so that the estimated propensity scores have similar values within each subclass. We assume there are N_{tk} treated subjects, $Y_{tk1},\ldots,Y_{tkN_{tk}}$, and N_{ck} control subjects, $Y_{ck1},\ldots,Y_{ckN_{ck}}$, in the kth subclass, $k=1,\ldots,K$. We let $\bar{Y}_{tk}=\sum_{i=1}^{N_{tk}}Y_{tki}/N_{tk}$ and $\bar{Y}_{ck}=\sum_{i=1}^{N_{ck}}Y_{cki}/N_{ck}$ be the mean responses for the treated and the control subjects in the kth subclass, and $n_t=\sum_{k=1}^{K}N_{tk}$ and $n_c=\sum_{k=1}^{K}N_{ck}$ be the total numbers of treated and control subjects in the entire experiment, respectively. We note that while n_t and n_c are considered as fixed in a given experiment, N_{tk} and N_{ck} are random quantities that depend on the estimated propensity scores.

From (2) we see that in order to estimate δ we need to estimate the fraction of propensity scores in each subclass, $P(e(\mathbf{X})\in I_k)$, and the expected values of the outcome in the two treatment groups within each subclass, $E(Y(1)\mid e(\mathbf{X})\in I_k, Z=1)$ and $E(Y(0)\mid e(\mathbf{X})\in I_k, Z=0)$. Using $(N_{tk}+N_{ck})/(n_t+n_c)$, $\bar{Y}_{tk}$, and $\bar{Y}_{ck}$ to estimate these three respective components, we have the following estimate for δ:

$$\hat{\delta}=\sum_{k=1}^{K}\frac{N_{tk}+N_{ck}}{n_t+n_c}\left(\bar{Y}_{tk}-\bar{Y}_{ck}\right). \tag{5}$$

The estimate $\hat{\delta}$ given by (5) uses cut-off points determined by the quantiles of the estimated propensity scores of the combined treatment groups.

D'Agostino (1998) gives an excellent and detailed survey on the different stratification strategies. The investigator must also decide K, the number of subclasses used in the analysis. A useful guideline on the selection of K can be found in the Appendix A of Rosenbaum and Rubin (1984), where they stated that five subclasses based on the propensity score would remove over 90% of the bias due to unbalanced covariates.

As in the analysis of stratified random samples, the variance of the average treatment effect estimate (5) is often calculated using formula

$$\mathrm{Var}(\hat{\delta}) = \sum_{i=1}^{K} \left(\frac{N_{tk} + N_{ck}}{n_t + n_c} \right)^2 \mathrm{Var}(\bar{Y}_{tk} - \bar{Y}_{ck}). \tag{6}$$

The inferences about the true treatment effect are then made by assuming that $\hat{\delta}$ follows a normal distribution with the variance given by (6), as one would do in the analyses of stratified random samples (Perkins et al. 2000).

The validity of such inferences hinges not only on the assumption of the normality of the point estimator $\hat{\delta}$, but also on several rather implicit assumptions: 1) the cut points of the subclasses are fixed; 2) the responses are independent across all the subclasses; and 3) within each subclass, the responses from the treated and control subjects are independent. In his Ph.D thesis, Du (1998) has shown that these assumptions are not true because the subclassification is based on orders of estimated propensity scores and hence introduces an order statistics structure into the problem. Consequently, the resulting subclassification destroys the original independent data structure (both within and between subclasses) and the variance estimate in (6) becomes incorrect. Tu et al. (1999) further showed that even if the independent structure were maintained, the variance formula for stratified random sample should not be used in propensity score based inferences because it also failed to account for the uncertainty associated with the estimation of the propensity scores.

Therefore, in order to achieve correct inferences, one must be able to find a method that accommodates the dependent structure caused by the subclassification, and to account for the additional source of variation in the propensity score estimation.

3 A BCa confidence interval for the treatment effect

Attempts have been made by several authors to alleviate the aforementioned deficiencies in the current practice. For example, Tu et al. (1999) discussed

the construction of a fully parametric likelihood ratio test for the treatment effect under an alternative stratification scheme with random subclass sizes. One of the limitations of the likelihood ratio test is the multivariate normal assumption on the covariate vector. Since many observational studies have categorical explanatory variables, the assumption appears to be too restrictive in most applications. In this research, we consider a bootstrap confidence interval for the average treatment effect that is conceptually straightforward and relatively easy to implement.

Several bootstrap confidence interval procedures have been proposed in the past two decades (Efron 1985, Efron and Tibshrani 1986, Hall 1992). Among the methods discussed in the literature, percentile-t and bias-corrected accelerated (BCa) bootstrap methods have been shown to possess better accuracy. Since the percentile-t method requires a variance measure for the estimated treatment effect that is not readily available in the propensity score subclassification situation, we focus on the BCa approach. Efron (1987) has shown that the BCa bootstrap interval has three highly desirable properties: (1) it is of second order accuracy, (2) it is transformation-respecting, and (3) it is range-preserving. The original one-sample BCa bootstrap method, however, is not directly applicable here because the causal inference on a treatment effect is in essence a two-sample problem. In an extension to Efron's one-sample BCa procedure, Hall and Martin (1988) described a two-sample BCa procedure that is both second order accurate and second order correct, as defined by Efron and Tibshirani (1993). In this research, we apply Hall and Martin's (1988) procedure to the propensity score subclassification situation.

To better describe the proposed bootstrap procedure, we re-introduce the notation with the subscription reserved for the subject: Let $n = n_t + n_c$ be the total number of subjects; $(Y_i, \mathbf{X}_i, Z_i)$ be a vector containing the response variable, the covariate vector for the true propensity model, and the treatment assignment for the ith subject, where $i = 1, \ldots, n$.

Our procedure starts with the fitting of propensity model (3) using the original sample, $(\mathbf{X}_i, Z_i)$ for $i = 1, \ldots, n$. After fitting the model, we estimate the propensity score $(\hat{e}_i)$ of each subject using (4). Based on these estimates, we partition the subjects into K homogeneous subclasses. The post-stratification balance of each covariate is then examined, a point estimate for the average treatment effect $(\hat{\delta})$ is obtained using (5).

For each bootstrap iteration b, $b = 1, \ldots, B$, we resample with replacement n_t treated and n_c control subjects separately from the treated and control subjects in the original sample. Let $(Y_{i'}^{(b)}, \mathbf{X}_{i'}^{(b)}, Z_{i'}^{(b)})$ be the bth

bootstrap, $i' = 1, \ldots, n = n_t + n_c$. Using the resampled data $(\mathbf{X}_{i'}^{(b)}, Z_{i'}^{(b)})$, for $i = 1, \ldots, n$, we re-fit the same logistic model and re-estimate the propensity score for each of the resample subjects, $\hat{e}_{i'}^{(b)}$. Then we stratify the bootstrapped responses $Y_{i'}^{(b)}$, compute the mean treatment effect, and denote it as $\hat{\delta}^b$.

After repeating this process a large number (B) of times, we compute the $100(1-2\alpha)\%$ BCa confidence interval for the average treatment effect δ following Hall and Martin's (1988) procedure. The bias-correction constant is computed as

$$\hat{d} = \Phi^{-1}\left(\frac{\#[\hat{\delta} < \hat{\delta}^b]}{B}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. The two-sample acceleration parameter is computed as

$$\hat{a} = \frac{1}{6}\hat{\sigma}^{-3/2}(n_t^{-2}\hat{\gamma}_t - n_c^{-2}\hat{\gamma}_c),$$

where $\hat{\sigma}^2 = \hat{\sigma}^2_{tjack} n_t^{-1} + \hat{\sigma}^2_{cjack} n_c^{-1}$; here we use the jackknife variance estimates from the original sample $\hat{\sigma}^2_{tjack}$ and $\hat{\sigma}^2_{cjack}$ to estimate the unknown variances of the two treatment groups, as suggested by Efron and Tibshirani (1993); $\hat{\gamma}_t$ and $\hat{\gamma}_c$ are the sample skewnesses of the respective groups.

Sorting the bootstrap treatment effect estimates into increasing order, $\hat{\delta}^{(1)} \leq \cdots \leq \hat{\delta}^{(B)}$, the resulting $100(1-2\alpha)\%$ confidence interval is defined as

$$\left(\hat{\delta}^{[B(\hat{\beta}_{a(\alpha)})]}, \hat{\delta}^{[B(\hat{\beta}_{a(1-\alpha)})]}\right),$$

where $\hat{\beta}_{a(\alpha)} = \Phi\{\hat{d} + (\hat{d} + z_\alpha)\{1 - \hat{a}(\hat{d} + z_\alpha)\}^{-1}\}$ and $[x]$ is the largest integer less than or equals to x.

4 A Simulation Study

To assess the finite sample performance of the proposed procedure, we conducted a simulation study. We focus primarily on the coverage of the bootstrap confidence intervals by reporting the empirical coverage probabilities of the proposed procedure under a set of pre-determined parameter settings.

For each configuration, we first generate the covariates $\mathbf{X}$. In our simulation, we consider a logistic regression propensity model with three covariates: a continuous covariate X_1 and two binary covariates X_2 and X_3.

In order to simulate situations where the covariate distributions differ systematically in the two treatment groups, we generate the covariate deviates for the treatment and control groups separately: For the control group ($Z = 0$), we assume that $X_1 \sim \mathrm{N}(0, \sigma_1^2)$, $X_2 \sim \mathrm{Bernoulli}(p_{2c})$, and $X_3 \sim \mathrm{Bernoulli}(p_{3c})$; for the intervention group ($Z = 1$), we assume that $X_1 \sim \mathrm{N}(d, \sigma_1^2)$, $X_2 \sim \mathrm{Bernoulli}(p_{2t})$, and $X_3 \sim \mathrm{Bernoulli}(p_{3t})$. By controlling d and the probabilities p_{2c}, p_{2t}, p_{3c}, and p_{3t} in the Bernoulli distributions, we will be able to simulate situations with varying level of differential covariate distributions.

With the pseudo-random covariate deviates $\mathbf{X}$ and the treatment assignment Z, we then generate responses Y from a linear relationship $Y = Z\delta + \mathbf{X}^t\boldsymbol{\beta} + \epsilon$ using pre-specified values of δ, $\boldsymbol{\beta}$ and the independently generated normal errors $\epsilon \sim \mathrm{N}(0, \sigma_\epsilon^2)$. Herein, δ represents the true treatment effect. Since the covariates $\mathbf{X}$ have different distributions in the two treatment groups, the effect of the treatment (δ) can not be directly estimated from the response Y without first adjusting for the effects of $\mathbf{X}$. For simplicity, we set the coefficients of the covariates to be $(\beta_1, \beta_2, \beta_3) = (0.5, 0.4, 0.4)$ throughout the simulation; the values of the rest of the parameters that we used in the simulation are listed in Table 1. It should be noted that a number of factors contribute to the extend of confounding effects of $\mathbf{X}$ on Y, including: 1) magnitudes of $\beta_1, \beta_2, \beta_3$, which affect the level of confounding directly; and 2) the differential distributions of $\mathbf{X}$ in the two treatment groups, which have an indirect effect on the level of confounding.

To evaluate the coverage probability of the proposed BCa Bootstrap confidence interval procedure based on the propensity score subclassification, we apply the proposed procedure to the simulated data. For each parameter configuration, we conduct 1000 simulations in an iterative fashion. Within each iteration, we use 2000 bootstraps to construct a 95% confidence interval. The empirical coverage probabilities under different parameter configurations are reported for the assessment of the performance of the proposed procedure. To understand how the coverage property changes in various sample size situations, we consider three sample sizes, $n_C = n_t = 500$, 1000, and 2000, for each of the parameter settings in Table 1. The simulation results are reported in Table 2.

The preliminary simulation results indicate reasonably good coverage probabilities in the proposed procedure. This coverage appears to be better when the sample sizes are larger. Further observations of the simulation study are discussed in Section 6.

5 Data Example

We now illustrate the procedure described in Section 3 with a pharmacoepidemiological example.

Non-steroidal anti-inflammatory drugs (NSAIDs) are widely prescribed in the US. One potential adverse effect related to the long term use of these drugs is the risk of renal insufficiency. Several studies have indicated that this risk may vary for different NSAIDs. For example, Bunning and Barth (1982) and Ciabatonni, Cinotti and Pierucci (1984) reported that Sulindac, one of the prescription NSAIDs, may be "renal sparing", comparing to some of the over-the-counter NSAIDs, such as Ibuprofen. Clinical evidence on the renal sparing effect of Sulindac, however, is rather mixed (Murray and Brater 1993, Murray, Greene and Kuzmik 1995). In an ideal situation, a randomized clinical trial would have provided a more definitive answer. But due to ethical concerns, it is often not feasible for the physicians to prescribe a non-renal sparing drug to patients with known histories of renal problems or other serious diseases. Therefore, the investigation can only be carried out as an observational study (Perkins et al. 2000). In this article, it is not our intention to address the clinical question on the renal sparing effect of Sulindac. Instead, we focus on the methodological issues associated with the causal inferences in observational studies. In particular, we will use the NSAIDs data to illustrate the bootstrap confidence interval procedure described in Section 3. Herein, we restrict our attention to the comparison of the renal effects of two popular NSAIDs, Ibuprofen and Sulindac. The renal effects of NSAIDs are measured by the differences of serum creatinine concentrations taken pre- and post-treatments. The goal of analysis is to compare these differences, adjusting for potential confounders.

The data are extracted from the Regenstrief Medical Record System (McDonals, Overhage and Tierney 1999). For the purpose of illustration, we only use the records of 1946 patients with complete medical records. Among them, 1694 had Ibuprofen and 252 had Sulindac in the study period. The outcome of interest is the change in serum creatinine concentrations before and after the use of the NSAIDs. The data set also contains 31 explanatory variables providing relevant demographic and clinical information of the study subjects. These variables are tabulated in Table 3. Careful examination of the explanatory variables reveals the apparent covariate imbalance between the two treatment groups. For example, the average age of the study subjects in the Sulindac group is 70.714, while that of the Ibuprofen group is only 57.824, with a difference of 12.89 years in the mean between the two groups. To formally check whether observed differences in

the explanatory variables between the two treatment groups are statistically significant, we use t-tests for continuous variables and chi-square tests for discrete variables. The p values of the tests are reported in Table 3, which show significant imbalance in 18 of the 31 covariates at $\alpha = 0.05$ significance level. Generally, the Sulindac patients are older and have poorer health conditions than the Ibuprofen patients. For example, 20.24% of the Sulindac users suffered from chronic heart failures, while only 10.27% of the Ibuprofen users suffered from the same condition. If left unadjusted, these unbalanced covariates could confound with the true treatment effect and lead to a bias in the point estimation.

– Insert Table 3 here –

To adjust for the covariates, we first use the information provided by the 31 explanatory variables to fit a logistic regression model for the probability of a study subject receiving Sulindac. Following the common practice of using 5 subclasses, we then stratify the entire data set according to the quintiles of the estimated propensity scores. In our data example, propensity score subclassification greatly improves the balance of the covariates.

The post-stratification balance of the continuous covariates are re-examined using a two-way analysis variance model with propensity score quintiles and the treatment assignment as the main effects, the interaction between the two main effects is also included in the model. For the binary covariates, we fit logistic models with the same main effects and interaction, as suggested by Rosenbaum and Rubin (1984), and D'Agostino (1998). The re-examination reveals a great improvement in the balance of covariates within each subclass. Of the 31 covariates considered in the model, only 4 still show significant imbalances (marked by $*$ in Table 2), comparing to 18 covariates before the subclassification. Further examining the data, we found that even though some of the imbalances were still statistically significant as our t tests and chi-square tests had suggested, the magnitudes of the differences were greatly reduced. For example, the covariate "Age" still tested significant. But the magnitude of differences in age of the two treatment groups within the subclasses were all less than 3 years, as comparing to the 12.89 years before the subclassification. It is our opinion that though such imbalances were statistically significant, they may no longer represent any clinically meaningful differences between the two groups. Figure 1 shows the balance of the covariate "Age" for all 5 subclasses after the stratification.

– Insert Figure 1 here –

Applying the proposed bootstrap procedure described in the previous section based on 2000 bootstrap iterations, we obtain a 95% confidence interval of $(-0.2453, 0.2752)$ for the difference in the serum creatinine changes

between the Ibuprofen and Sulindac treated patients. This interval is slightly wider than the currently used 95% normal confidence interval of $(-0.2122, 0.2413)$ based on the the variance estimate given in (6). Though both intervals point to the same conclusion that the renal effects of the two NSAIDs are not significantly different, the increased interval length in the new procedure may suggest the need for a variance adjustment by the bootstrap, which could lead to an improvement in the coverage probability.

6 Discussion

Since Rubin and Rosenbaum's early groundbreaking work, propensity score methodology has been successfully applied to many clinical and epidemiological studies. It has become a widely used tool for reducing the potential bias in treatment effect estimation in observational data analysis. In this paper, we have proposed a bootstrap based inference procedure for the treatment effect within the framework of propensity score subclassification.

Our study suggests that the proposed method provides valid causal inferences in large observational studies. In summary, it has several advantages over the existing methods: First, it does not require a variance estimate. Our experience indicates that a direct analytical derivation of the variance of the treatment effect estimate in a general situation is difficult, if not entirely impossible. Secondly, the method does not rely on any restrictive distributional assumption on the covariates. This is particularly important in practice because one rarely has all normally distributed explanatory variables. Thirdly, the bootstrap interval accounts for the variation that arises from the estimation of propensity scores, and it accommodates the dependency among the responses both within and between subclasses due to the ordering structure introduced by the subclassification. Finally, the new bootstrap procedure can be implemented relatively easily in most computing platforms.

Our simulation shows that the empirical coverage of the procedure are reasonably good. While the empirical coverage of the probabilities are below the nominal level (95%), they are generally close to the nominal level when the sample sizes are greater than 1000 per group. The simulation results also show that the coverage probability the BCa confidence interval is not affected by the size of the treatment effect δ. This should not be a surprise because in our simulation scheme, the size of the treatment effect is an additive component in a linear model; when the responses are generated from this model, δ simply represents a shift in the central locations between the two treatment groups. The simulation also shows that the proposed method

adjusts for the effects of the systematically different covariates quite effectively, when all of the covariates are used in the logistic regression model to estimate the unknown propensity scores (ie, strongly ignorable assumption holds).

Although the preliminary simulation results are promising, a more extensive simulation study is apparently needed to establish the operating characteristics of the proposed procedure under various practical data situations. In this respect, the current simulation has several limitations: First, it considers only balanced designs while most observational data have unbalanced group sizes. For example, the NSAIDs data that we used to illustrate our method have a rather substantial imbalance between the Ibuprofen group ($n_I = 1694$) and the Sulindac group ($n_S = 252$). Second, the range of values of the parameters used in the current simulation is still limited. For example, only one set of $\boldsymbol{\beta}$ values in linear relationship $Y = Z\delta + \mathbf{X}^t\boldsymbol{\beta} + \epsilon$ were used to generate random responses. Since $\boldsymbol{\beta}$ directly affects the level of confounding between the observed covariates $\mathbf{X}$ and the treatment assignment Z, it would be of interested to examine the performance of the proposed procedure under many different $\boldsymbol{\beta}$ values. Holding other parameters constant, a smaller value $\boldsymbol{\beta}$ decreases the level of confounding (in the most extreme case of $\boldsymbol{\beta} = \mathbf{0}$, we have $Y = Z\delta + \epsilon$, implying no confounding effect from $\mathbf{X}$.) In addition, for a set of pre-selected $\boldsymbol{\beta}$ values, parameters d, p_{2c}, p_{3c}, p_{2t}, and p_{3t} control the differential distributions between the treatment groups. The magnitude of d, and the difference between p_{2c} and p_{2t} (or that between p_{3c} and p_{3t}) reflect the separation in covariate distribution between the two treatment groups. In the current simulation, only one d value, and a limited number of binomial probabilities were considered. In light of these observations, we feel that further investigation is certainly needed in order to have a fuller understanding of the new method's operating characteristics. The authors plan to expand the simulation study to include: 1) unbalanced designs by altering n_t and n_c; 2) different levels of confounding by adjusting $\boldsymbol{\beta}$ values; and 3) different magnitudes of separation in covariate distributions between the two groups by controlling d, p_{2c}, p_{3c}, p_{2t}, and p_{3t}. We will also investigate the sensitivity of our procedure when the propensity model is mis-specified.

Our current work focuses on a resampling based approach for the construction of a simple confidence interval of an unknown treatment effect. Several related issues have yet to be explored. For example, the adjustment of covariates in treatment effect estimates via regression (instead of subclassification) has not been studied. Treatment effects summarized by other measures, such as odds ratios, also need to be discussed. These issues will

be at the center of our future exploration.

Acknowledgment

The authors' work was partially funded by NIH R01MH58875 and a research award from the Drug Information Association. They wish to thank the two reviewers for their constructive comments. Thanks also goes to Professor Donald Rubin for drawing the authors' attention to the work of Dr. Jiangtao Du. Finally, the authors are grateful to Dr. Michael D. Murray for providing the NSAIDs data.

References

Bunning, R. and Barth, W. (1982). Sulindac: A potentially renal-sparing nonsteroidal anti-immflammatory drug, *Journal of the American Medical Association* **248**: 2864–2867.

Ciabatonni, G., Cinotti, G. and Pierucci, A. (1984). Effects of sulindac and ibuprofen in patientcs with chronic glomerrular disease: Evidence for the dependence of renal function on prostacylin, *New England Journal of Medicine* **310**: 279–283.

D'Agostino, R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group, *Statistics in Medicine* **17**: 2265–2281.

Du, J. (1998). *Valid Inferences after propensity score subclassification using maximum number of subclasses as building blocks*, PhD thesis, Harvard University, Cambridge, Massachusetts.

Efron, B. (1985). Bootstrap confidence interval for a class of parametric problems, *Biometirka* **72**: 45–58.

Efron, B. (1987). Better bootstrap confidence intervals, *Journal of the American Statistical Association* **82**: 171–200.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, New York.

Efron, B. and Tibshrani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, *Statistical Science* **1**: 54–75.

Greenland, S., Robins, J. M. and Pearl, J. (1999). Confounding and collapsibility in causal inference, *Statistical Science* **14**: 29–46.

Hall, P. (1992). *The Bootstrap and Edgeworth Expension*, Springer-Verlag, New York.

Hall, P. and Martin, M. (1988). On the bootstrap and two-sample problems, *Australian Journal of Statistics* **30A**: 179–192.

McDonals, C., Overhage, J. and Tierney, W. (1999). The Regenstrief Medical Record System: A quarter century experience, *Int J Med Inf* **54**: 225–253.

Murray, M. and Brater, D. C. (1993). Renal toxicity of nonseroidal antiimflammatory drugs, *Annual Reviews of Pharmacological Toxicology* **32**: 435–465.

Murray, M. D., Greene, P. K. and Kuzmik, D. (1995). Effects of nonsteroidal anti-inflammatory drugs on glomerular filtration rate in elderly patients without and with renal insufficiency, *American Journal of Medical Sciences* **310**: 188–197.

Perkins, S. M., Tu, W., Underhill, M. G., Zhou, X.-H. and Murray, M. D. (2000). The use of propensity scores in pharmacoepidemiologic research, *Pharmacoepidemiology and Drug Safety* **9**: 93–101.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score, *Journal of the American Statistical Association* **79**: 516–524.

Rosenbaum, P. and Rubin, D. B. (1983). Central role of the propensity score in observational studies for causal effects, *Biometrika* **70**: 41–55.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies, *Journal of the American Statistical Association* **74**: 318–328.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores, *Annals of Internal Medicine* **127**(8): 757–763.

Sauerbrei, W. (1999). The use of resampling methods to simplify regression modelss in medical statistics, *Journal of the Royal Statistical Society Series C - Applied Statistics* **48**(3): 313–329.

Sauerbrei, W. and Schumacher, M. (1992). A bootstrap resampling procedure for model building - application to the cox regression model, *Statistics in Medicine* **11**(16): 2093–2109.

Stone, R. A., Obrosky, D. S. and Singer, D. E. (1995). Propensity score adjustment for pretreatment differences between hospitalized and ambulatory patients with community-acquired pneumonia, *Medical Care* **33**(4): 56–66.

Tu, W., Perkins, S. M., Zhou, X.-H. and Murray, M. D. (1999). Testing treatment effect using propensity score stratification, *1999 Proceedings of Section on Statistics in Epidemiology of the American Statistical Association*, American Statistical Association, pp. 105–107.

Table 1: Parameter Settings for Simulation

Setting	δ	d	σ_1	σ_ϵ	p_{2c}	p_{3c}	p_{2t}	p_{3t}
1	0.0	0.5	1.0	1.0	0.3	0.7	0.8	0.4
2	0.5	0.5	1.0	1.0	0.3	0.7	0.8	0.4
3	1.0	0.5	1.0	1.0	0.3	0.7	0.8	0.4
4	2.0	0.5	1.0	1.0	0.3	0.7	0.8	0.4
5	0.0	0.5	1.25	1.0	0.4	0.8	0.7	0.3
6	0.5	0.5	1.25	1.0	0.4	0.8	0.7	0.3
7	1.0	0.5	1.25	1.0	0.4	0.8	0.7	0.3
8	2.0	0.5	1.25	1.0	0.4	0.8	0.7	0.3

Table 2: Coverage Probabilities of the 95% BCa Confidence Intervals

Parameter Setting	Sample size $n_t = n_c$	Coverage Probabilities
1	500	0.911
	1000	0.931
	2000	0.945
2	500	0.910
	1000	0.933
	2000	0.947
3	500	0.915
	1000	0.935
	2000	0.945
4	500	0.912
	1000	0.934
	2000	0.947
5	500	0.908
	1000	0.935
	2000	0.943
6	500	0.912
	1000	0.938
	2000	0.946
7	500	0.909
	1000	0.935
	2000	0.945
8	500	0.908
	1000	0.936
	2000	0.948

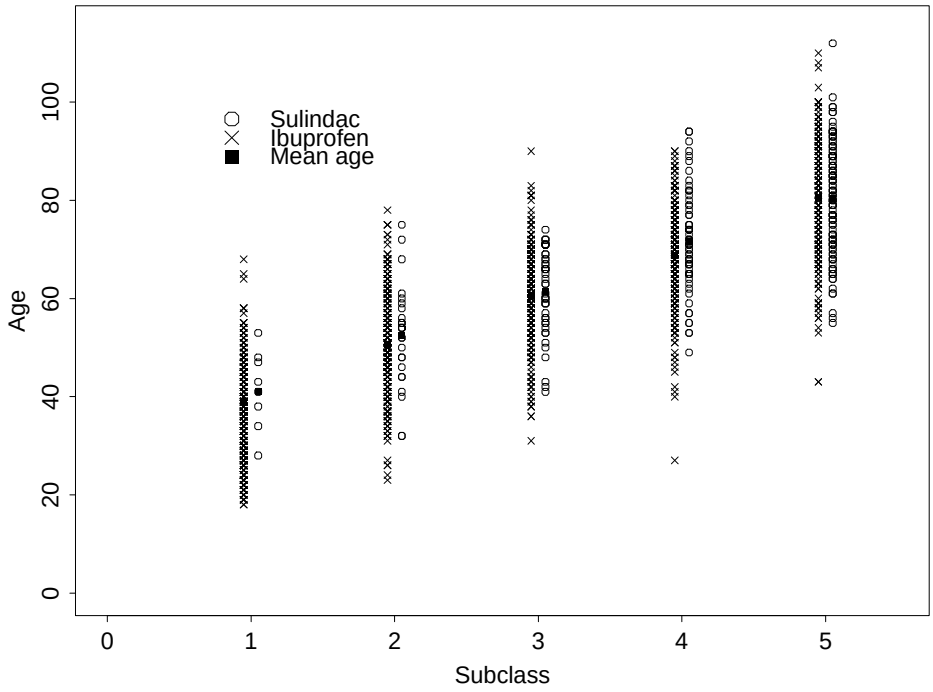

Figure 1: Balance within subclasses for Age

Table 3: Explanatory Variables in the NSAIDs Data

	Ibuprofen ($n = 1694$)	Sulindac ($n = 252$)	Unadjusted p value
Patient demographics			
Age (years)	57.824	70.714	<0.0001∗
Sex (% male)	25.86	31.75	0.0484
Race (% white)	40.38	37.30	0.3523
Clinical Variables			
Prior average systolic BP (mmHg)	131.5	138.22	<0.0001
Last systolic BP prior to NSAIDs	130.94	136.90	<0.0001
Prior average serum potassium (mEq/l)	4.0500	4.0511	0.9693
Last serum potassium prior to NSAIDs	4.0611	4.0599	0.9702
Prior average weight (lb)	179.71	183.00	0.3382
Last weight prior to NSAIDs	180.08	182.48	0.4920
Last SCC prior to NSAIDs (mg/dl)	0.9747	1.0226	0.0149
Disease (% of patient having the disease)			
Arrhythmia	6.91	16.67	<0.0001∗
Ascities	1.30	1.19	0.8868
Asthma	3.96	1.98	0.1220
CAD	13.22	18.65	0.0202
CHF	10.27	20.24	<0.0001∗
Cirrhosis	9.21	5.56	0.0429
COPD	8.97	13.49	0.0228
Diabetes	29.40	38.49	0.0035
Hypertension	73.49	87.70	<0.0001
Liver diseases	0.71	1.59	0.1494
Myocardial infraction	5.84	9.13	0.0449
Osteoarthritis	8.97	17.06	<0.0001∗
Rheumatoid arthritis	0.41	1.98	0.0030
Stroke	4.55	5.56	0.4786
ACE inhibitors	10.28	9.13	0.5741
Beta adrenergic antagonists	14.64	26.59	<0.0001
Blood pressure medications	53.96	73.03	<0.0001
Calcium channel antagonists	9.09	9.52	0.8240
Diuretic	45.99	67.46	<0.0001
Insulin	12.04	15.87	0.0868
Oral hypoglycemics	8.56	9.52	0.6121

Asterisk (*) indicates significant imbalance after propensity score stratification.