Estimation of the average causal effect via multiple propensity score stratification

Takanobu Nomura[1,2] and Satoshi Hattori[3]

[1] Graduate School of Medicine, Kurume University, Fukuoka, Japan

[2] Medical Affairs, Kyowa Hakko Kirin, Co. Ltd., Tokyo, Japan

[3] Department of Biomedical Statistics, Graduate School of Medicine, Osaka University, Osaka, Japan

Running title: Multiple propensity score stratification

Word count for abstract: 100

Word count for text (excluding references, figures and tables): 5232

Corresponding author:

Satoshi Hattori

Department of Biomedical Statistics, Graduate School of Medicine, Osaka University,

2-2 Yamadaoka, Suita-city, Osaka 565-0871, Japan.

Phone: +81-942-31-7835

Fax: +81-942-31-7865

E-mail: hattoris@biostat.med.osaka-u.ac.jp

Abstract

Suppose we are interested in estimating the average causal effect (ACE) for the population mean from observational study. Because of simplicity and ease of interpretation, stratification by a propensity score (PS) is widely used to adjust for influence of confounding factors in estimation of the ACE. Appropriateness of the estimation by the PS stratification relies on correct specification of the PS. We propose an estimator based on stratification with multiple PS models by clustering techniques instead of model selection. If one of them correctly specifies, the proposed estimator removes bias and thus is more robust than the standard PS stratification.

Key words

Propensity score; Stratification; Clustering; Multiple robustness; Confounding; Model misspecification;

1 Introduction

In order to draw inferences about the treatment effect from observational studies, it is a crucial issue how to control the effects of confounding factors. Rubin's causal model is a widely accepted framework for causal inference from observational studies (Rosenbaum et al., 1983). We follow this framework. Under the strongly ignorable treatment assignment (SITA) assumption, various methods have been proposed to estimate the average causal effect. Outcome regression describing the relationship between the outcome and covariates is one of the standard methods to control confounding. Alternatively, one can use the methods based on the propensity score (PS) proposed by Rosenbaum and Rubin (1983), including stratification, matching and regression. The inverse probability weighted estimator (IPW) has an advantage in that it does not suffer from residual confounding, whereas the stratified analysis does (Rosenbaum, 1987). Recent theoretical advances have been made through the inverse weighting: the PS is incorporated into the doubly robust estimator, which has desirable properties of robustness and efficiencies (Lunceford et al., 2004; Bang and Robins, 2005). Namely, if at least one of the propensity score model or the outcome regression is correctly specified, it is consistent and when both of the propensity score model and the outcome regression are correctly specified, the doubly robust estimator is more efficient than the IPW estimator.

On the other hand, the PS stratification is also widely applied in practice (Barker et al., 1988; Coyte et al., 2000; Bateman et al., 2013). One advantage of the stratified analysis is ease of interpreting the results of statistical analysis, in particular, for non-statisticians. The idea behind the PS stratification is very simple. First, subjects are classified by a PS into several strata, within which no factors are confounding due to a balancing property of the PS, next, the stratum-specific treatment effects are estimated using the

difference of simple sample means between two groups, and finally all the stratum-specific estimators are combined across strata. In theory, the number of strata must be large enough to ensure that the values of the propensity score are close within a stratum. However, a relatively small number of strata e.g., 5 is often employed since most biases can be removed when relatively few strata are involved.

In practice, the PS is unknown and thus must be estimated. Estimation of ACE by misspecified PS may have a serious bias (Drake, 1993; Kang et al., 2007). As argued by Kang and Schafer (2007), both the IPW and the doubly robust estimator may provide seriously biased or unstable estimates with highly variable PSs. The stratified estimator by a PS is likely to be stable even with highly variable PS, and is robust against misspecification of the link function in modeling the PS by the generalized linear model (Williamson, 2012; Drake, 1993).

In this paper, we focus on estimators based on stratification. To reduce uncertainty in modeling of the PS, some model selection procedures are often employed. Instead, we propose conducting stratification accounting for multiple PS models simultaneously. In principle, one can easily construct such strata. For example, if we have three PS models and stratify subjects into 5 strata with respect to each PS, then 5^3=125 strata are created. However, some of them may be empty or have only a small number of subjects and estimator with the stratification may be unstable. We propose to apply clustering techniques in order to classify subjects efficiently into relatively small number of strata, within which each of the PSs is homogeneous.

The rest of this manuscript is organized as follows. In Section 2, we introduce our proposed method. In Section 3, we examine the performance of our proposed method, comparing it with several alternatives; the stratified estimator by a propensity score, that by the propensity score selected by BIC, that by the model-averaged propensity score, that by a clustering method applied to a vector of covariates directly. Furthermore, comparison with the inverse probability treatment weighting estimator and the doubly robust estimator were also performed. In Section4, our proposal is illustrated with a dataset from the Tone study, which is a community survey conducted in Japan.

2 Methods

2.1 Preliminaries

Suppose we are interested in estimating the average causal effect (ACE) by comparing treatment and control groups in an observational study. Let Z be an indicator of treatment allocation (Z = 1 for the treatment group and Z = 0 for the control group) and X denote a p-dimensional covariate vector. Each subject has a pair of potential outcomes $\left(Y^{(1)}, Y^{(0)}\right)$, where, $Y^{(r)}$, r = 0, 1 is the outcome of the subject that would be observed if he or she were assigned to $Z = r$ (r = 0, 1). Suppose we have observations from n subjects. Let $\left(Y_i^{(1)}, Y_i^{(0)}, Z_i, X_i\right)$, i = 1, 2, ..., n be n i.i.d copies of $\left(Y^{(1)}, Y^{(0)}, Z, X\right)$, where the subscript i implies the i-th subject. As a fundamental problem in causal inference, only one of $Y^{(1)}$ and $Y^{(0)}$

would be observed for each subject and the observed outcome would be denoted by $Y = ZY^{(1)} + (1 - Z)Y^{(0)}$. We have observations (Y_i, Z_i, X_i), i = 1, 2, ..., n. The ACE is defined as $\Delta = \mu^{(1)} - \mu^{(0)}$, where $\mu^{(r)} = \mathrm{E}(Y^{(r)})$ for r = 0, 1. To draw an inference for the ACE from an observational study, we assume the SITA condition (Rosenbaum and Rubin, 1983), $(Y^{(1)}, Y^{(0)}) \perp Z|X$, where, for arbitrary random variables A_1, A_2 and A_3, $A_1 \perp A_2|A_3$ implies that A_1 is conditionally independent of A_2 given A_3. The PS is defined as $\mathrm{P}(Z = 1|X)$ as a device for controlling confounding under the SITA condition, which satisfies $0 < \mathrm{P}(Z = 1|X) < 1$.

2.2 Stratified estimator via multiple propensity scores

We begin with summarizing the idea of the standard stratified estimator (Rosenbaum and Rubin, 1984). Suppose subjects are stratified into S strata in accord with the value of the PS. The number of subjects in the s-th stratum is denoted by n_s. Let $\bar{Y}_s^{(r)}$ denote the sample means of the observed outcomes for the subjects assigned to $Z = r$ in the s-th stratum, and $\hat{\Delta}_s = \bar{Y}_s^{(1)} - \bar{Y}_s^{(0)}$. The stratified estimator by the PS is defined by $\hat{\Delta}^{STR} = \sum_{s=1}^{S} n_s \hat{\Delta}_s / n$. If the PS is common within each stratum, $\hat{\Delta}^{STR}$ consistently estimates the ACE Δ. In practice, the PS is unknown and is estimated by regression models such as logistic regression to this end. We suppose the PS is estimated by a regression model and the estimated propensity score is denoted by $\hat{\mathrm{e}}(X)$. Define a sequence $0 = C_0 < C_1 < \cdots < C_S = 1$. A subject is classified into the s-th stratum if the PS satisfies $\hat{\mathrm{e}}(X) \in (C_{S-1}, C_S]$. In principle, if the stratification boundaries $0 = C_0 < C_1 < \cdots < C_S = 1$ are taken sufficiently precisely, the PSs of the subjects in each stratum are expected to be close. However, too-precise stratification boundaries produce unstable stratum-specific estimates and may lead to a biased estimate of the ACE by construct of strata with a small number of samples. In practice, the stratified estimator by the PS removes more than 90 percent bias even with a relatively small number of strata, for example 5 (Rosenbaum and Rubin, 1984; Williamson, 2012).

Misspecification in modeling of the PS may lead to seriously biased estimates (Drake, 1993; Kang and Schafer, 2007). To avoid misspecification, one can prepare several candidates of the PS model and can select the best by a model selection procedure such as Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) (Claeskens and Hjort, 2008). Instead, we sought to construct strata, within each of which all the estimated PSs were close. Suppose we have K candidates of PS models. The PS of the i-th subject based on the k-th PS model is denoted by $\hat{\mathrm{e}}^{(k)}(X_i)$, k = 1, 2, ..., K, and define $\boldsymbol{f_i} = (\hat{\boldsymbol{e}}^{(1)}(\boldsymbol{X_i}), \hat{\boldsymbol{e}}^{(2)}(\boldsymbol{X_i}), \ldots, \hat{\boldsymbol{e}}^{(\mathbf{K})}(\boldsymbol{X_i}))$. We propose to construct strata by applying a hierarchical clustering method to the vector $\boldsymbol{f_i}$. Clustering techniques classify subjects into several clusters (strata) and many clustering algorithms have been proposed (Gan et al., 2007). Although any clustering algorithm may be applied, we employ Ward's minimum variance method, which is one of hierarchical clustering methods (Ward, 1963). Ward's minimum variance method performs cluster integration which attains the minimum increase of the cluster sum of squares that indicate the total amount of distance for each object from

center. See among others for details of clustering techniques (Gan et al., 2007; Mirkin, 2012). Let C be the number of clusters. If we select sufficiently large C, $\boldsymbol{f_i}$ of subjects in each cluster are close to each other with respect to a Euclidian distance on the K-dimensional Euclidian space. Then, subjects in the same cluster have a close value for each element of $\boldsymbol{f_i}$.

The estimator $\widehat{\Delta}^{MPS}$ based on multiple PS stratification (MPS) is defined in a similar way to the standard stratified estimator stratifying subjects into the strata based on clustering. The MPS is expected to estimate the ACE well if one of the candidates for the PS model is correctly specified. In practice, too precise stratification may lead to unstable estimation. The number of strata should be determined accounting that the number of subjects in each stratum is not so small. To construct a confidence interval of the ACE, regarding strata as fixed, one may use a variance formula $\widehat{\mathrm{Var}}(\widehat{\Delta}^{MPS}) = \sum_{c=1}^{C}(n_c/n)^2\left\{\widehat{\mathrm{Var}}\left(\bar{Y}_c^{(1)}\right) + \widehat{\mathrm{Var}}\left(\bar{Y}_c^{(0)}\right)\right\}$, where n_c is the number of subjects in the c-th strata (cluster) and $\widehat{\mathrm{Var}}(Y^{(r)})$ is the sample variance of the averaged observe outcomes for the subjects assigned to $Z = r$ in the c-th stratum. In this paper, this method is called the naïve method. Alternatively, one may use the bootstrap method to construct confidence interval.

3 Simulation study

3.1 Performance when one of candidate models is correctly specified

In this subsection, we report results of simulation studies examining the performance of our proposed method. We generated 5,000 datasets as follows. Let $X^{(1)}, X^{(2)}$ and $X^{(3)}$ be independent random variables, which represent baseline covariates and follow U[−10, 10], on U[−4, 4] and the binomial distribution with $\mathrm{P}\left(X^{(3)} = 1\right) = 0.5$, respectively, where U[a,b] is the uniform distribution on [a, b]. The sample size was set as 200 or 500.

Dataset A:

The treatment allocations and outcomes were generated as follows:

$$\mathrm{logit}\{\mathrm{P}(Z = 1|X)\} = -1.5 + 1.6I(X^{(1)} > 0) + X^{(3)} + 0.6I(X^{(1)} > 0)X^{(3)},$$
$$Y = 2 + 2Z + I(X^{(1)} > 0) + 0.5X^{(3)} + 2I(X^{(1)} > 0)X^{(3)} + \epsilon,$$

where $\mathrm{logit}(x) = \log x/(1-x)$ and $X = (X^{(1)}, X^{(2)}, X^{(3)})$ and ϵ is a random error following the standard normal distribution.

To Dataset A, we applied our proposed method. We prepared three candidates PS models, denoted as PS1, PS2 and PS3 as which PS1 designated the true model for the PS. They are the logistic regression models with the following explanatory variables, respectively:

PS1 : $I\left(X^{(1)} > 0\right), X^{(3)}, I\left(X^{(1)} > 0\right)X^{(3)},$

PS2 : $I\left(t_1 < X^{(1)} \leq t_2\right), I\left(t_2 < X^{(1)}\right), X^{(3)}, I\left(t_1 < X^{(1)} \leq t_2\right)X^{(3)}, I\left(t_2 < X^{(1)}\right)X^{(3)},$

PS3 : $X^{(1)}, \left(X^{(1)}\right)^2, X^{(3)}, X^{(1)}X^{(3)}, \left(X^{(1)}\right)^2 X^{(3)},$

where t_1 and t_2 are the 33*th* and 67*th* percentiles of $X^{(1)}$, respectively. With these three PS models, we calculated the proposed estimators for the ACE with 2, 5, 7, 10 or 20 strata based on clustering by Ward's minimum variance method with the Euclid metric, which are denoted by MPS2, MPS5, MPS7, MPS10 and MPS20, respectively. We expected that with relatively small number, say 5, 7 and 10, of strata, our proposed method worked well since in the standard stratified estimator by the propensity score, 5 strata often work well. The reason for inclusion of MPS2 and MPS20 in the simulation study was to evaluate whether or not strata with large heterogeneity (MPS2) and strata with possibly small number of subjects (MPS20) might lead poor performance. For comparison, the standard stratified estimators with PS1, PS2 and PS3 were calculated. Five strata are defined according to 20*th*, 40*th*, 60*th*, and 80*th* percentiles of each PS. The estimators themselves are denoted by PS1, PS2 and PS3, respectively.

As an alternative to the proposed method, we also considered to constructing strata by applying a clustering method to covariates directly, which is called the direct clustering method. To vectors of covariates (not those of propensity scores), we applied Ward's minimum variance method with the Mahalanobis metric and constructed strata. We call this estimator the stratified estimator with the direct clustering (DC). The number of strata was 5, 7, 10 and 20, which were denoted by DC5, DC7, DC10 and DC20, respectively.

We also calculated the stratified estimator by a model-averaged estimate of the PS models. To be precise, the PS was estimated by the weighted average of PS1, PS2 and PS3, in which weights defined by the BIC according to the formula given in Example 7.2 of the textbook (Claeskens and Hjort, 2008). That is, let BIC_i be the BIC of PS*i* and the model averaged PS is defined as $PS_{MA} = C_{BIC}(1) \times \mathrm{PS}_1 + C_{BIC}(2) \times \mathrm{PS}_2 + C_{BIC}(3) \times \mathrm{PS}_3$, where $C_{BIC}(i) = \exp(-BIC_i/2)/\sum_{i=1}^{3} \exp(-BIC_i/2)$. The stratified estimator by it is denoted by model averaging (MA). We also evaluated performance of the stratified estimator by the PS selected by the model selection criteria BIC. We calculated the stratified estimator with the PS of the minimum BIC among the three models, which is denoted by model selection (MS).

Inverse weighting by the PS is alternative to stratification by the PS (Rosenbaum, 1987). The Inverse probability weighting estimator is defined as $\hat{\mu}_{IPW} = \hat{\mu}_{IPW}^{(1)} - \hat{\mu}_{IPW}^{(0)}$, where $\hat{\mu}_{IPW}^{(1)} = n^{-1}\sum_{i=1}^{n} Z_i Y_i / \hat{e}(X_i)$, $\hat{\mu}_{IPW}^{(0)} = n^{-1}\sum_{i=1}^{n}(1 - Z_i)Y_i/\left(1 - \hat{e}(X_i)\right)$ and $\hat{e}(X_i)$ is an estimate of the PS. The doubly robust estimator (DR) is a hybrid estimator of the IPW and a regression model for the outcome (Lunceford and Davidian, 2004). It is doubly robust in the sense that it estimates the ACE consistently if at least one of the PS model and the regression model for the outcome is correctly specified. Here we consider a doubly robust estimator $\hat{\mu}_{DR} = \hat{\mu}_{DR}^{(1)} - \hat{\mu}_{DR}^{(0)}$, where

$\hat{\mu}_{DR}^{(1)} = n^{-1}\sum_{i=1}^{n}\left\{Z_i Y_i - \left(Z_i - \hat{e}(X_i)\right)\hat{m}_1(X_i)\right\}/\hat{e}(X_i),$

$\hat{\mu}_{DR}^{(0)} = n^{-1}\sum_{i=1}^{n}\left[(1-Z_i)Y_i - \left\{(1-Z_i) - \left(1-\hat{e}(X_i)\right)\right\}\hat{m}_0(X_i)\right]/\left(1-\hat{e}(X_i)\right)$,

We defined the following outcome regression models for $Z = r$ ($r = 0, 1$):
$\hat{m}_r(X_i) = {\alpha_0}^{(r)} + {\alpha_1}^{(r)}X^{(1)} + {\alpha_2}^{(r)}X^{(3)}$, which is a misspecified model for $E(Y^{(r)}|X, Z = r)$.
We applied to our proposed method, MA, MS, the standard stratified estimators, DC with $\left(X^{(1)}, X^{(3)}\right)$, IPW and DR with PS1, PS2 or PS3. Let IPW and DR with PS*i* denoted by IPW*i* and DR*i* for *i*=1, 2, 3, respectively.

Dataset B:

The treatment allocation and the outcome were generated as follows:

$\text{logit}\{P(Z = 1|X)\} = 0.4 + 0.12X^{(1)} - 0.16X^{(2)} - 0.02X^{(1)}X^{(2)}$,
$Y = 1.8 + 2Z + 0.12X^{(1)} - 0.16X^{(2)} - 0.01X^{(1)}X^{(2)} + \epsilon$,

where ϵ follows the standard normal distribution. To Dataset B, we prepared three candidates of the propensity score model, which are denoted by PS1, PS2 and PS3 and are the logistic regression with the following explanatory variables, respectively:

PS1: $X^{(1)}, X^{(2)}, X^{(1)}X^{(2)}$,
PS2: $\exp\left(X^{(1)}\right), \exp\left(X^{(2)}\right), \exp\left(X^{(1)}\right)\exp\left(X^{(2)}\right)$,
PS3: $\log\left(X^{(1)}\right)^2, \log\left(X^{(2)}\right)^2, \log\left(X^{(1)}\right)^2\log\left(X^{(2)}\right)^2$,

We applied our proposed method, MA, MS, and the standard stratified estimators, DC with $\left(X^{(1)}, X^{(2)}\right)$, IPW and DR with PS1, PS2 or PS3 In DR, we applied the model:
$\hat{m}_r(X_i) = {\alpha_0}^{(r)} + {\alpha_1}^{(r)}X^{(1)} + {\alpha_2}^{(r)}\left(X^{(1)}\right)^2 + {\alpha_3}^{(r)}X^{(2)}$, which is a misspecified model for the outcome.
The results for Datasets A and B are summarized in Tables 1 and 2, in which PS1 correctly specified the true model for the PS. In both simulation studies, as anticipated, the stratified estimator with a misspecified PS model (PS2, PS3) has a serious bias. The MPS successfully removed biases except for the MPS2 and MPS20. The poor performance of MPS2 suggests that stratification with insufficient homogeneity within each stratum may lead to poor estimates, and that of MPS20 suggests that too precise stratification may lead to unstable estimation. Thus, determination of the number of strata is very important. In order to determine the number of strata, it is very helpful to check the distribution of the PSs in each stratum. Thus, similar to the standard stratified estimator, a relatively small number of strata are recommended. The MA and MS removed biases in Dataset B, but not in Dataset A with n=500. We counted frequencies in which each model had the minimum BIC among the three models for 5,000 simulated realizations. In Dataset B, PS1, which is the correctly specified PS model, attained the

minimum BIC in 4,201 of 5,000 cases, or 84.02 percent of the time. On the other hand, in Dataset A, PS1 attained the minimum BIC in 55.02 percent [2,751/5,000] and the result was in 2,249/5,000 for 44.98 percent realization, PS3, which is incorrectly specified, had the lowest BIC. This indicates that BIC may select suboptimal models with moderate sample size and this may have led to biases in MA and MS in Dataset A. Since the MPS is free from any model-selection procedures, it worked well even for Dataset A. Although the estimated result of MA and MS was good in Dataset B for n= 500, arising of bias was observed in n= 200. It is considered to be the cause that the cases where PS1, which is the correctly specified model of PS, under few sample situations has the minimum BIC decrease from 84.02 percent [4,201/5,000] to 6.3 percent [315/5,000] remarkably. In n= 200 of Dataset A, as for the case where PS1 has the minimum BIC, reduction was similarly observed from 55.02 percent [2,751/5,000] to 2.54 percent [127/5,000]. In both Dataset, considerable biases were observed with DC. Moreover, also in IPW2, IPW3, DR2, and DR3, non negligible bias was observed in the estimated result by the incorrectly PS specificified model (PS2, PS3).

3.2 Performance in the presence of an outlying observation

Our proposed estimator is free from any model-selection procedures, whereas the MA and the MS based methods discussed in the previous subsection rely on the BIC. The BIC is based on a likelihood function and thus may be sensitive to outlying observations. Then, in this subsection, we ask whether or not the model-averaging and the model-selection may be more sensitive to outlying observations than the proposed method. To examine this hypothesis, we replaced the last subject of the Dataset B with a subject of an outlying observation. The subject has $X^{(1)} = 20$ and $X^{(2)} = 15$, and the PS less than 0.01. We assign this subject to $Z = 1$, which hardly occurs. We apply the same method as in Dataset B, and call this variation Dataset B*. The results are presented in Table 3. Although the dataset is same as that in Dataset B except for one observation, the performances of the MA and the MS in Dataset B* were very different from those in Dataset B. The MA and the MS did not work well in Dataset B*. Being contaminated with an outlying observation, the frequency with which BIC selected the correct model PS1 decreased from 84.02 percent [4,201/5,000] to 19.38 percent [969/5,000] leading to biased estimation of the MA and MS as presented in Table 3. On the other hand, the MPS worked well, except for MPS2, both in Dataset B and B* and seemed to be stable against outlying data. In common with Datasets A and B, non negligible bias was observed in DC5, DC7, DC10 and DC20. We also observed that if the PS is correctly specified, both the IPW and the DR have only a negligible bias, whereas they have a considerable increase in the MSE compared with the proposed method.

3.3 Robustness against misspecification of the link function in estimation of the propensity score

The use of the stratified PS estimator with the generalized linear model is robust against misspecification

of the link function (Drake, 1993). We wondered whether or not the MPS would maintain this property. We generated $X^{(1)}$, $X^{(2)}$, Z and the outcome in the same way as in Dataset B as follows:

Datasets C-E:

$$\mathrm{P}(Z = 1|X) = G\left(\beta_0 + \beta_1 X^{(1)} + \beta_2 X^{(2)} + \beta_3 X^{(1)} X^{(2)}\right),$$

$$Y = 1.8 + 2Z + 0.12X^{(1)} - 0.16X^{(2)} - 0.01X^{(1)}X^{(2)} + \epsilon,$$

where G is the cumulative distribution function of the *t*-distribution of degree of freedom 0.7 (Dataset C : $\beta_0 = 0.4, \beta_1 = 0.35, \beta_2 = -0.16, \beta_3 = -0.09$), the quadratic function $G(x) = 0.004(x + 7.2)^2$ (Dataset D : $\beta_0 = 0.3, \beta_1 = 0.23, \beta_2 = -0.08, \beta_3 = -0.04$), and probit (Dataset E : $\beta_0 = 0.09$, $\beta_1 = 0.04, \beta_2 = -0.06, \beta_3 = -0.02$), and ϵ follows the standard normal distribution. Again, 5,000 datasets were generated. To these datasets, the same sets of PS models were considered as in Simulation study B. Table 2 summarizes the results for n=500. The MPS, as well as the standard stratified estimator with the correctly specified PS model (PS1), have only negligible biases indicating that they are robust against misspecification of the link function in the PS model. In spite of having used PS1 specified surely, we observed that the IPW1 and the DR1 are not robust in Datasets C. Among all simulation studies, non ignorable bias was observed in the DC. Moreover, in Datasets D and E, considerable biases were observed in the average and the MSE of MA and MS.

3.4 Coverage probability

Empirical coverage probabilities of a Wald-type confidence interval based on the variance formula, which is called the naïve method, were calculated with the datasets used in the previous subsections. We also evaluated empirical coverage probabilities of the percentile-based bootstrap confidence interval. That is, a confidence interval was constructed by the 2.5th and 97.5th percentiles of the estimated treatment effects over 1,000 bootstrapped samples in Table 5. We observed that coverage probabilities of the MPS evaluated by the naïve method was anti-conservative, and that coverage probabilities of the bootstrap confidence intervals are close to the nominal level of 95 percent for MPS5, MPS7 and MPS10. We also observed that with 20 strata, both methods provide substantially anti-conservative confidence intervals. Then, too much strata is not recommended again, and use of the bootstrap confidence interval is recommended.

3.5 Performance with more covariates

In this subsection, we report results of an additional simulation study with more covariates and more candidate PS models, which would be more practical.

Let $X^{(k)}$, k = 1, 2, ..., 10, be mutually independent random variables: for $X^{(1)}$ ~U$[-10, 10]$, $X^{(2)}$

~U[−4, 4], $X^{(3)}$ ~U[−7, 7], $X^{(4)}$ ~N(5, 4), $X^{(5)}$ ~N(10, 9), $X^{(6)}$ ~N(3, 1), and $X^{(k)}$, k = 7, 8, 9, 10 follow the binomial distribution with $\mathrm{P}(X^{(7)} = 1) = 0.5$, $\mathrm{P}(X^{(8)} = 1) = 0.5$, $\mathrm{P}(X^{(9)} = 1) = 0.6$ and $\mathrm{P}(X^{(10)} = 1) = 0.3$ respectively. The sample size was set as 200 or 500.

Dataset F:

Denote ${I_M}^{(k)} = I(X^{(k)} > m^{(k)})$, ${I_{12}}^{(k)} = I(t_1{}^{(k)} < X^{(k)} \le t_2{}^{(k)})$ and ${I_2}^{(k)} = I(t_1{}^{(k)} < X^{(k)})$, where $m^{(k)}$ is the median of $X^{(k)}$ and $t_1{}^{(k)}$, $t_2{}^{(k)}$ are the 33*th* and 67*th* percentiles of $X^{(k)}$, respectively. The treatment allocations and outcomes were generated as follows.

$$\begin{aligned}\mathrm{logit}\{\mathrm{P}(Z=1|X)\} = {} & -1.5 + 0.8{I_M}^{(1)} + 1.2{I_M}^{(2)} - 1.2{I_M}^{(3)} + 0.8{I_M}^{(4)} - 1.0{I_M}^{(5)} + 1.6{I_M}^{(6)} + 0.6X^{(7)} \\ & -1.32X^{(8)} + 0.48X^{(9)} + 0.36X^{(10)} + 0.4{I_M}^{(1)}X^{(9)} + 0.4{I_M}^{(3)}X^{(7)} + 0.4{I_M}^{(5)}X^{(9)} \\ & +0.24{I_M}^{(2)}X^{(7)} + 0.32{I_M}^{(6)}X^{(8)} + 0.16{I_M}^{(4)}X^{(10)},\end{aligned}$$

$$\begin{aligned}Y = {} & 6 + 2Z + 0.24{I_M}^{(1)} + 0.42{I_M}^{(2)} - 0.36{I_M}^{(3)} + 0.24{I_M}^{(4)} - 0.3{I_M}^{(5)} + 0.48{I_M}^{(6)} + 0.75X^{(7)} \\ & -1.95X^{(8)} - 0.6X^{(9)} + 0.45X^{(10)} + 0.2{I_M}^{(1)}X^{(9)} + 0.2{I_M}^{(3)}X^{(7)} + 0.2{I_M}^{(5)}X^{(9)} + 0.12{I_M}^{(2)}X^{(7)} \\ & +0.16{I_M}^{(6)}X^{(8)} + 0.08{I_M}^{(4)}X^{(10)} + \epsilon,\end{aligned}$$

where $\mathrm{logit}(x) = \log x/(1-x)$ and $X = (X^{(1)}, X^{(2)}, \ldots, X^{(10)})$, and ϵ is a random error following the standard normal distribution. 5,000 datasets were generated.

To the datasets, we prepared five candidates PS models, denoted as PS1(median), PS2(tertile), PS3(quadratic), PS4(linear) and PS5(exponential) as which PS1 designated the true model for the PS. They are the logistic regression models with the following explanatory variables, respectively:

PS1 : ${I_M}^{(1)}, {I_M}^{(2)}, {I_M}^{(3)}, {I_M}^{(4)}, {I_M}^{(5)}, {I_M}^{(6)}, X^{(7)}, X^{(8)}, X^{(9)}, X^{(10)}, {I_M}^{(1)}X^{(9)}, {I_M}^{(3)}X^{(7)}, {I_M}^{(5)}X^{(9)}, {I_M}^{(2)}X^{(7)}, {I_M}^{(6)}X^{(8)}, {I_M}^{(4)}X^{(10)},$

PS2 : ${I_{12}}^{(1)}, {I_2}^{(1)}, {I_{12}}^{(2)}, {I_2}^{(2)}, {I_{12}}^{(3)}, {I_2}^{(3)}, {I_{12}}^{(4)}, {I_2}^{(4)}, {I_{12}}^{(5)}, {I_2}^{(5)}, {I_{12}}^{(6)}, {I_2}^{(6)}, X^{(7)}, X^{(8)}, X^{(9)}, X^{(10)}, {I_{12}}^{(1)}X^{(9)}, {I_2}^{(1)}X^{(9)}, {I_{12}}^{(3)}X^{(7)}, {I_2}^{(3)}X^{(7)}, {I_{12}}^{(5)}X^{(9)}, {I_2}^{(5)}X^{(9)}, {I_{12}}^{(2)}X^{(7)}, {I_2}^{(2)}X^{(7)}, {I_{12}}^{(6)}X^{(8)}, {I_2}^{(6)}X^{(8)}, {I_{12}}^{(4)}X^{10}, {I_2}^{(4)}X^{(10)},$

PS3 : $X^{(1)}, (X^{(1)})^2, X^{(2)}, (X^{(2)})^2, X^{(3)}, (X^{(3)})^2, X^{(4)}, (X^{(4)})^2, X^{(5)}, (X^{(5)})^2, X^{(6)}, (X^{(6)})^2, X^{(7)}, X^{(8)}, X^{(9)}, X^{(10)}, X^{(1)}X^{(9)}, (X^{(1)})^2X^{(9)}, X^{(3)}X^{(7)}, (X^{(3)})^2X^{(7)}, X^{(5)}X^{(9)}, (X^{(5)})^2X^{(9)}, X^{(2)}X^{(7)}, (X^{(2)})^2X^{(7)}, X^{(6)}X^{(8)}, (X^{(6)})^2X^{(8)}, X^{(4)}X^{(10)}, (X^{(4)})^2X^{(10)},$

PS4 : $X^{(1)}, X^{(2)}, X^{(3)}, X^{(4)}, X^{(5)}, X^{(6)}, X^{(7)}, X^{(8)}, X^{(9)}, X^{(10)}, X^{(1)}X^{(9)}, X^{(3)}X^{(7)}, X^{(5)}X^{(9)}, X^{(2)}X^{(7)}, X^{(6)}X^{(8)}, X^{(4)}X^{(10)},$

PS5 : $\exp(X^{(1)}), \exp(X^{(2)}), \exp(X^{(3)}), \exp(X^{(4)}), \exp(X^{(5)}), \exp(X^{(6)}), X^{(7)}, X^{(8)}, X^{(9)}, X^{(10)},$

The results for Dataset F are summarized in Table 4, in which PS1 correctly specified the true model for the PS. We observed that DC had substantial biases with relatively large number of covariates. Despite an

increment of the PS candidates, proposed method observed successfully removed biases under the complicated situation. Non ignorable bias was observed for MA and MS.

4 Example

In this section, we illustrate our proposed method using a dataset from the Tone study, which is a community survey conducted in Japan (Miyamoto, 2009). Subjects' baseline covariates were collected from 2001 to 2002 by interviews using a structured questionnaire recording age, sex, education and assessing previous medical and psychiatric diseases and dementia risk factors. After completing the interview, all participants underwent a group assessment which used a set of five tests measuring the following cognitive domains: attention, memory, visuospatial function, language and reasoning (The Five-Cog test). The Five-Cog test can evaluate the levels of mild cognitive impairment called aging associated cognitive decline and can be used to screen for elderly subjects who are at high risk of developing dementia. All participants underwent the same cognitive assessment at the 3-year follow up.

We enrolled 935 subjects with baseline measurements and the follow-up data at 2005 in our dataset. The primary objective of the Tone study was to examine whether a physical examination contributes to the prevention of dementia. Subjects were assigned to either the physical examination or the observational groups. The assignment was not determined randomly, but according to the subject's preference. Two hundred and thirty-four and 701 subjects were assigned to the physical examination and the observational groups, respectively. We use the memory score as the outcome variable for illustration. The baseline memory score was unbalanced between the two groups: the physical examination group had a median score of 13.1 (the lower and upper 25% percentiles: 9, 16) and the observational group had 9.8 (6, 13). Educational status was also unbalanced: subjects receiving <9 years, 9-12 years, and >12 years of education comprised 46/234 = 19.6%, 130/234 = 55.6% and 58/234 = 24.8%, of the physical examination group, versus 339/701 = 48.3%, 228/701 = 41.1% and 74/701 = 10.6% of the observational group. These covariates may be associated with the outcome, the memory score at 2004. Therefore, they must be adjusted in estimating the ACE. We applied the following logistic regression for modeling the PS:

$$\text{logit}\{P(Z=1|X)\} = \alpha_0 + \alpha_1 AGE + \alpha_2 GENDER + \alpha_{31} EDU1 + \alpha_{32} EDU2 + \alpha_4 SMOKE + \alpha_5 DRINK + \alpha_6 SLEEP + h(MEN),$$

where $AGE, GENDER, SMOKE, DRINK, SLEEP$ and MEN are age at 2001, gender, smoking status, drinking status, napping status, the memory score at 2001, and $h(.)$ is a function. $EDU1$ and $EDU2$ are dummy variables for educational status of less than 9 years and that of 9 to 12 years, respectively. Considering three functions as $h(.)$, we defined three candidates for the PS model:

PS1: $h(MEN) = \alpha_{81} MEN + \alpha_{82} MEN^2$,

PS2: $h(MEN) = \alpha_{81} I(t_1 < MEN \leq t_2) + \alpha_{82} I(t_2 < MEN)$,
PS3: $h(MEN) = \alpha_8 I(m_{0.5} < MEN)$,

where $m_{0.5}$ is the empirical median of MEN. With these three PS models, we applied our proposed method with 5, 7, 10 or 20 strata constructed by Ward's minimum variance method. They are denoted by MPS5, MPS7, MPS10 and MPS20, respectively. It is very important to determine how many strata are used in our proposed method. The simulation studies presented in Section 3 indicates that too many strata may lead to biased estimates, and that a relatively small number of strata are more effective. The R^2 measure in clustering (Massart and Kaufman, 1983) was 0.951 with 7 strata and 0.969 with 10 strata. With 10 strata, one of the 10 strata had only 10 subjects (7 in the physical examination group and 3 in the observational group) and the stratum-specific average may be unstable. From these observations, we determined to estimate the ACE with 7 strata. Estimates are presented in Table 6, together with the stratum-specific difference of means, in which confidence intervals by the naïve method is presented. In the standard stratified estimation by the propensity score, it is very important to check the overlap of the distributions of the PS between the two groups. This is true in applying the proposed stratified estimation with the multiple PSs. Table 6 also shows ranges of the three PSs in each group in each stratum. It indicates that within each of 7 strata, the two groups had a good overlap for all three PSs. We also observed that ranges given in Table 6 are similar to those by the standard stratified analysis of each PS with 5 strata by 20*th*, 40*th*, 60*th*, and 80*th* percentiles. Thus, the proposed estimator is anticipated to work better than the standard stratified estimator regardless of the choice of the PS model. In Table 7, we summarize estimates with our proposed method, together with the crude estimate (no adjustment) and the standard stratified estimators with PS1, PS2 or PS3 respectively. The bootstrap 95 percent confidence intervals given in Table 7 were based on 1000 replicates. The difference of simple sample means of the physical examination and the observational groups was 7.06 (95%*CI*: 6.18, 7.91). PS1, PS2 and PS3 were 3.91(3.03, 4.71), 4.02 (3.17, 4.87) and 4.59 (3.67-5.41) respectively. Estimate with PS3 was larger than that by PS1 or that by PS2. Then, one may wonder which result was most reliable. MPS7 was 4.01 (3.17, 4.84), indicating that PS3 is not reliable.

5 Conclusions

Our estimator can incorporate multiple PS models and removes bias if one of them specifies the PS correctly. In the standard stratified estimator by the PS, one can easily stratify subjects according to the PS since it is scalar. This simplicity is lost in our proposed method. However, our method is still simple and the use of clustering techniques enables us to construct strata based on multiple PSs efficiently. Stratified analysis by the PS has an advantage over the IPW estimator and the doubly robust estimator in its simplicity and robustness: it is easy for non-statisticians to understand and is robust against highly variable weights, outlying observations and misspecification of the link function in modeling of the PS.

Through the simulation studies, we observed that the simple percentile-based bootstrap confidence interval works satisfactorily, although more complicated variants such as BCa bootstrap can be employed for the standard stratified estimator by the PS (Tu and Zhou, 2002).

We considered only Ward's minimum variance method in clustering PS vectors. Any hierarchical clustering techniques can be applied (Gan et al., 2007). Although we did not show the results, our simulation study for comparison of performances with several clustering methods suggest that performance of the estimator does not rely strongly on the choice of the clustering method. Regardless of choice of a clustering method, as done in Table 6, it is important to check the overlap of each PS included in clustering.

Non-linear regression techniques such as the generalized additive model (Hasti and Tibsirani, 1993) or machine learning techniques (Lee et al., 2010) may be useful to against biases due to misspecification. Performance of these methods strongly relies on smoothing parameters and stopping rules, respectively, and selection of them is crucial. Our method can reduce risk of misspecification by incorporating several candidate models in a simple way without relying on model-selection criteria, which may not work well in practice as demonstrated in Simulation section.

By combining several PS models with nonlinear or machine learning techniques, risk can be further reduced. Recently, Han and Wang (2013) proposed an estimator of multiple robustness. Their empirical-likelihood-based estimator can incorporate multiple PS models and outcome regression models, and if at least one of the PS models and the outcome regression models hold, the estimator is shown to be consistent. To execute their estimator, one must solve an equation, which may suffer from a multiple roots issue or a non-convergence issue. Although this issue was tackled by the latest paper Han (2014), which proposed the algorithm with easy solving and implementing the multiple roots issue in Han and Wang (2013), it is still complicated. Although formal theoretical justification of consistency has not been made, our method provides a very simple way to apply multiply robust estimation. Indeed, our method can be easily implemented using a standard statistical software covering logistic regression and clustering.

The PS matching has been widely used in practice for a very long time (Connors et al., 1996; Ayanian et al., 2002; Abidov et al., 2005; Shishehbor et al., 2006). Our idea to utilize multiple PSs jointly can provide some benefits to matching analysis. That is, one may construct matched samples robust against misspecification of the PS model based on a vector of several PSs by using a distance such as the Mahalanobis distance. Recently, Leacy and Stuart (2014) proposed new matching approach based on a pair of the PS and an alternative balancing score called the prognostic score. Their idea can be generalized to multiple robust matching incorporating the multiple PSs and prognostic scores.

Acknowledgements

The research of the second author was partly supported by a grant-in-aid (C) from the Ministry of Education, Science, Sports and Technology of Japan (#21500286).

Conflict of Interest

The authors have declared no conflict of interest.

References

Abidov, A., Rozanski, A., Hachamovitch, R. (2005). Prognostic Significance of Dyspnea in Patients Referred for Cardiac Stress Testing. *N Engl J Med* 353: 1889-1898.

Ayanian, J., Landrum, M. B., Guadagnoli, E. (2002). Specialty of Ambulatory Care Physicians and Mortality among Elderly Patients after Myocardial Infarction. *N Engl J Med* 347:1678-1686.

Bang, H., Robins, JM. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* 61: 962-73.

Barker, F. GII., Chang, S. M., Gutin, P. H. (1988). Survival and functional status after resection of recurrent glioblastoma multiforme. *Neurosurgery* 42: 709-720.

Bateman, B., Hernandez-Diaz, S., Huybrechts, K. (2013). Outpatient calcium-channel blockers and the risk of postpartum haemorrhage: a cohort study. *An International Journal of Obstetrics & Gynaecology* 120(13): 1668-1677.

Claeskens G, Hjort NL. (2008). *Model Selection and Model Averaging.* Cambridge: Cambridge University Press.

Connors, A. F., Speroff, T, Dawson, N. V. (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. SUPPORT Investigators. *JAMA* 276(11):889-97.

Coyte, P. C., Young, W., Croxford, R. (2000). Costs and outcomes associated with alternative discharge strategies following joint replacement surgery: analysis of an observational study using a propensity score. *Journal of Health Economics* 192: 907-929.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics* 49: 1231-1236.

Gan, G., Ma, C., Wu, J. (2007). *Data clustering: theory, algorithm, and applications*. Society for Industrial and Applied Mathematics, and American Statistical Association.

Han, P., Wang, L. (2013). Estimation with missing data: beyond double robustness. *Biometrika* 100: 417-430.

Han, P. (2014). A further study of the multiply robust estimator in missing data analysis. *Journal of Statistical Planning and Inference* 148: 101-110.

Hasti, T. J., Tibshirani, R. J. (1993). *Generalized additive models*. Chapman & Hall/CRC.

Kang, D. Y., Schafer, J. L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 22: 523-539.

Leacy, F. P., Stuart, E. A. (2014). On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Statistics in Medicine* 20: 3488-3508.

Lee, B. K., Lessler, J., Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine* 29: 337-346.

Lunceford, J. K., Davidian, M. (2004). Stratification and weighting via the propensity score in estimation

of causal treatment effects: a comparative study. *Statistics in Medicine* 23: 2937-2960.

Massart, D. L., Kaufman, L. (1983). *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*. New York: John Wiley & Sons.

Mirkin, B. (2012). *Clustering: A Data Recovery Approach*, Second Edition. Computer Science & Data Analysis: Chapman & Hall/CRC.

Miyamoto, M., Kodama, C. (2009). Dementia and mild cognitive impairment among non-responders to a community survey. *Journal of Clinical Neuroscience* 16: 270-276.

Rosenbaum, P. R., Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70: 41-55.

Rosenbaum, P. R, 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. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association* 82: 387-394.

Shishehbor, M. H., Litaker, D., Pothier, C. E. (2006). Association of socioeconomic status with functional capacity, heart rate recovery, and all-cause mortality. *JAMA* 15;295(7):784-92.

Tu, W., Zhou, X. H. (2002). A bootstrap confidence interval procedure for the treatment effect using propensity score subclassification. *Health Services & Outcome Research Methodology* 3: 135-147.

Ward, J. H. (1963). Hierarchical Grouping to Optimize an objective Function. *Journal of the American Statistical Association* 58: 236-244.

Williamson, E. J., Morley, R., Lucas, A. (2012). Variance estimation for stratified propensity score estimators. *Statistics in Medicine* 31: 1617-1632.

Table 1 Summary of results of the simulation study for evaluation of empirical biases (Average), mean-squared errors (MSE) in n=500:PS1 correctly specifies the propensity score.

		Dataset A		Dataset B	
Method		Average (true=2)	MSE	Average (true=2)	MSE
Stratification	PS1	2.00	0.012	2.04	0.012
	PS2	2.35	0.138	2.25	0.073
	PS3	2.22	0.064	2.55	0.316
Clustering	MPS2	2.27	0.113	2.19	0.047
	MPS5	2.03	0.014	2.08	0.017
	MPS7	2.01	0.012	2.05	0.014
	MPS10	2.00	0.013	2.03	0.012
	MPS20	1.98	0.015	1.97	0.013
Direct Clustering	DC5	2.26	0.096	2.15	0.034
	DC7	2.20	0.060	2.11	0.023
	DC10	2.14	0.036	2.07	0.017
	DC20	2.05	0.018	2.02	0.013
BIC Based	MA	2.17	0.043	2.05	0.014
	MS	2.09	0.030	2.07	0.018
Weighting	IPW1	2.02	0.012	2.01	0.012
	IPW2	2.33	0.125	2.36	0.147
	IPW3	2.17	0.048	2.55	0.316
Doubly Robust	DR1	2.00	0.013	2.00	0.013
	DR2	2.24	0.073	2.30	0.164
	DR3	2.20	0.059	2.46	0.220

Table 2 Summary of results of the simulation study for evaluation of empirical biases (Average), mean-squared errors (MSE) in small samples (n=200):PS1 correctly specifies the propensity score.

		Dataset A		Dataset B	
Method		Average (true=2)	MSE	Average (true=2)	MSE
Stratification	PS1	2.00	0.032	2.04	0.031
	PS2	2.34	0.156	2.26	0.099
	PS3	2.22	0.085	2.55	0.336
Clustering	MPS2	2.27	0.113	2.19	0.047
	MPS5	2.02	0.035	2.06	0.034
	MPS7	1.99	0.035	2.02	0.032
	MPS10	1.96	0.039	1.98	0.032
	MPS20	1.83	0.071	1.86	0.051
Direct Clustering	DC5	2.25	0.116	2.14	0.050
	DC7	2.18	0.079	2.09	0.040
	DC10	2.10	0.054	2.05	0.034
	DC20	1.90	0.055	1.92	0.040
BIC Based	MA	2.21	0.081	2.19	0.071
	MS	2.21	0.080	2.24	0.091
Weighting	IPW1	2.07	0.040	2.03	0.031
	IPW2	2.37	0.170	2.39	0.187
	IPW3	2.22	0.094	2.55	0.337
Doubly Robust	DR1	2.00	0.036	2.01	0.035
	DR2	2.24	0.093	2.31	0.141
	DR3	2.18	0.110	2.46	0.238

Table 3 Summary of results of the simulation study (n=500) for evaluation of empirical biases (Average), mean-squared errors (MSE) in the presence of misspecification of the link function in estimation of the propensity score: misspecification in PS1 lies only on the link function.

		Dataset B* (with outlier)		Dataset C (PS1 has a misspecified link function)		Dataset D (PS1 has a misspecified link function)		Dataset E (PS1 has a misspecified link function)	
Method		Average (true=2)	MSE	Average (true=2)	MSE	Average (true=2)	MSE	Average (true=2)	MSE
Stratification	PS1	2.04	0.012	2.03	0.014	2.03	0.015	2.04	0.012
	PS2	2.25	0.074	2.38	0.167	2.20	0.057	2.16	0.038
	PS3	2.54	0.304	2.75	0.577	2.36	0.152	2.36	0.141
Clustering	MPS2	2.19	0.049	2.24	0.069	2.14	0.035	2.11	0.022
	MPS5	2.08	0.017	2.06	0.018	2.04	0.016	2.04	0.012
	MPS7	2.06	0.014	2.04	0.015	2.03	0.015	2.02	0.011
	MPS10	2.04	0.013	2.01	0.015	2.01	0.015	2.01	0.010
	MPS20	1.99	0.012	1.94	0.020	1.98	0.017	1.98	0.011
Direct Clustering	DC5	2.14	0.030	2.19	0.053	2.09	0.023	2.10	0.021
	DC7	2.11	0.023	2.14	0.033	2.06	0.018	2.08	0.016
	DC10	2.07	0.017	2.09	0.022	2.04	0.016	2.05	0.013
	DC20	2.02	0.012	2.01	0.017	1.99	0.016	2.02	0.011
BIC Based	MA	2.19	0.053	2.03	0.014	2.16	0.045	2.07	0.016
	MS	2.20	0.059	2.03	0.014	2.19	0.055	2.11	0.024
Weighting	IPW1	2.00	0.028	1.87	0.053	2.01	0.013	2.01	0.010
	IPW2	1.72	1.601	2.37	0.269	2.26	0.084	2.25	0.074
	IPW3	2.54	0.304	2.75	0.576	2.37	0.153	2.36	0.141
Doubly Robust	DR1	2.00	0.046	1.76	0.135	2.01	0.015	2.01	0.010
	DR2	2.29	0.175	2.18	7.142	2.26	0.085	2.29	0.065
	DR3	2.46	0.219	2.71	0.512	2.32	0.118	2.23	0.097

Table 4 Summary of results of the simulation study under the situation with more covariates (Dataset F).

Method		n=200		n=500	
		Average (true=2)	MSE	Average (true=2)	MSE
Stratification	PS1	2.05	0.095	2.10	0.033
	PS2	2.24	0.144	2.28	0.098
	PS3	2.20	0.131	2.26	0.086
	PS4	2.24	0.123	2.26	0.087
	PS5	2.43	0.219	2.43	0.194
Clustering	MPS5	2.04	0.092	2.14	0.040
	MPS7	1.98	0.097	2.10	0.035
	MPS10	1.89	0.107	2.08	0.033
	MPS20	1.61	0.227	2.00	0.032
Direct Clustering	DC5	2.81	0.701	2.81	0.674
	DC7	2.76	0.622	2.76	0.600
	DC10	2.69	0.528	2.71	0.527
	DC20	2.49	0.285	2.61	0.387
BIC Based	MA	2.41	0.212	2.24	0.079
	MS	2.42	0.216	2.25	0.086

Table 5 Summary of results of the simulation study for evaluation of coverage probabilities.

	Method	Naïve	Bootstrap
Dataset A (n=500)	MPS5	0.930	0.947
	MPS7	0.947	0.961
	MPS10	0.945	0.963
	MPS20	0.908	0.942
Dataset A (n=200)	MPS5	0.922	0.973
	MPS7	0.918	0.968
	MPS10	0.888	0.948
	MPS20	0.672	0.579
Dataset B (n=500)	MPS5	0.885	0.912
	MPS7	0.915	0.947
	MPS10	0.931	0.968
	MPS20	0.912	0.941
Dataset B (n=200)	MPS5	0.913	0.971
	MPS7	0.914	0.975
	MPS10	0.897	0.955
	MPS20	0.775	0.703

Table 6 Stratum-specific treatment effects and the average causal effect estimated by the proposed method with seven clustering-based strata in the Tone study: confidence intervals are based on the naïve method.

Strata	group	n	PS1(range)	PS2(range)	PS3(range)	Estimate (95%CI)
1	Z=1	15	0.051 - 0.114	0.057 - 0.094	0.057 - 0.108	5.33(2.86, 7.80)
	Z=0	157	0.025 - 0.116	0.036 - 0.120	0.036 - 0.131	
2	Z=1	19	0.075 - 0.187	0.106 - 0.177	0.095 - 0.189	2.83(0.22, 5.43)
	Z=0	164	0.070 - 0.198	0.091 - 0.197	0.080 - 0.194	
3	Z=1	26	0.162 - 0.242	0.147 - 0.238	0.160 - 0.266	4.73(2.48, 6.99)
	Z=0	93	0.108 - 0.264	0.141 - 0.249	0.152 - 0.283	
4	Z=1	33	0.237 - 0.305	0.201 - 0.319	0.201 - 0.337	3.97(2.00, 5.94)
	Z=0	83	0.165 - 0.300	0.193 - 0.319	0.199 - 0.364	
5	Z=1	28	0.279 - 0.404	0.269 - 0.410	0.242 - 0.441	3.74(1.94, 5.54)
	Z=0	76	0.272 - 0.399	0.244 - 0.405	0.249 - 0.430	
6	Z=1	48	0.356 - 0.484	0.347 - 0.494	0.334 - 0.485	3.10(1.22, 4.99)
	Z=0	59	0.350 - 0.486	0.336 - 0.485	0.327 - 0.533	
7	Z=1	63	0.437 - 0.697	0.418 - 0.648	0.426 - 0.616	4.24(2.59, 5.89)
	Z=0	60	0.437 - 0.671	0.472 - 0.625	0.399 - 0.616	
Pooled	Z=1	232	0.051 - 0.697	0.057 - 0.648	0.057 - 0.616	4.01(3.13, 4.88)
	Z=0	692	0.025 - 0.671	0.036 - 0.625	0.036 - 0.616	

Table 7 Summary of the estimated average causal effects with a bootstrap confidence interval in the Tone study.

Method	PS	# of strata	ACE (95%CI)
Crude			7.06(6.18, 7.91)
Stratification by PS	PS1	5	3.91(3.03, 4.71)
	PS2	5	4.02(3.17, 4.87)
	PS3	5	4.59(3.67, 5.41)
Proposed	MPS5	5	4.02(3.29, 4.99)
	MPS7	7	4.01(3.17, 4.84)
	MPS10	10	3.97(3.04, 4.75)
	MPS20	20	3.89(2.81, 4.61)