# Adjusting Complex Heterogeneity in Treatment Assignment in Observational Studies

Jichang Yu[1], Haibo Zhou[2,*], Xianchen Liu[3], Fei Zou[2] and Richard J. Willke[3]

[1]*School of Statistics and Mathematics, Zhongnan University of Economics and Law, Hubei, 430073, China*

[2]*Department of Biostatistics, University of North Carolina at Chapel Hill, NC, 27599-7420, USA*

[3]*Global Health Economics & Outcomes Research, Pfizer, Inc, NY, 10017, USA*

**Abstract:** Treatment assignment in observational studies is complex and can be influenced by many factors that include patient characteristics, physician practices, and health care systems. These influences can present heterogeneity or clustering effects in the treatment assignment. If those heterogeneity or clustering effects are not appropriately adjusted, the estimated treatment effect may be severely biased. Through a series of models that mimic various level of heterogeneity in treatment assignment in observational studies, we evaluate, through simulation study, the performance of several estimators under the impact of different types of heterogeneity. These estimators include propensity score stratification, propensity score inverse probability weighting, propensity score regression and the partial least squares method. Our results suggest that the partial least squares method is most robust while the dummy variable adjustment method in propensity regression also performs fairly consistently. We use the proposed method to analyze a data set from the German Breast Cancer Study Group study.

**Keywords:** Heterogeneity, partial least squares, propensity score.

## 1. INTRODUCTION

Correctly identifying a treatment interventions effectiveness for a disease can help patients, health care professionals, and purchasers to make informed decisions. Randomized controlled trials (RCT) have been the gold standard for demonstrating the efficacy and safety of a novel intervention for many years. In randomized controlled trials, patients are selected and randomly assigned to an intervention according to pre-determined inclusion, exclusion criteria and a random assignment protocol to ensure there is no confounding due to the baseline covariates when comparing treatment effects in two or more groups. While RCT are desirable in this regard, it is not always feasible to conduct them in practice due to ethical, time, or cost considerations. For example, it would be unethical to expose patients deliberately to less effective treatments in the case of vaccines in the face of a threat of an influenza pandemic; researchers cannot randomize people to smoking or gender to study their effects on lung cancer or design experimental studies to determine the effects of pollution or global warming on people.

Observational studies, on the other hand, have gathered much attention recently due to the availability of real world data [1-4]. The treatment examined in observational studies is not randomly assigned, but is influenced by the patients and their health care providers. The observational data sets could come from several sources that include insurance claim data, electronic medical records, prescription records, patient's records and so on. These data sets, usually large and rich in information, cover a much broader scope than RCT data can provide. For example, the Women's Health Initiative (WHI) observational study enlisted 93, 676 postmenopausal women between the ages of 50 to79, and the participants were tracked over an average of eight years. It was conducted to provide reliable estimates of the extent to which known risk factors to predict heart disease, cancers and fractures, to identify new risk factors for these and other diseases in women, and to create a future resource to identify biological indicators of disease, especially substances and factors found in blood.

The challenge in analyzing the observational studies for treatment effect is that the treatment assignment in observational studies is deliberate choice made by physicians, patients, and/or the payer and is far from random. There may be systematic differences between the treated and untreated groups that are not be fully measured, or, are simply unmeasurable. Different diseases, physicians, hospitals and insurance plans all can introduce different levels of heterogeneity or clustering effects in the treatment assignment. Hence, if not adjusted appropriately, the treatment effect estimated from observational data could be biased due to prognostically important baseline differences among

*Address correspondence to this author at the Department of Biostatistics, University of North Carolina at Chapel Hill, NC 27599-7420, USA; Tel: 919-966-3885; Fax: 919-966-3804; E-mail: zhou@bios.unc.edu

E-ISSN: 1929-6029/14 

patients, along with physicians' knowledge of unmeasured prognostic variables. For example, even though RCT data show prophylaxis reduces VTE risk among medically ill inpatients, attempts to generalize the trial findings to real world patients have been inconclusive [5, 6]. Establishing the equivalence of treatment effect between RCT and observational studies requires statistical methods that properly identify and handle the underlying confounding factors and heterogeneity. Recently, Alemayehu *et al*. [7] reviewed some statistical issues in analyzing non-randomized studies; Willke and Mullins [8] provided a practically useful checklist for achieving the goal of developing credible and germane comparative effectiveness research studies.

The propensity score (PS) method [9] is the most commonly used method to address confounding problems in observational studies in practice. Recent work on PS methods includes Stukel *et al*. [10], Chen *et al*. [11], Hong and Yu [12], Ye and Kaskutas [13], Wyse, Keesler and Schneider [14], Staff *et al*. [15], Maciejewski *et al*. [16], among others. The propensity score, which is generally estimated using logistic regression, is defined as the probability of receiving a treatment conditional on a set of observed covariates, i.e., $PS = \Pr(Trt = 1 \mid X)$.

The validity of the PS method is built on the following two assumptions:

$$\left(Y(1), Y(0)\right) \perp Trt \mid X, \tag{1}$$

$$0 < \Pr(Trt = 1 \mid X) < 1, \tag{2}$$

where $Y(1)$, $Y(0)$ are the outcomes under the active treatment and the control treatment; $Trt$ is an indicator variable denoting the treatment assignment ($Trt = 1$ for active treatment, $Trt = 0$ for control treatment). $X$ is observed baseline covariates. The first condition says that the treatment assignment is independent of the potential outcomes conditional on the observed baseline covariates. Rosenbaum and Rubin [9] had shown that conditional on the propensity score, the distribution of measured baseline covariates is similar between treated and untreated subjects. Thus, for subjects with the same propensity score, the distributions of their baseline covariates will be the same between the treated and untreated subjects. They demonstrated that if treatment assignment is strongly ignorable (conditions (1) and (2)), conditioning on PS, one also has $\left(Y(1), Y(0)\right) \perp Trt$. Hence, conditions (1) and (2) are also referred to as the "no unmeasured confounders" assumption: all variables that affect treatment assignment and outcome have been measured.

The assumption of no unobserved confounding covariates may be too simplistic or ideal in observational study data settings, especially when there is complex heterogeneity or clustering in treatment assignment. For example, different hospitals may have different preferences or limitations in assigning a treatment. Patients with similar demographics and the same insurance plans may get similar treatment assignments. Physicians' skills and personal preferences may introduce another layer of heterogeneity in determining what treatment to give. All those factors could also cause clustering effects in the outcome models. Under complex scenarios, how existing PS methods perform and how to modify them to properly account for heterogeneity remain to be investigated. We conduct an extensive simulation study in this paper to explore the performance of several methods under various levels of heterogeneity in treatment assignment. Methods considered include PS stratification, PS inverse probability weighting, PS regression, PS with random effect and partial least squares (PLS). These methods are selected for their ready availability in standard computing packages and/or their robust nature against underlying model assumptions.

The rest of the article is organized as follows. In Section 2, we introduce the outcome model and several treatment assignment models that include traditional models and more close to realistic complex models. In Section 3, we briefly describe the different methods considered. The simulation study to evaluate these methods under various models is presented in Section 4. The real world data analysis is presented in Section 5. Final remarks are presented in Section 6.

## 2. MODELS FOR SIMULATION STUDY

There are two models that are associated with our study: the outcome model that links treatment with a response and the treatment assignment model that defines the conditional probability of treatment assignment given some covariates. In this Section, we consider one outcome model and three treatment assignment models under complex heterogeneity situations. Suppose $K$ is the number of clusters. To fix notation, let $Y_{ij}$, $Trt_{ij}$ (0 or 1) and $Z_{ij}$ denote the response, the treatment assignment and the covariates of subject $j$ in the $i^{th}$ cluster, respectively, where $i = 1, \ldots, K$; $j = 1, \ldots, n_i$, and $n_i$ is the number of $i^{th}$ cluster.

### 2.1. The Outcome Model

We consider the following outcome model:

$$Y_{ij} = \beta_0 + \alpha Trt_{ij} + \beta_1 Age_{ij} + \beta_2 Sev_{ij} + \beta_3 Test_{ij} + \beta_4 Insu_{ij} + \gamma_1 Z_1 + \cdots + \gamma_{20} Z_{20} + b\eta_i + \varepsilon_{ij} \quad (3)$$

Here $i = 1,\ldots,n$ denotes physician and $j = 1,\ldots,n_i$ denotes the patients of physician $i$. The five named variables are age ($Age$), disease severity ($Sev$), diagnostic test score ($Test$), insurance type ($Insu$, depending on age), and treatment ($Trt$). Specifically, $Age_{ij} \sim Unif(10,60)$, $Sev_{ij}(0,1,2) \sim \text{Mult}(p_0 = 0.49, p_1 = 0.42, p_2 = 0.09)$, $Test_{ij} \sim N(1, 2.25)$. If $Age_{ij} \geq 40$, $Insu_{ij} \sim Bern(1, p = 0.8)$, otherwise, $Insu_{ij} \sim Bern(1, p = 0.2)$. The 20 unnamed covariates $(Z_1,\ldots,Z_{20})$ are nuisance variables and all follow the standard normal distribution. The cluster effect is $b\eta_i$, where $\eta_i \sim N(0,1)$ and $b$ is constant. The random error $\varepsilon_{ij}$ is where $\varepsilon_{ij} \sim N(0,1)$. The primary goal is to estimate $\alpha$, the regression coefficient associated with $Trt$.

### 2.2. Treatment Assignment Models

We consider three random effects models in the treatment assignment. Treatment models 1 and 2, denoted by ($T_1$) and ($T_2$), are simple extensions of standard models used in the PS approach, except that we allow heterogeneity in the models. Both ($T_1$) and ($T_2$) are logistic link functions. In model ($T_1$), we allow the same heterogeneity variable $\eta_i$ in the outcome model (3) to impact the treatment assignment. In model ($T_2$), we introduce an independent random effect $U_i$, where $U_i \perp \eta_i$, in the treatment assignment model. ($T_1$) and ($T_2$) are defined as following:

$$\Pr(Trt_{ij} = 1 \mid X_{ij}, \eta_i) = (1 + \exp(-(\theta_1 Age_{ij} + \theta_2 Sev_{ij} + \theta_3 Test_{ij} + \theta_4 Insu_{ij} + \eta_i)))^{-1}, \ (T_1)$$

$$\Pr(Trt_{ij} = 1 \mid X_{ij}, U_i) = (1 + \exp(-(\theta_1 Age_{ij} + \theta_2 Sev_{ij} + \theta_3 Test_{ij} + \theta_4 Insu_{ij} + U_i)))^{-1}, \ (T_2)$$

where $U_i \sim N(0,1)$, $\eta_i \sim N(0,1)$, $U_i \perp \eta_i$ are cluster effects.

In the third treatment model ($T_3$, also see illustration and diagram in Figure **1**), we mimic the complexity of the real world treatment assignment process and introduce the following treatment assignment scheme where the system (e.g., Insurance), patients and physician's effects collectively determine the treatment assignment. Most importantly, these factors introduce

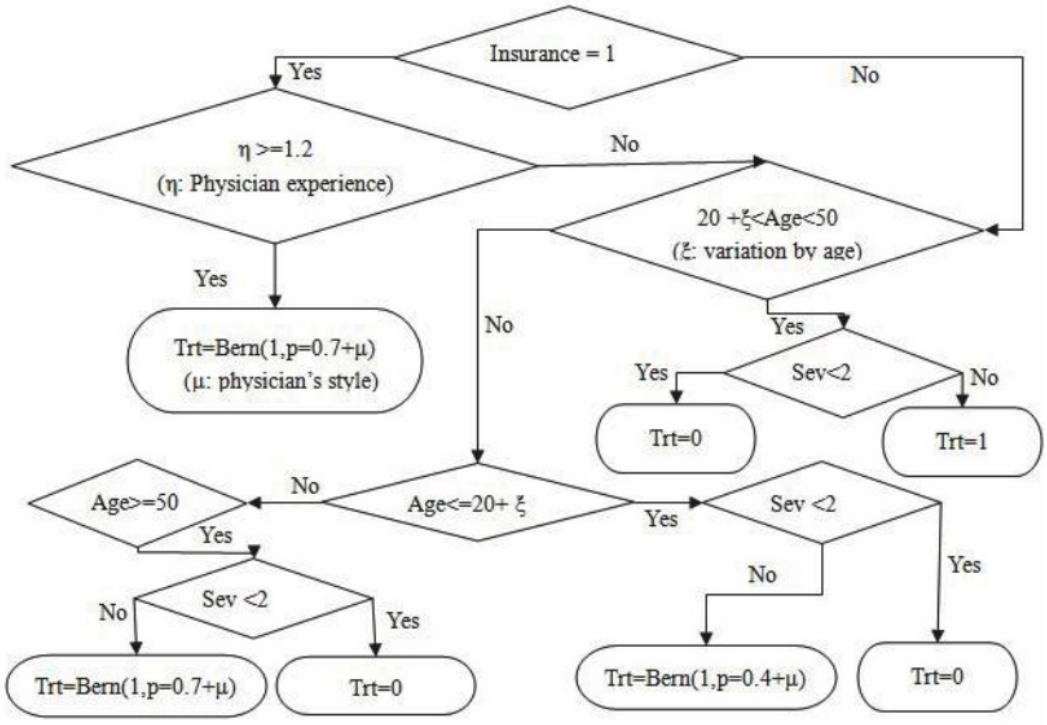


**Figure 1:** Flow chart of treatment assignment model $T_3$.

several unobserved heterogeneity parameters $(\mu,\eta,\xi)$, where $\mu \sim Unif(-0.15,0.15)$ is the random effect for probability, representing random variation of chances in assigning the treatment; $\xi \sim Unif(-5,5)$ is the random effect for patient's age variation in getting the treatment.

Model ($T_3$) below shows this dynamic treatment assignment scheme:

$$Trt_{ij} = \begin{cases} Bern(1, p = 0.7 + \mu_i), Insu_{ij} = 1 \,\&\, \eta_i \geq 1.2 \\ 0, (Insu_{ij} = 0 or \eta_i < 1.2) \,\&\, Age_{ij} \in (20 + \xi_i, 50) \,\&\, Sev_{ij} < 2 \\ 1, (Insu_{ij} = 0 or \eta_i < 1.2) \,\&\, Age_{ij} \in (20 + \xi_i, 50) \,\&\, Sev_{ij} = 2 \\ 0, (Insu_{ij} = 0 or \eta_i < 1.2) \,\&\, Age_{ij} \leq 20 + \xi_i \,\&\, Sev_{ij} < 2 \\ Bern(1, p = 0.4 + \mu_i), (Insu_{ij} = 0 or \eta_i < 1.2) \,\&\, Age_{ij} \\ \leq 20 + \xi_i \,\&\, Sev_{ij} = 2 \\ 0, (Insu_{ij} = 0 or \eta_i < 1.2) \,\&\, Age_{ij} \geq 50 \,\&\, Sev_{ij} < 2 \\ Bern(1, p = 0.7 + \mu_i), (Insu_{ij} = 0 or \eta_i < 1.2) \,\&\, Age_{ij} \\ \geq 50 \,\&\, Sev_{ij} = 2 \end{cases} . \quad (T_3)$$

To help understand why the model in ($T_3$) is closer to a real world data simulation, we provide the following illustration. The above scheme depicts a situation where the physician's skill or practice style (represented by random effect $\eta_i$), the insurance plan coverage, the patient's age (young, middle, and old), and the severity of the patient's condition all affect the treatment assignment. For example, line 1 of ($T_3$) can be interpreted as following: if a doctor is a more aggressive type ($\eta_i \geq 1.2$) and the patient's insurance coverage is good, then the patient is more likely to get the more expensive new treatment, rather than the less expensive traditional care. Here the average chance for getting the treatment is 70%, though different doctors have their own chance, i.e., range 55% ~ 85%, with $\mu_i$'s range considered. The rest of the lines in ($T_3$) further allow patient's age and severity to influence the treatment assignment. Furthermore, we allow doctors to have their own views of old and young ($\xi_i$), as well as the chances of assigning the treatment ($\mu_i$). The link function between $Trt$ and those factors is not a logistic function in contrast to the link function in ($T_1$) and ($T_2$). We believe this model is more realistic and closer to a real world treatment assignment process than the simple logistic link function, which is chosen mainly for its mathematical convenience.

## 3. METHODS CONSIDERED IN SIMULATION STUDY

We will make inference on the data from three treatment assignment models in Section 2 with the PS stratification method, PS inverse probability weighting, PS regression, and its extensions under random effects. We also consider the partial least squares methods. Under the PS methods framework, how to model the clustered data is an open question. We consider several possible approaches, including ignoring the heterogeneity, adjusting the heterogeneity by dummy variables, and a mixed effects model.

### PS Methods Under Heterogeneity

Define $Dummy_{ij} = 1$ if subject $j$ is in the $i^{th}$ cluster, and $Dummy_{ij} = 0$, otherwise. We proceed with the existing PS in the following three ways:

(1) ignoring the heterogeneity (denoted by $PS_I$) and using the standard logistic regression based PS method

$$Logit(\Pr(Trt_{ij} = 1 \mid X_{ij})) = \theta_0 + \theta_1 X_{ij}; \quad (4)$$

(2) using a dummy variable method (denoted by $PS_D$) to adjust for heterogeneity

$$Logit(\Pr(Trt_{ij} = 1 \mid X_{ij}, Dummy_i)) = \theta_0 + \theta_1 X_{ij} + \theta_2 Dummy_i; \quad (5)$$

(3) using a mixed effects model (denoted by $PS_M$) for heterogeneity

$$Logit(\Pr(Trt_{ij} = 1 \mid X_{ij}, \zeta_i)) = \theta_0 + \theta_1 X_{ij} + \zeta_i, \quad (6)$$

where $\zeta_i$ is a random cluster effect.

After fitting PS from one of three models above, we can proceed to estimate the treatment effect ($\alpha$) by the following methods.

### PS Stratification

The PS stratification method is to stratify subjects into mutually exclusive subsets based on their estimated propensity score. Within each stratum, the treatment effect can be estimated by comparing outcomes between treated and untreated subjects. The stratum-specific treatment effects are weighted by the proportion of subjects lying within that stratum to obtain the overall treatment effect. In this article, we divide subjects into four equal-size groups using the quantiles of the estimated propensity score. Let denote the

estimated treatment effect from the stratification method with the propensity score estimated from models (4), (5) and (6), respectively.

### PS Inverse Probability Weight

The PS inverse probability weight method is to use inverse of the propensity score as weight to create a synthetic sample in which the distribution of measure baseline covariates is independent of treatment assignment. The weight can be defined as following:

$$W_{ij} = \frac{Trt_{ij}}{PS_{ij}} + \frac{1 - Trt_{ij}}{1 - PS_{ij}}.$$

The treatment effect can be estimated by

$$\frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{n_i}\frac{Y_{ij}Trt_{ij}}{PS_{ij}} - \frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{n_i}\frac{Y_{ij}(1-Trt_{ij})}{1-PS_{ij}},$$

where $N = \sum_{j=1}^{K} n_j$ is the total number of subjects. Let $\hat{\alpha}_{IPW\cdot I}$, $\hat{\alpha}_{IPW\cdot D}$ and $\hat{\alpha}_{IPW\cdot M}$ denote the estimated treatment effects from the inverse probability weight method with the propensity score estimated from models (4), (5) and (6), respectively.

### PS Regression

The PS regression method is to estimate the treatment effect ($\alpha$) from one of following three regression models.

(a) Ignoring heterogeneity model:

$$Y_{ij} = \beta_0 + \alpha Trt_{ij} + \beta_1 \widehat{PS} + e_{ij}, \quad (7)$$

(b) Using dummy variable adjustment model:

$$Y_{ij} = \beta_0 + \alpha Trt_{ij} + \beta_1 \widehat{PS} + \beta_2 Dummy_i + e_{ij}, \quad (8)$$

(c) Using mixed model:

$$Y_{ij} = \beta_0 + \alpha Trt_{ij} + \beta_1 \widehat{PS} + \varphi_i + e_{ij}, \quad (9)$$

where $\varphi_i$ is the cluster effect.

In the PS regression, we have 9 estimators from the factorial combination of three treatment models (models 4-6) and three outcome models (models 7-9). These estimators are denoted as three class estimators $\left(\hat{\alpha}_{I\cdot I}, \hat{\alpha}_{I\cdot D}, \hat{\alpha}_{I\cdot M}\right)$, $\left(\hat{\alpha}_{D\cdot I}, \hat{\alpha}_{D\cdot D}, \hat{\alpha}_{D\cdot M}\right)$, $\left(\hat{\alpha}_{D\cdot I}, \hat{\alpha}_{D\cdot D}, \hat{\alpha}_{D\cdot M}\right)$, where the first letter in the subscript represents how heterogeneity is handled in calculating PS scores and the second letter in the subscript represents how the heterogeneity in the outcome model is handled. Table **1** below summarizes these estimators.

**Table 1: Extending Existing PS Method to Handle Heterogeneity**

| Treatment Model | Outcome Model | | |
|---|---|---|---|
| | Ignore | Mixed | Dummy |
| Ignore | $\hat{\alpha}_{I\cdot I}$ | $\hat{\alpha}_{I\cdot M}$ | $\hat{\alpha}_{I\cdot D}$ |
| Mixed | $\hat{\alpha}_{M\cdot I}$ | $\hat{\alpha}_{M\cdot M}$ | $\hat{\alpha}_{M\cdot D}$ |
| Dummy | $\hat{\alpha}_{D\cdot I}$ | $\hat{\alpha}_{D\cdot M}$ | $\hat{\alpha}_{D\cdot D}$ |

### Partial Least Squares (PLS)

PLS is a wide class of methods for modeling relations between sets of observed variables by means of latent variables [17-19]. In its general form, PLS creates orthogonal score vectors (also called latent vectors or components) by maximizing the covariance between different sets of variables. PLS can be naturally extended to regression problems. The predictor and predicted (response) variables are each considered as a block of variables. PLS then extracts the score vectors which serve as a new predictor representation and regresses the response variables on these new predictors. PLS does not depend on the model specification, hence is robust to nonlinear, clustering, and interactions. We consider the following PLS algorithm. We first center the covariates matrix $X$ and response matrix $Y$, then set $u$ to the first column of $Y$ and repeat a sequence of the following steps until convergence. 1. $w = X^T u / \left(u^T u\right)$, coefficients of regressing $u$ on $X$; 2. $\|w\| \to 1$, scale $w$ to be of length one; 3. $t = Xw$; 4. $c = Y^T t / \left(t^T t\right)$, coefficients of regressing $t$ on $Y$; 5. $\|c\| \to 1$, scale $c$ to be of length one; 6. $u = Yc$. If $Y$ is a vector, $u$ is equal to $Y$.

It can be shown that the weight vector $w$ corresponds to the first eigenvector of the following eigenvalue problem:

$$X^T Y Y^T X w = \lambda w.$$

The latent variables $t$, $u$ are then given as $t = Xw$ and $u = Yc$. The vectors of loadings $p$ and $q$ are computed as coefficients of regressing $X$ on $t$ and $Y$ on $u$, respectively. That is, $p = X^T t / \left(t^T t\right)$ and $q = Y^T u / \left(u^T u\right)$. After the extraction of the score vectors $t$ and $u$, the matrix $X$ and $Y$ are deflated as $X = X - tp^T$ and

$Y = Y - uq^T$. If $Y$ is a vector, we need not subtract above approximation from $Y$. The iterations can continue until a stopping criteria is met or $X$ becomes the zero matrix. For the PLS method, we let $\widehat{\alpha}_{PLS\cdot I}$, $\widehat{\alpha}_{PLS\cdot D}$, $\widehat{\alpha}_{PLS\cdot M}$ denote three corresponding PLS estimators under the three outcome models (7-9), respectively.

**Intuitive Version of Partial Least Squares**

The intention of PLS is to form a set of components (named latent variable) that capture most of the information in the $X$ variables that is useful for predicting response, while reducing the dimensionality of the regression problems by using fewer components than the number of $X$ variables. While the description we presented above may not be immediately intuitive, Garthwaite [20] provided a more intuitive and interpretable version of PLS, simplified to the case of just one response variable $Y$. Under Garthwaite's framework of PLS, it is easily seen that the PLS is robust to not only the underlying model link structure, but also to the form of the covariates. His version of PLS is

1. 1st component: $S_1$ and $V_{11},\ldots,V_{1m}$ are centered response $Y$ and covariates $X_1,\ldots,X_m$, respectively. Perform univariate regressions $\widehat{S}_1 = b_{1j}V_{1j}, j = 1,\ldots,m$. The first component $L_1$ is constructed as $L_1 = \sum_{j=1}^{m} W_{1j}b_{1j}V_{1j}$, where $W_{1j}$ are pre-determined weight functions;
2. 2nd component: run a univariate regression with $L_1$ being the covariate and $S_1$, $X_1,\ldots,X_m$ being response. Denote the residuals by $S_2$, $V_{21},\ldots,V_{2m}$, respectively. Similarly, perform univariate regressions: $S_2 = b_{2j}V_{2j}, j = 1,\ldots,m$. Then $L_2$ is constructed as $L_2 = \sum_{j=1}^{m} W_{2j}b_{2j}V_{2j}$, where $W_{2j}$ are pre-determined weight functions;
3. 3rd component: run a univariate regression with $L_1$ and $L_2$ being the covariate and $S_1$, $X_1,\ldots,X_m$ being response. Denote the residuals by $S_3$, $V_{31},\ldots,V_{3m}$, respectively. Similarly, perform univariate regressions: $\widehat{S}_3 = b_{3j}V_{3j}, j = 1,\ldots,m$. Then $L_3$ is constructed as $L_3 = \sum_{j=1}^{m} W_{3j}b_{3j}V_{3j}$, where $W_{3j}$ are pre-determined weight functions;
4. Subsequent components are constructed likewise.

From Garthwaite's interpretation, we can see that the PLS components are obtained without any model specification. Further, inspecting the key step 3 reveals that the subsequent components are computed from residuals of regressing the original $Y$ and $X$ on the already derived components i.e., these components are like "orthogonal" components and they are robust to potential nonlinearity, clustering, interactions, and complex heterogeneity existing in the treatment assignment.

## 4. SIMULATION STUDY

We generate the data from the outcome model (3):

$$Y_{ij} = \beta_0 + \alpha Trt_{ij} + \beta_1 Age_{ij} + \beta_2 Sev_{ij} + \beta_3 Test_{ij} + \beta_4 Insu_{ij} + \gamma_1 Z_1 + \cdots + \gamma_{20} Z_{20} + b\eta_i + \varepsilon_{ij}.$$

We set $\beta_0 = 0$, $\beta_1 = 0.2$, $\beta_2 = 0.4$, $\beta_3 = 0.7$, $\beta_4 = 0.3$; and $b = 0,1$, $\alpha = 0, 0.5$, respectively. We set $\gamma_1 = \cdots = \gamma_{20} = 0$. We randomly generate 500 data sets, each with $n = 100$ (i.e., 100 physicians), $n_i = 20 \sim 80$ (i.e., 20-80 patients per doctor) for a total patient number $N = 4700$. The treatment assignments are generated from one of the three models, ($T_1$), ($T_2$) and ($T_3$), respectively. We present the result for $\Pr(Trt = 1) = 0.1$ and leave the result for $\Pr(Trt = 1) = 0.5$ as an online supplement. We use the R packages "glmmML" (generalized linear mixed model) to estimated propensity score, and "lmm" (linear mixed model) to estimate treatment effect, respectively. In the propensity score method, the covariates $X$ from $PS = \Pr(Trt = 1 \mid X)$ is the matrix $(Age, Sev, Test, Insu, Z_1, \ldots Z_{20})$. We also use the package "pls" (PLS method) to compute the latent variables. In the PLS method, we use $Trt$ as the response and $X$ as the covariates matrix, then we can obtain the latent variables matrix $W$. In this article, the latent variables matrix $W$ is constituted by 5 PLS components. Then we can obtain the treatment effect estimates from the regression $Y$ on $Trt$ and $W$. Below, we present our results in the order of ($T_1$), ($T_2$) to ($T_3$).

**Treatment Assignment Generated From Model ($T_1$)**

We choose the parameters $\theta_1 = -0.1(Age)$, $\theta_2 = 0.1(Sev)$, $\theta_3 = 0.2(Test)$, $\theta_4 = 0.3(Insu)$ to get

**Table 2: Treatment assignment model ($T_1$)**

| Method | $b = 0$ | | | | $b = 1$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $\alpha = 0$ | | $\alpha = 0.5$ | | $\alpha = 0$ | | $\alpha = 0.5$ | |
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| $\hat{\alpha}_{S\cdot I}$ | -0.100 | 0.193 | 0.400 | 0.193 | -0.104 | 0.236 | 0.396 | 0.236 |
| $\hat{\alpha}_{S\cdot M}$ | 0.140 | 0.066 | 0.640 | 0.066 | -0.121 | 0.230 | 0.379 | 0.230 |
| $\hat{\alpha}_{S\cdot D}$ | -0.032 | 0.443 | 0.468 | 0.443 | -0.028 | 0.463 | 0.472 | 0.463 |
| $\hat{\alpha}_{IPW\cdot I}$ | -0.257 | 0.525 | 0.232 | 0.545 | -0.246 | 0.540 | 0.225 | 0.560 |
| $\hat{\alpha}_{IPW\cdot M}$ | 3.979 | 1.504 | 4.469 | 1.576 | 3.969 | 1.515 | 4.483 | 1.586 |
| $\hat{\alpha}_{IPW\cdot D}$ | -0.470 | 1.403 | -0.000 | 1.464 | -0.466 | 1.417 | 0.004 | 1.477 |
| $\hat{\alpha}_{I\cdot I}$ | -0.025 | 0.060 | 0.475 | 0.060 | 0.844 | 0.142 | 1.344 | 0.142 |
| $\hat{\alpha}_{I\cdot M}$ | -0.026 | 0.061 | 0.474 | 0.061 | 0.105 | 0.081 | 0.605 | 0.081 |
| $\hat{\alpha}_{I\cdot D}$ | -0.038 | 0.080 | 0.462 | 0.080 | -0.038 | 0.080 | 0.462 | 0.080 |
| $\hat{\alpha}_{M\cdot I}$ | -0.189 | 0.087 | 0.311 | 0.087 | 0.679 | 0.135 | 1.179 | 0.135 |
| $\hat{\alpha}_{M\cdot M}$ | -0.190 | 0.087 | 0.310 | 0.087 | -0.058 | 0.086 | 0.442 | 0.086 |
| $\hat{\alpha}_{M\cdot D}$ | -0.207 | 0.088 | 0.293 | 0.088 | -0.207 | 0.088 | 0.293 | 0.088 |
| $\hat{\alpha}_{D\cdot I}$ | 0.001 | 0.062 | 0.501 | 0.062 | 0.001 | 0.057 | 0.499 | 0.057 |
| $\hat{\alpha}_{D\cdot M}$ | 0.002 | 0.068 | 0.502 | 0.068 | 0.002 | 0.068 | 0.502 | 0.068 |
| $\hat{\alpha}_{D\cdot D}$ | 0.003 | 0.070 | 0.503 | 0.070 | 0.003 | 0.070 | 0.503 | 0.070 |
| $\hat{\alpha}_{PLS\cdot I}$ | -0.004 | 0.051 | 0.496 | 0.051 | 0.840 | 0.139 | 1.340 | 0.139 |
| $\hat{\alpha}_{PLS\cdot M}$ | -0.004 | 0.051 | 0.496 | 0.051 | 0.016 | 0.054 | 0.516 | 0.054 |
| $\hat{\alpha}_{PLS\cdot D}$ | -0.003 | 0.054 | 0.497 | 0.054 | -0.003 | 0.054 | 0.497 | 0.054 |

Note: $(\hat{\alpha}_{S\cdot I},\hat{\alpha}_{S\cdot D},\hat{\alpha}_{S\cdot M})$ : The PS stratification estimates. $(\hat{\alpha}_{IPW\cdot I},\hat{\alpha}_{IPW\cdot D},\hat{\alpha}_{IPW\cdot M})$ : PS inverse probability weight estimates. $\hat{\alpha}_{\cdot\cdot}$ PS regression with PS from the model of using ignoring (I), dummy variable (D) and mixed model (M) to adjust the heterogeneity in propensity score and the second subscript of $\hat{\alpha}$ denotes the response from the model of using ignoring (I), dummy variable (D) and mixed model (M) to adjust the heterogeneity in response. $\hat{\alpha}_{PLS\cdot}$ : the PLS method and the second denotes the response model.

$\Pr(Trt=1)=0.1$. The simulation results are summarized in Table **2**.

Inspecting Table **2** reveals that when $b=0$, i.e., when there is no random effect in the outcome model, we have:

(i) $(\hat{\alpha}_{S\cdot I},\hat{\alpha}_{S\cdot M})$, $(\hat{\alpha}_{IPW\cdot I},\hat{\alpha}_{IPW\cdot D},\hat{\alpha}_{IPW\cdot M})$ and $(\hat{\alpha}_{M\cdot I},\hat{\alpha}_{M\cdot D},\hat{\alpha}_{M\cdot M})$ estimators are biased, $\hat{\alpha}_{S\cdot D}$ is unbiased;

(ii) $\hat{\alpha}_{PLS\cdot I}$, $\hat{\alpha}_{PLS\cdot D}$ and $\hat{\alpha}_{PLS\cdot M}$ are unbiased and are the most efficient estimators for $\alpha$.

(iii) $(\hat{\alpha}_{D\cdot I},\hat{\alpha}_{D\cdot D},\hat{\alpha}_{D\cdot M})$ and $(\hat{\alpha}_{I\cdot I},\hat{\alpha}_{I\cdot D},\hat{\alpha}_{I\cdot M})$ are all unbiased but are less efficient than PLS estimators.

When $b=1$, i.e., when $\eta$ appears in both the treatment and the outcome model, where the physician's skill level affects both treatment choice and outcome.

(i) $(\hat{\alpha}_{I\cdot I},\hat{\alpha}_{I\cdot D},\hat{\alpha}_{I\cdot M})$ class estimators now become biased;

(ii) $(\hat{\alpha}_{S\cdot I},\hat{\alpha}_{S\cdot M})$, $(\hat{\alpha}_{IPW\cdot I},\hat{\alpha}_{IPW\cdot D},\hat{\alpha}_{IPW\cdot M})$ and $(\hat{\alpha}_{M\cdot I},\hat{\alpha}_{M\cdot D},\hat{\alpha}_{M\cdot M})$ estimators are still biased, $\hat{\alpha}_{S\cdot D}$ is still unbiased;

(iii) $\left(\hat{\alpha}_{D\cdot I},\hat{\alpha}_{D\cdot D},\hat{\alpha}_{D\cdot M}\right)$ class estimators are again unbiased and still less efficient than $\left(\hat{\alpha}_{PLS\cdot I},\hat{\alpha}_{PLS\cdot D},\hat{\alpha}_{PLS\cdot M}\right)$ class;

(iv) among $\left(\hat{\alpha}_{PLS\cdot I},\hat{\alpha}_{PLS\cdot D},\hat{\alpha}_{PLS\cdot M}\right)$ class estimators, $\hat{\alpha}_{PLS\cdot I}$ is now biased, but $\hat{\alpha}_{PLS\cdot D}$ and $\hat{\alpha}_{PLS\cdot M}$ are still the most efficient estimators.

### Treatment Assignment Generated from Model ( $T_2$ )

We choose the parameters $\theta_1=-0.1\left(Age\right)$, $\theta_2=0.1\left(Sev\right)$, $\theta_3=0.2\left(Test\right)$, $\theta_4=0.3\left(Insu\right)$ to get $\Pr\left(Trt=1\right)=0.1$. The simulation results are summarized in Table **3**.

Here, the random effect in the treatment model is independent of that in the outcome model, i.e., $U_i \perp \eta_i$.

The basic observation outlined in Table **2** mostly stay the same. The only difference is that $\hat{\alpha}_{PLS\cdot I}$ is now unbiased in Table **3**, even when $b=1$ and $\hat{\alpha}_{S\cdot D}$ is biased, when $b=1$. Again, $\left(\hat{\alpha}_{D\cdot I},\hat{\alpha}_{D\cdot D},\hat{\alpha}_{D\cdot M}\right)$ class estimators are pretty robust and unbiased, yet they are not as efficient as $\hat{\alpha}_{PLS\cdot D}$ and $\hat{\alpha}_{PLS\cdot M}$.

### Treatment Assignment Generated from Model ( $T_3$ )

Again, $\Pr\left(Trt=1\right)=0.1$ in model ( $T_3$ ). The other parameter values are given in ( $T_3$ ). The simulation results are presented in Table **4**.

Table **4** results reveal that:

(i) $\left(\hat{\alpha}_{S\cdot I},\hat{\alpha}_{S\cdot M}\right)$, $\left(\hat{\alpha}_{IPW\cdot I},\hat{\alpha}_{IPW\cdot D},\hat{\alpha}_{IPW\cdot M}\right)$ and $\left(\hat{\alpha}_{I\cdot I},\hat{\alpha}_{I\cdot D},\hat{\alpha}_{I\cdot M}\right)$ class estimators are all biased;

**Table 3: Treatment Assignment Model ($T_2$)**

| Method | $b$ = 0 | | | | $b$ = 1 | | | |
|---|---|---|---|---|---|---|---|---|
| | $\alpha$ = 0 | | $\alpha$ = 0.5 | | $\alpha$ = 0 | | $\alpha$ = 0.5 | |
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| $\hat{\alpha}_{S\cdot I}$ | -0.100 | 0.193 | 0.400 | 0.193 | 0.796 | 0.244 | 1.296 | 0.244 |
| $\hat{\alpha}_{S\cdot M}$ | -0.117 | 0.190 | 0.383 | 0.190 | 0.782 | 0.244 | 1.282 | 0.244 |
| $\hat{\alpha}_{S\cdot D}$ | -0.032 | 0.443 | 0.468 | 0.443 | 0.136 | 0.470 | 0.636 | 0.470 |
| $\hat{\alpha}_{IPW\cdot I}$ | -0.257 | 0.524 | 0.232 | 0.545 | 0.629 | 0.582 | 1.118 | 0.602 |
| $\hat{\alpha}_{IPW\cdot M}$ | 3.979 | 1.504 | 4.694 | 1.576 | 5.280 | 1.790 | 5.994 | 1.862 |
| $\hat{\alpha}_{IPW\cdot D}$ | -0.470 | 1.403 | -0.000 | 1.464 | -0.380 | 1.309 | 0.090 | 1.370 |
| $\hat{\alpha}_{I\cdot I}$ | -0.025 | 0.060 | 0.475 | 0.060 | -0.030 | 0.126 | 0.470 | 0.126 |
| $\hat{\alpha}_{I\cdot M}$ | -0.026 | 0.061 | 0.474 | 0.061 | -0.036 | 0.077 | 0.464 | 0.077 |
| $\hat{\alpha}_{I\cdot D}$ | -0.038 | 0.080 | 0.462 | 0.080 | -0.038 | 0.080 | 0.462 | 0.080 |
| $\hat{\alpha}_{M\cdot I}$ | -0.196 | 0.083 | 0.304 | 0.083 | -0.201 | 0.136 | 0.299 | 0.136 |
| $\hat{\alpha}_{M\cdot M}$ | -0.197 | 0.083 | 0.303 | 0.083 | -0.210 | 0.085 | 0.290 | 0.085 |
| $\hat{\alpha}_{M\cdot D}$ | -0.213 | 0.086 | 0.287 | 0.086 | -0.213 | 0.086 | 0.287 | 0.086 |
| $\hat{\alpha}_{D\cdot I}$ | 0.001 | 0.062 | 0.501 | 0.062 | 0.001 | 0.062 | 0.501 | 0.062 |
| $\hat{\alpha}_{D\cdot M}$ | 0.002 | 0.068 | 0.502 | 0.068 | 0.002 | 0.068 | 0.502 | 0.068 |
| $\hat{\alpha}_{D\cdot D}$ | 0.003 | 0.070 | 0.503 | 0.070 | 0.003 | 0.070 | 0.503 | 0.070 |
| $\hat{\alpha}_{PLS\cdot I}$ | -0.004 | 0.051 | 0.496 | 0.051 | 0.002 | 0.120 | 0.502 | 0.120 |
| $\hat{\alpha}_{PLS\cdot M}$ | -0.004 | 0.051 | 0.496 | 0.051 | -0.003 | 0.054 | 0.497 | 0.054 |
| $\hat{\alpha}_{PLS\cdot D}$ | -0.003 | 0.054 | 0.497 | 0.054 | -0.003 | 0.054 | 0.497 | 0.054 |

Note: The notations are the same as in Table **2**.

**Table 4: Treatment Assignment Model ($T_3$)**

| Method | $b = 0$ | | | | $b = 1$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $\alpha = 0$ | | $\alpha = 0.5$ | | $\alpha = 0$ | | $\alpha = 0.5$ | |
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| $\hat{\alpha}_{S\cdot I}$ | -0.184 | 0.949 | 0.316 | 0.949 | 1.200 | 0.992 | 1.700 | 0.992 |
| $\hat{\alpha}_{S\cdot M}$ | -0.186 | 0.920 | 0.314 | 0.920 | 1.209 | 0.938 | 1.709 | 0.938 |
| $\hat{\alpha}_{S\cdot D}$ | 0.543 | 0.492 | 1.043 | 0.492 | 1.480 | 0.504 | 1.980 | 0.504 |
| $\hat{\alpha}_{IPW\cdot I}$ | 13.689 | 12.162 | 14.856 | 12.818 | 17.480 | 14.387 | 18.647 | 15.044 |
| $\hat{\alpha}_{IPW\cdot M}$ | 85.691 | 103.801 | 90.806 | 109.566 | 102.685 | 121.762 | 107.800 | 127.532 |
| $\hat{\alpha}_{IPW\cdot D}$ | -8.513 | 4.037 | -8.294 | 4.446 | -8.751 | 5.027 | -8.532 | 5.037 |
| $\hat{\alpha}_{I\cdot I}$ | -0.292 | 0.096 | 0.208 | 0.096 | 0.774 | 0.165 | 1.274 | 0.165 |
| $\hat{\alpha}_{I\cdot M}$ | -0.294 | 0.097 | 0.206 | 0.097 | -0.085 | 0.147 | 0.415 | 0.147 |
| $\hat{\alpha}_{I\cdot D}$ | -0.343 | 0.159 | 0.157 | 0.159 | -0.343 | 0.159 | 0.157 | 0.159 |
| $\hat{\alpha}_{M\cdot I}$ | -0.015 | 0.165 | 0.485 | 0.165 | 1.090 | 0.206 | 1.590 | 0.206 |
| $\hat{\alpha}_{M\cdot M}$ | -0.015 | 0.164 | 0.485 | 0.164 | 0.246 | 0.173 | 0.746 | 0.173 |
| $\hat{\alpha}_{M\cdot D}$ | -0.020 | 0.167 | 0.480 | 0.167 | -0.020 | 0.167 | 0.480 | 0.167 |
| $\hat{\alpha}_{D\cdot I}$ | -0.006 | 0.085 | 0.494 | 0.085 | -0.006 | 0.096 | 0.494 | 0.096 |
| $\hat{\alpha}_{D\cdot M}$ | -0.005 | 0.086 | 0.495 | 0.086 | -0.005 | 0.091 | 0.495 | 0.091 |
| $\hat{\alpha}_{D\cdot D}$ | -0.004 | 0.090 | 0.496 | 0.090 | -0.004 | 0.090 | 0.496 | 0.090 |
| $\hat{\alpha}_{PLS\cdot I}$ | -0.003 | 0.055 | 0.497 | 0.055 | 0.733 | 0.158 | 1.233 | 0.158 |
| $\hat{\alpha}_{PLS\cdot M}$ | -0.003 | 0.055 | 0.497 | 0.055 | 0.015 | 0.059 | 0.515 | 0.059 |
| $\hat{\alpha}_{PLS\cdot D}$ | -0.003 | 0.059 | 0.497 | 0.059 | -0.003 | 0.059 | 0.497 | 0.059 |

Note: The notations are the same as in Table **2**.

(ii) $\left(\hat{\alpha}_{M\cdot I}, \hat{\alpha}_{M\cdot M}\right)$ are biased under $b=0$, however they are biased under $b=1$ and the estimator $\hat{\alpha}_{M\cdot D}$ is unbiased;

(iii) $\left(\hat{\alpha}_{D\cdot I}, \hat{\alpha}_{D\cdot D}, \hat{\alpha}_{D\cdot M}\right)$ class estimators are consistently robust estimator for $\alpha$;

(iv) $\hat{\alpha}_{PLS\cdot D}$ and $\hat{\alpha}_{PLS\cdot M}$ are unbiased and are the most efficient estimators among all.

The study of different model combinations above shows that $\left(\hat{\alpha}_{IPW\cdot I}, \hat{\alpha}_{IPW\cdot D}, \hat{\alpha}_{IPW\cdot M}\right)$ are biased estimators where there is heterogeneity in the treatment and outcome models. $\hat{\alpha}_{PLS}$ class estimators, especially the $\hat{\alpha}_{PLS\cdot D}$ and $\hat{\alpha}_{PLS\cdot M}$ estimators are unbiased and most efficient among all estimators considered. Perhaps most intriguing are the results for the $\left(\hat{\alpha}_{D\cdot I}, \hat{\alpha}_{D\cdot D}, \hat{\alpha}_{D\cdot M}\right)$ class estimator. They are surprisingly unbiased and robust, not too far below the $\hat{\alpha}_{PLS}$ class estimators in term of efficiency.

## 5. REAL DATA ANALYSIS

The German Breast Cancer Study Group (GBSG) study was conducted between November 1983 and November 1989 to evaluate the treatment effect of breast preservation compared with radical breast cancer surgery, under non-randomized, real-world treatment conditions. There were 63 university and community hospitals and the total number of the women patients with breast cancer of

pathophysiological tumor stage pT1 N0 M0 was 646 in the GBSG study [21, 22]. Our main purpose is to evaluate the performance status ( $PST$ ) in relation to the two treatment modalities, which are simple mastectomy ( $Trt=0$ ) and lumpectomy (BC, breast conservation, $Trt=1$ ). $PST$ is a score between 0 and 100 comprising several items of the quality of life (QoL) questionnaire, with higher scores reflecting better QoL. Patient age ( $Age$ ) and tumor size ( $Tmass$ ) are considered as possible confounders.

We consider the following response model:

$$PST_{ij} = \beta_0 + \alpha Trt_{ij} + \beta_1 Age_{ij} + \beta_2 Tmass_{ij} + \beta_3 Age_{ij} \cdot Tmass_{ij} + \gamma_1 Z_{ij1} + \cdots + \gamma_{15} Z_{ij15} + \varepsilon_{ij},$$

where $i$ denotes the hospital, $j$ denotes the patient in the $i^{th}$ hospital and $\varepsilon$ is a random error. The covariates $Z_1,\ldots,Z_{10}$ each has a standard normal distribution, and $Z_{11},\ldots,Z_{15}$ are discrete variable with corresponding distribution being $Z_{11} \sim Bern(0.4)$, $Z_{12}(0,1,2) \sim Mult(0.4,0.5,0.1)$, $Z_{13} = I(W_1 > 0)$ with $W_1 \sim N(0.2,1)$, $Z_{14} = I(W_2 > 0.2)$ with $W_2 \sim N(0.5,1.5)$ and $Z_{15}(0,1,2,3,4) \sim Mult(0.2,0.3,0.2,0.2,0.1)$, respectively. The covariates $Z_1,\ldots,Z_{15}$ are noise variables and $\gamma_1 = \cdots = \gamma_{15} = 0$.

We consider the models (4), (5) and (6) in Section 3 to estimate propensity scores with $X = (Age, Tmass, Age \cdot Tmass, Z_1,\ldots,Z_{15})$. We consider four methods to estimate the treatment effect: PS stratification, PS inverse probability weighting, PS regression and partial least squares methods. We use the bootstrap method to obtain variance estimate and the number of bootstrap iterations is 300. We use the cross-validation method to select the number of PLS components. The results are summarized in Table **5**.

Inspecting results in Table **5**, we observed similar conclusion to our simulation results, i.e. the IPW class estimators of the main effect estimate $(\hat{\alpha}_{IPW\cdot I}, \hat{\alpha}_{IPW\cdot D}, \hat{\alpha}_{IPW\cdot M})$ is biased. The fact that the IPW-D and IPW-M are so far apart and one is a positive effect (24.143) and the other one is a negative (-27.353 and significant) also suggests that the estimated results from this method cannot be trusted. The PS stratification class estimator and the PS regression class estimator yield about the same order of magnitude of the effect size as the proposed PLS estimator. Inspecting the column 3 in Table **5**, our proposed partial least squares method has the smallest variance 1.302 and a more precise 95% confidence interval (-2.158, 2.945) compared with other three methods. This suggest that in this real data analysis, the PLS method is more efficient than the other methods. We would conclude based on the proposed method that the proposed partial least squares method confirms that lumpectomy treatment is not significantly

**Table 5: Analysis Results for German Breast Cancer Study Group Study**

| Method | $\hat{\alpha}$ | SE( $\hat{\alpha}$ ) | 95% CI |
|---|---|---|---|
| $\hat{\alpha}_{S\cdot I}$ | -0.169 | 1.416 | (-2.944, 2.607) |
| $\hat{\alpha}_{S\cdot M}$ | -0.251 | 1.672 | (-3.439,2.937) |
| $\hat{\alpha}_{S\cdot D}$ | 2.943 | 2.666 | (-2.283, 8.170) |
| $\hat{\alpha}_{IPW\cdot I}$ | 1.612 | 3.257 | (-4.772, 7.997) |
| $\hat{\alpha}_{IPW\cdot M}$ | -27.353 | 24.189 | (-74.763, 20.056) |
| $\hat{\alpha}_{IPW\cdot D}$ | 24.143 | 7.496 | (9.450, 38.836) |
| $\hat{\alpha}_{R\cdot I}$ | 0.489 | 1.396 | (-2.246,3.225) |
| $\hat{\alpha}_{R\cdot M}$ | 0.416 | 1.376 | (-2.263,3.094) |
| $\hat{\alpha}_{R\cdot D}$ | 0.580 | 1.779 | (-2.906,4.067) |
| $\hat{\alpha}_{PLS\cdot 2}$ | 0.393 | 1.302 | (-2.158,2.945) |

Note: $(\hat{\alpha}_{S\cdot I}, \hat{\alpha}_{S\cdot D}, \hat{\alpha}_{S\cdot M})$ : The PS stratification estimates. $(\hat{\alpha}_{IPW\cdot I}, \hat{\alpha}_{IPW\cdot D}, \hat{\alpha}_{IPW\cdot M})$ : PS inverse probability weight estimates. $(\hat{\alpha}_{R\cdot I}, \hat{\alpha}_{R\cdot D}, \hat{\alpha}_{R\cdot M})$ : PS regression estimators. $\hat{\alpha}_{PLS\cdot 2}$ : partial least squares estimator by using 2 components.

related to improved performance status, but has a nominal positive impact on it.

## 6. DISCUSSION AND CONCLUDING REMARKS

In this article, we designed several treatment assignment mechanisms that we believe to be more realistic and closer to real world medical practice than the commonly used standard logistic model. Treatment selection in observational studies is a complex process that involves many factors derived from patients, disease status, health care providers, and health care systems. Incompleteness in measuring the underlying treatment assignment process can create heterogeneity that is complex and intractable, and could bias the estimate of the true treatment effects. We extended the traditional PS method to allow different ways of handling random effects (see Table **1**) and consider the traditional PS stratification, PS inverse probability weight and PS regression methods. We also considered that non model-dependent approaches, the PLS methods under the data generated. All methods are used to analyze the same data set generated. To our knowledge, our paper is the first to compare these methods under real world heterogeneity situations.

Overall, we found that two classes of estimators are unbiased under all situations studied: the class of estimators $\left(\hat{\alpha}_{D\cdot I},\hat{\alpha}_{D\cdot D},\hat{\alpha}_{D\cdot M}\right)$ where the dummy variable method was used in the treatment model and the class of estimators based on PLS, namely, $\hat{\alpha}_{PLS\cdot D}$ and $\hat{\alpha}_{PLS\cdot M}$. The best performance goes to the class of estimators based on PLS, as they are more efficient than the $\left(\hat{\alpha}_{D\cdot I},\hat{\alpha}_{D\cdot D},\hat{\alpha}_{D\cdot M}\right)$ class of estimators. The $\left(\hat{\alpha}_{IPW\cdot I},\hat{\alpha}_{IPW\cdot D},\hat{\alpha}_{IPW\cdot M}\right)$ do not yield unbiased estimators under the situations considered. Our results shed some light on what method to use when heterogeneity in the underlying process is suspected. More importantly, it shows that ignoring such heterogeneity will result in biased estimates for the true treatment effect. Our result confirmed that where there is complex heterogeneity in the treatment assignment, simply using the existing method that ignores the heterogeneity in the treatment assignment will lead to biased estimators generally. We provide below a few remarks and insights on the performance of each method.

Interesting questions remain for future research. This includes developing a hypothesis testing method for detecting potential heterogeneity structure in observational studies, and investigating the properties of the PLS method. The robust properties of those estimators need to be developed theoretically and tested with real data. Chen, Zhang and Davidian [23] proposed a Monte Carlo EM (MCEM) algorithm for generalized linear mixed models with flexible random effects distribution. This algorithm allows the density to be skewed, multi-modal, fat- or thin-tailed relative to the normal distribution and includes the normal distribution as a special case. Next, we will consider using the MCEM algorithm for PS with random effects, which will be more robust. In conclusion, we suggest using the dummy adjusted PS method, i.e., $\left(\hat{\alpha}_{D\cdot I},\hat{\alpha}_{D\cdot D},\hat{\alpha}_{D\cdot M}\right)$ and/or the PLS method in practice whenever there is suspicion that there is heterogeneity in the underlying treatment assignment. When the number of clusters is small and the number of observations in each cluster is large, using the dummy variables to handle the fixed cluster effect is reasonable.

## ACKNOWLEDGEMENT

This work is supported in part by Pfizer, Inc. Drs. Liu and Willke are employees of Pfizer.

## SUPPORTING MATERIALS

The supporting materials can be downloaded from the journal website along with the article.

## REFERENCES

[1] Renehan AG, Tyson M, Egger M, Heller RF, Zwahlen M. Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. Lancet 2008; 371: 569-78. http://dx.doi.org/10.1016/S0140-6736(08)60269-X

[2] Silverman SL. From randomized controlled trials to observational studies. AM J Med 2009; 122: 114-20. http://dx.doi.org/10.1016/j.amjmed.2008.09.030

[3] Thompson S, Ekelund U, Jebb S, Lindroos AK, Mander A, Sharp S, Turner R, Wilks D. A proposed method of bias adjustment for meta-analyses of published observational studies. Int J Epidemiol 2011; 40: 765-77. http://dx.doi.org/10.1093/ije/dyq248

[4] Hair JF, Sarstedt MC, Ringle CM, Mean JA. An assessment of the use of partial least squares structural equation modeling in marketing research. J Acad Mark Sci 2012; 40: 414-33.

[5] Kucher N, Koo S, Quiroz R, Cooper J, Paterno M, Soukonnikov B, Goldhaber S. Electronic Alerts to Prevent Venous Thromboembolism among Hospitalized Patients. N Engl J Med 2005; 352: 969-77. http://dx.doi.org/10.1056/NEJMoa041533

[6] Sharpe N. Clinical Trials and the Real World: Selection Bias and Generalizability of Trial Results. Cardiovasc Drugs Therapy 2002; 16: 75-7.

[7] Alemayehu D, Ma J, Jones B, Willke R. Statistical issue with the analysis of non- randomized studies in comparative effectiveness research. J Manag Care Pharm 2011; 17: 22-26.

[8] Willke R, Mullins C. “Ten commandments” for conducting comparative effectiveness research using ”Real-World Data”. J Manag Care Pharm 2011; 17: 10-15.

[9] Rosenbaum R, Rubin D. The central role of the propensity score in observational studies for causal effects. Biometrika 1983; 70: 41-55.

[10] Stukel T, Fisher E, Wennberg D, Alter D, Gottlieb D, Vermeulen M. Analysis of Observational Studies in the Presence of Treatment Selection Bias. JAMA 2007; 297: 278-85.

[11] Chen Y, Lin J, Yu H, Ko W, Jerng J, Chang W, Chen W, Huang S, Chi N, Wang C, Chen L, Tsai P, Wang S, Hwang J, Lin F. Cardiopulmonary resuscitation with assisted extracorporeal life-support versus conventional cardiopulmonary resuscitation in adults with in-hospital cardiac arrest: an observational study and propensity analysis. Lancet 2008; 372: 554-61. http://dx.doi.org/10.1016/S0140-6736(08)60958-7

[12] Hong J, Yu B. Effects of kindergarten retention on children's social-emotional de- velopment: An application of propensity score method to multivariate, multilevel data. Dev Psychol 2008; 44: 407-21. http://dx.doi.org/10.1037/0012-1649.44.2.407

[13] Ye Y, Kaskutas L. Using propensity scores to adjust for selection bias when assessing the effectiveness of Alcoholics Anonymous in observational studies. Drug Alcohol Depend 2009; 104: 56-64. http://dx.doi.org/10.1016/j.drugalcdep.2009.03.018

[14] Wyse A, Keesler V, Schneider B. Assessing the effects of small school size on mathematics achievement: A propensity score-matching approach. Teachers College Record 2008; 110: 1879-900.

[15] Staff J, Patrick M, Loken E, Maggs J. Teenage alcohol use and educational attain- ment. J Stud Alcohol Drugs 2008; 69: 848-58.

[16] Maciejewski M, Livingston E, Smith V, Kavee A, Kahwati L, Henderson W, Arter- burn D. Survival Among High-Risk Patients After Bariatric Surgery. JAMA 2011; 305: 2419-26. http://dx.doi.org/10.1001/jama.2011.817

[17] Hoskuldsson A. PLS regression methods. J Chemometr 1988; 2: 211-28. http://dx.doi.org/10.1002/cem.1180020306

[18] Frank I, Friedman J. A Statistical View of Some Chemometrics Regression Tools. Technometrics 1993; 35: 109-35. http://dx.doi.org/10.1080/00401706.1993.10485033

[19] Rosipal R, Kramer N. Overview and Recent Advances in Partial Least Squares. Lecture Notes Comp Sci 2006; 3940: 34-51.

[20] Garthwaite P. An interpretation of partial least squares. J Am Statist Assoc 1994; 89: 122-27. http://dx.doi.org/10.1080/01621459.1994.10476452

[21] Rauschecker H, Sauer R, Schauer A, Schumacher M, Olschewski M, Sauerbrei W, Seegenschmiedt M, Schmoor C. Therapy of small breast cancer-four year results of a prospective non-randomized study. Breast Cancer Res Treat 1995; 34: 1-13.

[22] Senn S, Graf E, Caputo A. Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment or exposure. Stat Med 2007; 26: 5529-44.

[23] Chen J, Zhang D, Davidian M. A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution. Biostatistics 2002; 3: 347-60.

Received on 15-04-2014 Accepted on 29-04-2014 Published on 14-05-2014

http://dx.doi.org/10.6000/1929-6029.2014.03.02.13