ROBUST AND EFFICIENT STATISTICAL INFERENCE FOR CLUSTERED OBSERVATIONAL DATA IN COMPARATIVE EFFECTIVENESS RESEARCH

Baiming Zou

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the School of Public Health.

Chapel Hill
2013

Approved by:

Haibo Zhou

Jianwen Cai

Gary Koch

Amy Herring

Hongtu Zhu

Til Sturmer

ABSTRACT

Baiming Zou: Robust and Efficient Statistical Inference for Clustered Observational Data in Comparative Effectiveness Research
(Under the direction of Professor Haibo Zhou)

Treatment allocations in observational studies are nonrandom and result in the confounding problem and potentially biase treatment effect estimates. Propensity score (PS) methods are commonly used in practice to address the confounding problem. Among different PS methods, PS regression is frequently used in clinical research. Even though the treatment effect estimate from the PS regression model is unbiased under the strongly ignorable treatment assignment assumption, the default variance estimate is biased. In the first topic of this dissertation, an improved variance estimator for the treatment effect estimate is proposed.

Many observational data are clustered, for example, by physicians, and are therefore, not independent. A few PS methods consider correlated or clustered samples using mixed effects models with a strong normality assumption on the cluster effects. In the second part of this dissertation, a robust semi-nonparametric propensity score (SNP-PS) regression model is proposed. We relax the normality assumption and model the complex heterogeneity structure in treatment allocation process nonparametrically. The proposed SNP-PS model is robust and provides unbiased treatment effect estimates while parametric mixed effects PS models fail to do so when the cluster effects are non-normally distributed. We establish the asymptotic result for the treatment effect estimate and propose an unbiased variance estimator for it. Computationally, we propose an adaptive quadrature integration EM (expectation-maximization) algorithm to avoid potential large Monte Carlo errors of existing Monte Carlo EM algorithms.

Many real world medical record data are not only clustered but also multilevel clustered with millions of samples and hundreds of thousands of clusters. The SNP-PS framework is in theory applicable to these large datasets. However, in practice, it is computationally prohibited. In the third topic of this dissertation, we propose a flexible mixed effects PS model (FM-PS) that is computationally efficient for large multilevel clustered data. The FM-PS model relaxes a critical independence assumption that the random effects are independent of the fixed effect covariates made in the standard mixed effects PS (SM-PS) models. The FM-PS model provides an unbiased treatment effect estimate regardless whether the independence assumption holds or not. Though the treatment effect estimate from the SM-PS model is biased when the independence assumption does not hold, it is unbiased and more efficient than the estimate from the FM-PS model when the independence assumption holds. We propose a likelihood ratio statistics for testing the independence assumption which allows us to choose between the FM-PS and SM-PS models. A cluster bootstrapping procedure to estimate the variance of treatment effect estimate is proposed. The FM-PS model is robust to various model misspecifications as demonstrated by our extensive simulations.

This dissertation is dedicated to my wife, Fei Zou, Ph.D. Her support, encouragement, and constant love have sustained me throughout my life. It is also dedicated to my two lovely children, my daughter Jennifer and my son Chris for their understandings.

ACKNOWLEDGEMENTS

I greatly appreciate and owe tremendous debts to many people who made this dissertation possible. Their support and assistance were indispensable in completing this research project.

The list begins with my mentor and doctoral supervisor, Professor Haibo Zhou for being a supportive advisor anyone would hope to have. I would like to thank him for his patience, time, expertise, wisdom and great guidance both academically and non-academically. His enthusiasm, support and encouragement have helped me to surpass many challenges I have faced in the past several years and to stretch my abilities to my full potential. Without him, this dissertation would not be possible.

I would also like to express my sincere gratitude to Professor Gary Koch. Whenever I have encountered any obstacles, he was always there and provided me great support and suggestions.

My sincere gratitude also goes to Professors Jianwen Cai, Amy Herring and Hongtu Zhu. I greatly appreciate them for serving in my dissertation committee. I have benefited from Dr. Cai's theory of linear models, Dr. Herring's longitudinal data analysis and Dr. Zhu's generalized linear regression courses. I would like to thank them for their great advice and suggestions on my dissertation, and many other supports that are too long to list.

Last but not the least, I would like to thank Professor Til Sturmer from Department of Epidemiology. His great insight and knowledge in comparative effectiveness research have greatly improved and expanded the applications of the methodologies developed in this dissertation.

Finally, I extend my sincere appreciation to my fellow classmates, Ruoqing Zhu, Shangbang Rao, Hongtao Zhang and to everyone who cheered me up with smiles and kind words.

Baiming Zou

September 10, 2013
Chapel Hill, NC 27599

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Medical Record Data

1.1.1 Features of Medical Record Data and Its Applications

Randomized controlled trial (RCT) is routinely used in clinical research for estimating treatment efficacy in drug development and is considered as the gold standard to establish causal effects of drugs. In randomized controlled trials, treatments are randomly assigned by researchers. Proper randomization procedures guarantee that there exists no systematic difference among subjects for their baseline features in different treatment groups. Therefore, valid treatment effects can be estimated easily by comparing the outcomes from different treatment groups directly without the need to adjust for any other covariates.

Even though RCTs are regarded as one of the best designs to evaluate efficacy of drug therapies or other medical interventions, RCTs have their limitations (e.g. Kramer and Shapiro 1984). First, RCTs are expensive to conduct and thus small sample sizes are commonly observed in RCTs (Johnston et al. 2006). Second, experiments in RCTs may have involved a specific group of people and conducted under certain situations within a short time period. Third, the conduction of a randomized controlled trial usually is time consuming (e.g. for recruiting enough participants). Fourth, RCT is not always applicable due to ethical considerations. Fifth, some side effects of medications may not be able to be fully detected in RCTs due to short time period, small sample size, and/or limited participants. With all these limitations of RCTs, observational data, including the medical record data, have been extensively used as alternatives for the evaluation of therapy effectiveness or even drug efficacy.

Medical record data is a systematic collection of medical information for individual patients. A medical record usually contains various information, including personal and physical information like age, weight and blood pressure, family disease history, treatment assignment, demographics, medical history, medication and allergies, immunization status, laboratory test results, radiology images, vital signs, and billing information, etc for each patient. Rich information included in medical record data can be used for various purposes including statistical reporting on quality improvement, resource management and public health communicable disease surveillance, etc.

In recent years, medical record data have been increasingly used for comparative effectiveness research (e.g. Cebul et al. 2011; Schneeweiss and Avorn 2005; Suissa and Garbe 2007; Lau et al. 2011). The advantage of using medical record data for treatment effect estimate is that it offers broader population spectrum than the traditional randomized controlled trials (RCTs) because often participating subjects are highly heterogeneous in terms of their medical and social backgrounds. In addition, in practical medical record data, the treatments are assigned based on patients' clinical needs. Therefore, treatment effects from medical record data more closely reflect daily clinical practice than randomized controlled trials and provide clinically relevant information that may not be provided by RCTs (Yang et al. 2010). Treatment effects estimates using practical medical record data may supplement the evidence obtained from randomized controlled trials. Furthermore, comparative effectiveness research using medical record data and conducted with well-established statistical methods can improve causal inference of treatment effects (Berger et al. 2009). The retrospective analysis of medical record data can bridge the inferential gap between what is proved to be effective for the selected groups of patients versus the complicated clinical decisions required for individual patients (Stewart 2006). Large medical record database can help to identify the rare adverse event for medications that was not able to be detected in RCTs. Medical record data have become a valuable resource for comparative effectiveness research that allows researchers to determine the inferiority, equivalence, or superiority of various interventions when compared with each other (e.g. Mitka 2010).

However, there exist controversy and challenges on whether the medical record data can provide reliable information on treatment effect estimate (e.g. Pocock and Elbourne 2000; Ioannidis and Lau 2001). Some studies reveal reasonable results comparable to that from randomized controlled trials. On the other hand, some comparisons identify inconsistent results with RCTs. For example, statin therapy was found to decrease the overall mortality and myocardial infarction based on Weiner et al. (2008) research. Weiner et al. (2008) used medical record data from the United Kingdom General Practice Research Database (GPRD) which includes the compiled information for over 8 million patients. These results were reasonably comparable to that from other RCTs. But, in the same study, they found disparity between the results based on GPRD database and that from RCTs about the effect of statin therapy for coronary revascularization. Another example, prophylaxis is shown to reduce the occurrence of venous thromboembolism among critically ill patients (The PROTECT Investigators 2011) in the randomized controlled trial. However, the trial findings extended to real world patients have been inconclusive (Sharpe et al. 2002). While some studies found that observational studies and RCTs are overall comparable and produced similar results (e.g. Benson and Hartz 2000; Concato et al. 2000), other research identified that discrepancies beyond chance do occur and differences in estimating the magnitude of treatment effect do often exist (e.g. Ioannidis et al. 2001).

Even though one could attribute some of the inconsistent results between observational studies and randomized controlled trials to the lack of rigorous inclusion and exclusion criteria, exposure definitions and outcomes identical to the RCT (Tannen et al. 2008), the key difference between RCT and the real world medical record data is that the treatment assignment in real world medical data is not random. In medical record data, the treatment allocation of each patient is primarily made by physicians according to the patient's physical condition, disease severity, the physician's preference of a therapy, etc. Nonrandom treatment assignment of medical record data could result in large differences in the baseline covariates between the treated and untreated groups. The imbalanced baseline covariates distributions may twist the treatment effect, i.e. a problem known as confounding.

Apart from its distinctive observational feature, another key feature of medical record data is the clustering feature, for example, by physicians, clinics, hospitals, and insurance agencies. The clustering feature of medical record data often reflects the heterogeneity in patients' health conditions, social economical status, and so on, which not only plays important roles in treatment assignment decision but also affect the disease outcomes. To obtain valid treatment effect estimate for clustered observational data, it is critical to take into consideration of confounding and sample heterogeneity features. The following two subsections review existing statistical methods for clustered observational data.

1.1.2 Review of Confounding Adjustment Methods

The large and heterogeneous populations included in medical record datasets provide ideal resources to examine treatment effects and outcomes under real world conditions over long periods. However, the bias induced by the confounding factors restricts medical records' capacity to distinguish treatment effects from the effects of patient-related, disease-related, and provider-related factors, etc. Ignoring or inappropriately adjusting the confounding may result in biased treatment effect estimate and erroneous conclusions. A large number of confounding adjustment methods have been proposed in the literature to assess the treatment effects for non-randomized data which usually fall into the following categories: matching, stratification, instrumental variable, multiple regression, and various versions of propensity score methods.

Matching

Matching is the simplest and most intuitive confounding adjustment method to account for selection bias under non-randomized design (e.g. Rubin 1973; Wacholder 1992; Greenland et al. 1981; Miettinen 1968; Kupper et al. 1981; McKinlay 1977; Rose and van der Laan 2009). The idea is similar to that of randomized trials in making the confounding factors balanced as much as possible between treated and untreated groups. Basically, matching procedure first identifies a set of confounding covariates , i.e. the factors that are potentially related to both the dependent outcomes and the independent variable of interest (i.e. the treatment assignment). For each covariate in the set of identified confounding covariates, a subject in the treated group is matched with another subject in the untreated group with (nearly) identical value of the confounding covariate considered. Then, by doing this way, the subjects matched will be almost balanced with respect of the confounding covariates between the treated and untreated groups. In this sense, matching can be viewed as a manually created RCT.

Matching is simple and straightforward to implement. In many cases, it provides reasonable solutions to control confounding factors and reduce bias in treatment effect estimates (Wunsch et al. 2006). Matching can also reach a balanced number of treated and untreated across the levels of the selected matching variables. This balance can reduce the variance in the parameter estimates of interest and improves statistical efficiency (Kupper et al. 1981; Rothman and Greenland 1998). However, intrinsic limitations exist for matching which restrict its practical applications. In situations where many confounding factors exist, matching can be difficult, and result in low sample overlap. Matching with low sample overlap will cause inefficiency because those unmatched subjects have to be thrown away without including in the analysis. More importantly, how to select a "right" set of confounding variables is tricky but important. Inappropriately selecting the covariates to match over will lead to overmatching issue. Overmatching could severely lower the statistical analysis power and lead to a new bias (Day et al. 1980).

Stratification

Stratification is another simple approach for confounding adjustment (Cochran 1968; Miettinen 1976). Similar to matching procedure, stratification procedure also identifies a list of potential confounding variables first. It identifies two or more mutually exclusive subgroups or strata within which the confounding variables are largely constant. For each of the identified confounding covariates, a set of subgroups are created such that the covariate is similar within each of the subgroup, for example classifying age into decades, or weight into quartiles. Subgrouping will be subsequently performed within each subgroup based on another remaining confounding factor from the list. The subgrouping process continues till no more remaining confounding variables left in the list or there is no more subjects to be classified. Stratification process usually generates a tree and the height of the tree and the number of nodes created depending on the number of confounding covariates and the number of samples. The order of nodes created from stratification depends on the order of confounding factors used in the subgrouping procedure. This means that different stratification order on the confounding variables could lead to different treatment effect estimate. After subgroups are created, stratify analysis is performed.

After the intervention effect is estimated within each stratum, a pooled estimate is calculated across strata to generate the final overall estimate. Weighting is commonly used for combining each stratum's estimate to obtain the overall estimate (noted as adjusted estimate). Mantel-Haenszel method (Mantel and Haenszel 1959) is the most popular approach for this purpose. It uses a weighted average of the stratum-specific estimates to obtain the overall estimate. The weights are inversely proportional to the variances of the stratum-specific estimates, i.e. the more precise the estimates are, the greater weights they get. Homogeneity of stratum-specific estimates can be tested via, for example χ^2 test. In this sense, stratification method is similar to the meta-analysis where the goal is to combine treatment effect estimates from different studies.

The algorithm of stratification is simple to implement and robust without specific assumptions for the distributions of confounding variables and no linear relationship between the outcome and confounding factors is assumed (Cochran 1968). Results based on stratification are clear and easy to interpret (Klungel et al. 2004). However, several limitations exist for this method. First, the computational workload could be intensive since the number of subgroups can increase exponentially with the increase of the number of confounding covariates and the number of subjects. Second, grouping based on continuous variables would be subjective where the original continuous confounder is replaced by a less accurate, categorical version. This may lead to residual confounding if the strata is not fine enough (Becher 1992). As a result, stratification usually is restricted to categorical variables. All these make stratification method less practical for large datasets with a number of confounders, such as medical record data.

Instrumental Variable

Instrumental variable (IV) method is another well-known confounding controlling approach and has been widely used in economical research (e.g. Angrist and Krueger 2001; Heckman 1997; Miguel et al. 2004). It is later introduced into clinical research (e.g. McClellan 1994; Permutt and Hebel 1989; Vansteelandt et al. 2011). Even though it is mainly used for confounding adjustment purpose under observational design, it is used for causal inference as well (Angrist et al. 1996). An instrumental variable is an observed covariate that is associated with the independent variable of interest (e.g. treatment assignment) but not with the random measurement error term. In other words, the instrumental variable is associated with the independent variable of interest but NOT associated with the outcome directly. That is the effect of an instrumental variable on the dependent variable is indirectly through another independent variable.

IV approach can be demonstrated by the following regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{1.1}$$

A variable z is called an instrumental variable of the regressors x if z is uncorrelated with the error term ϵ but correlated with x. A simple example can demonstrate how this scenario can happen. Suppose y has the following true linear relationship with x^*:

$$y_i = \beta_0 + \beta_1 x_i^* + \epsilon_i^* \tag{1.2}$$

Suppose x is observed via x^* with an error ξ such that $x_i = x_i^* + \xi_i$, where x_i^* and ξ_i are independent. Then the true regression equation (1.2) for regressor x^* can be rewritten by using its proxy x as equation (1.1) with $\epsilon_i = \epsilon_i^* - \beta_1 \xi_i$. Therefore, regressor x is correlated with the error term ϵ since $Cov(x, \epsilon) = Cov(x^* + \xi, \epsilon^* - \beta_1 \xi) = -\beta_1 var(\xi)$.

In the above regression equation (1.1), $\hat{\beta}$ from OLS is biased and inconsistent.

Suppose now we have another variable z which equals x^*. By the definition of IV, it is easy to prove that z is an IV of x. Based on z, we can get the instrumental variable estimator for β as:$\hat{\beta}_{IV} = (\mathbf{z'x})^{\mathbf{-1}}\mathbf{z'y}$. It can be shown that $E(\hat{\beta}_{IV}) = \beta$, thus $\hat{\beta}_{IV}$ is an unbiased estimator of β.

In practice, the IV estimator is obtained via two-stage linear regression. That is, for each confounding covariate x, identify the corresponding instrumental variable and do the first stage linear regression for the confounding covariate as dependent variable against the instrumental variable and obtain the prediction for x, i.e. $\hat{x}$. Substitute x with $\hat{x}$ in the original regression model and conduct the second stage linear regression to obtain the parameter estimate.

The IV estimator is consistent and the procedure to obtain IV estimate is simple and straightforward. Studies by Brookhart et al. (2006) show that the instrumental variable method could substantially reduce the bias due to unobserved confounding. However, there exist limitations for this method also. For example, the requirements for IV are difficult to satisfy and test in practice, such as the independent assumption on the instrumental variables with respect to the error term (Klungel et al. 2004). Therefore, it is not an easy task to identify an instrumental variable. In practice, the determination of an instrumental variable is subjective. Therefore, the generalization of the findings from the IV method is questionable (Klungel et al. 2004). As such, the validity of using IV for treatment effect estimate under non-randomized design remains debatable.

Multiple Regression

Multiple regression can be used to adjust the effects of confounding factors directly. Advantages of regression methods include allowing many confounding variables being included in the model and the possibility of incorporating quantitative continuous factors without categorization and the possibility of modeling trends in confounders measured on an ordinal scale. Multiple regression method is regarded as the gold standard method for adjusting confounding factors since it would provide the best linear unbiased estimates (BLUE) when the assumptions for the regression model hold. However, such efficiency gains would be at the risk, for examples when the number of confounding variables is not small (with respect to the number of samples) and the regression model is incorrect (such as covariate functional form misspecified). Furthermore, in regression analysis, limited overlapping of confounding covariates between treatment groups may lead to multicollinearity.

Propensity Score

As an alternative for multiple regression method, propensity score (PS) method by Rosenbaum and Rubin (1983) is often used in practice (e.g. Seeger et al. 2005; Huang et al. 2005; Sturmer et al. 2006; Hong and Yu 2008; Wyse et al. 2008; Staff et al. 2008; Lunt et al. 2009; Ye and Kaskutas 2009) to adjust for confounding factors and estimate the treatment effect under non-randomized design via stratification, matching, inverse probability weighting, or covariate adjustment.

A propensity score is defined as the conditional probability of a unit (e.g. person) being assigned to a particular treatment in a study given a set of known covariates. Let the binary variable trt refer the treatment assignment (with 1 for treated and 0 for untreated) and $\mathbf{x}$ refer the vector of the covariates.

$$PS = Pr(trt = 1 \mid \mathbf{x})$$

In randomized controlled trial, this probability is known and independent of covariates or the observed features $\mathbf{x}$, i.e. $trt \perp \mathbf{x}$ or $Pr(trt = 1 \mid \mathbf{x})$ is a constant (usually it is set as 0.5) regardless of $\mathbf{x}$. In observational studies, the propensity of receiving treatment is unknown but depends on $\mathbf{x}$, i.e. the treatment allocation depends on other covariates. Imbalance of propensity score indicates an imbalance in covariates between the two comparison groups. The goal of propensity score analysis is to balance two non-equivalent groups based on their propensity scores to reduce the selection bias in the treatment effect estimate.

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

$$(y(1), y(0)) \perp trt|\mathbf{x} \tag{1.3}$$

$$0 < P(trt = 1|\mathbf{x}) < 1. \tag{1.4}$$

where $y(1)$ and $y(0)$ are the potential outcomes of a particular unit under the treated and untreated, respectively. That is, $y(1)$ and $y(0)$ are the outcomes if the unit had been assigned to the treated and untreated group, respectively. They are never observed simultaneously in reality. Their relationship with the observed outcome y and the treatment assignment trt can be expressed as $y = trt * y(1) + (1 - trt) * y(0)$. The first condition (1.3) says that treatment assignment is independent of the potential outcomes conditional on the observed baseline covariates. Rosenbaum and Rubin (1983) had shown that conditional on the propensity score, the distribution of measured baseline covariates is similar between the treated and untreated subjects. Thus, for a set of subjects who have the same propensity scores, the distributions of the baseline covariates will be the same between the treated and untreated groups. They demonstrated that if treatment assignment is strongly ignorable(i.e. conditions (1.3) and (1.4) hold), conditioning on the propensity score allows one to obtain unbiased estimates of the treatment effects. This condition is also referred to as the no-unmeasured confounders assumption,i.e., all confounding variables that affect the treatment assignment and the outcome have been measured and included in $\mathbf{x}$. Under this assumption it has been shown that adjustment made with the propensity score is sufficient to remove the bias due to the non-random treatment assignment in both large and small sample scenarios (Rubin 1997).

In practice, the propensity score is unknown and commonly estimated via the logistic regression model:

$$logit(Pr(trt_i = 1 \mid \mathbf{x}_i)) = \beta_0 + \mathbf{x}_i\boldsymbol{\beta} \quad (i = 1, \cdots, n) \tag{1.5}$$

where $\mathbf{x}_i$ represents all the observed covariates other than the treatment assignment trt_i (1 for treated and 0 for untreated) of subject i. Denote the parameter estimates of β_0 and $\boldsymbol{\beta}$ (i.e. maximum likelihood estimates (MLE)) as $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$. Then, PS can be estimated as:

$$\widehat{PS}_i = \exp(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})/[1 + \exp(\hat{\beta}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}})]$$

A critical aspect of PS based method is to obtain valid PS estimation which often involves model and variable selection (e.g. Brookhart et al. 2006). Even though the prevailing propensity score estimation method is logistic regression, several other propensity score estimation methods have been proposed such as boosting and bagging (Lee et al. 2010; McCaffrey et al. 2004), random forests (Lee et al. 2010), neural networks (Setoguchi et al. 2008), and regression tree or partitioning methods (Lee et al. 2010; Setoguchi et al. 2008).

Based on the estimated propensity scores from equation (1.5), the treatment effect can be estimated via matching (Dehejia and Wahba 2002), stratification (Rosenbaum and Rubin 1984; He and McDermott 2012), inverse-probability-weighting (IPW) (Rosenbaum 1998), or covariate adjustment (Shepardson et al. 1999; Perkins et al. 2000).

In practice, an often used propensity score matching is one-to-one match which matches each treated (i.e. $trt = 1$) subject to an untreated (i.e. $trt = 0$) subject with identical propensity score. However, since the propensity score is a continuous variable, it is difficult to match a treated with an untreated with exactly identical propensity score. Several propensity score matching algorithms have been proposed in this regard. Commonly used propensity score matching algorithms include nearest neighbor matching and clipper matching (Dehejia and Wahba 1999). Propensity score matching is simple. In practice, PS matching is often used by researchers in various observational studies designs. However, shortcomings with propensity score matching include (Shadish et al. 2002): large samples are required and overlap must be substantial. In the case of low overlapping, a large number of samples will not be matched and have to be discarded which could result in low estimation efficiency. In addition, low overlap of propensity scores between the two groups may result in some severely imbalanced covariates after matching. Propensity score matching could also match two dissimilar subjects if the propensity score range used for matching is too broad and thus lead to inexact matching. In practice, matching by propensity score may fail to remove all bias due to confounders because samples may not be able to be matched sufficiently closely (Hill 2008) and the within-pair differences in covariate values may still be large (Rosenbaum and Rubin 1985).

Compared to the conventional stratified analysis method described previously, stratification using propensity score becomes straightforward since all the samples are stratified according to only ONE variable, i.e. the estimated propensity score, instead of stratifyin all possible confounding variables one by one. This process greatly reduces the complication and thus improves computation efficiency. If the treatment assignment is strongly ignorable, then conditional on propensity scores by stratification will allow us to obtain unbiased treatment effect estimate. By combining the stratum-specific treatment effect estimate, the weighting algorithm as shown below produces an overall adjusted estimate of treatment effect:

$$\hat{\alpha}_{trt} = \sum_{j=1}^{K} \frac{n_j}{n} \{ \frac{1}{n_{1j}} \sum_{i=1}^{n} trt_i y_i I(\widehat{PS_i} \in Q_j) - \frac{1}{n_{0j}} \sum_{i=1}^{n} (1 - trt_i) y_i I(\widehat{PS_i} \in Q_j) \}$$

where K is the total number of strata n_{1j} and n_{0j} are the number of subjects in treated and untreated groups within stratum j, respectively. $n_j = n_{1j} + n_{0j}$ is the number of subjects in stratum j, and Q_j denotes the propensity score range of stratum j. Identifying individuals having exactly the same propensity value may be infeasible in practice, stratification attempts to achieve groups where this at least holds approximately. Consequently, the treatment effect estimator via PS stratification may be a biased estimator of the treatment effect, as some residual confounding within strata may remain.

One benefit of propensity score stratification over matching is that it allows samples who might not have a close enough matching mates on their propensity scores to be included in some strata and not be discarded for the treatment effect estimation. No matter propensity score matching or stratification, both methods need to determine the propensity score cutoff for declaring participants and nonparticipants having exactly "identical" propensity scores which is subjective. In practice, researchers often use these two schemes due to their simplicity and easy to understand.

Inverse probability weighting (IPW) is another approach that weighs observations from each group (i.e. treated and untreated) by the inverse of the probability of being in that group (Rosenbaum 1998). Specifically, IPW estimates treatment effect as the following:

$$\hat{\alpha}_{trt} = \frac{1}{n} \sum_{i=1}^{n} \{ \frac{trt_i * y_i}{\widehat{PS_i}} - \frac{(1 - trt_i) * y_i}{1 - \widehat{PS_i}} \}$$

where $w_{1i} = \frac{trt_i}{n\widehat{PS}_i}$ and $w_{0i} = \frac{1-trt_i}{n(1-\widehat{PS}_i)}$ are corresponding weights. One problem in practice with this estimator is that the weights do not necessarily add up to 1. Therefore, another normalized version of IPW is used in practice:

$$\hat{\alpha}_{(trt,norm)} = \sum_{i=1}^{n}\{\frac{trt_i * y_i}{w_1\widehat{PS}_i} - \frac{(1-trt_i) * y_i}{w_0(1-\widehat{PS}_i)}\}$$

where $w_1 = \sum_{i=1}^{n} \frac{trt_i}{\widehat{PS}_i}$ and $w_0 = \sum_{i=1}^{n} \frac{1-trt_i}{1-\widehat{PS}_i}$. The IPW estimators, no matter normalized or not, are unbiased and consistent estimator of the treatment effect $\Delta = E(y(1)) - E(y(0))$ with $y(1)$ and $y(0)$ being the potential outcomes if the subjects have had been assigned to the treated and untreated groups, respectively. In addition, these estimators are also asymptotically normal under certain regularity conditions. Compared with the stratification estimator, IPW estimator is more efficient (Lunceford and Davidian 2004). In practice, in addition to being used as a confounding adjustment tool, IPW is also often used by researchers to describe missing and censoring data.

Another convenient approach using the propensity score for confounding adjustment is the covariate adjustment approach (e.g. D'Agostino 1998) by including the propensity score as a covariate in the following regression model

$$\mathbf{y} = \alpha_0 + \alpha_{trt}\mathbf{trt} + \alpha_{PS}\widehat{\mathbf{PS}} + \boldsymbol{\epsilon} \tag{1.6}$$

Rosenbaum and Rubin (1983) had shown that the treatment effect estimate obtained via the PS regression (1.6) is unbiased under the strongly ignorable treatment assignment assumptions (1.3) and (1.4).

Propensity score regression is less sensitive to the misspecification of the functional form (i.e. linear or quadratic) of the covariates as compared to the multiple regression method (Rubin 1997). Compared with other versions propensity score methods, propensity score regression needs more restrictive assumption on the linearity between response and propensity score in addition to the strongly ignorable treatment assignment assumption that the other propensity score methods also made. In contrast to the regression model where all covariates are incorporated in the regression analysis, propensity score regression reduces baseline information to a single composite summary of the covariates. In this sense, PS regression can be viewed as a dimension reduction technique also. As compared with propensity score matching and stratification, propensity score regression is simpler to use without the tedious procedure and burden to match samples with close propensity scores. Propensity score regression results in increased precision for continuous outcomes and increased statistical power for continuous, binary, and time-to-event outcomes (Steyerberg 2009). In the perspective of applications, PS matching, stratification, and inverse probability weighting can be used for observational study design and analysis while PS regression is mainly used for the analysis. There exists many review literatures on propensity score methods (e.g. D'Agostino 1998; P.C. Austin 2011). A review and comparison between different versions of PS methods and multiple regression method can be found in Sturmer et al. (2006).

1.1.3 Review of Clustered Data Analysis Methods

Besides the observational feature of medical record data where the treatment is not randomly assigned, another distinctive characteristic of medical record data is that they are clustered e.g. by physicians, clinics, hospitals, or insurance agencies. Furthermore, medical record data are not only clustered but also they could be multi-lever or hierarchically clustered. For example, patients are clustered with physicians who are also clustered by hospitals. As such, in the analysis of medical record data, both the observational and clustering features should be taken into account to obtain the valid treatment effect inference. The confounding control methods reviewed in the previous section are all for independent samples rather than clustered data.

The degree of clustering can be delineated in terms of correlation among the measurements on units within the same cluster. Appropriate statistical models for clustered data must explicitly describe and account for this correlation. With more and more repeated measurements and longitudinal designs being used in various biomedical and social economical studies, the interest in the analysis of clustered data continuously grows. Many clustered data analysis techniques have been developed to deal with different challenges. Mixed effects model (e.g. linear mixed effects model, generalized linear mixed effects model, etc) and the generalized estimating equations (GEE) method are the two most widely used methods for the analysis of correlated data, which we review below.

Mixed Effects Model

As noted in previous section, multiple regression model is a very versatile approach in describing the relationship between the mean response and a set of independent covariates. However, the straightforward application of general regression method to the clustered data like medical record data is not appropriate due to the lack of independence among samples.

Many researchers have incorporated random effects into a wide variety of regression models to account for dependent structures of responses and multiple sources of variations. A frequently used model for describing clustered continuous data is the linear mixed effects model (e.g. Laird and Ware 1982; Lindstrom and Bates 1988) where random effects are used to model the correlations among samples within each cluster:

$$y_{ij} = \alpha_0 + \alpha_{trt} trt_{ij} + \mathbf{x}_{ij}\boldsymbol{\alpha}_{\mathbf{x}} + \eta_i + \epsilon_{ij} \tag{1.7}$$

where $\mathbf{x}_{ij}$ and y_{ij} represent the covariates and the response outcome of subject j in cluster i, respectively. The random variable η_is are independent and identically distributed as $N(0, \sigma_\eta^2)$ which denotes the cluster specific effect to account for mean differences amongst clusters and the random error $\epsilon_{ij} \sim N(0, \sigma^2)$ is assumed to be independent of the η_is.

Examples of using mixed effects model for clustered data analysis can be found in many biomedical and life science research (e.g. Petkova and Teresi 2002; Vaida and Xu 2000). The above linear mixed model can be easily extended to generalized linear mixed models (GLMMs) (Schall 1991; Zeger and Karim 1991; Breslow and Clayton 1993; Davidian and Giltinan (1995,2003); McCulloch and Searle 2001) for other types of responses such as binary outcomes or count data.

Many real world data usually includes more than one level of clusters. For example, in medical record data, many patients are treated by a physician and many physicians work in a clinic/hospital. This leads to physicians as the first layer cluster who are nested in the second layer cluster, i.e. clinics or hospitals. In the scenario of multilevel or hierarchical clustering, a more complex mixed effects model is needed to account for the heterogeneity of each cluster level. Conventionally, the following multilevel mixed effects model (e.g. Sullivan et al. 1999; Goldstein et al. 2002) is used to model the clustered data structures:

$$y_{ijk} = \alpha_0 + \alpha_{trt} trt_{ijk} + \mathbf{x}_{ijk} \boldsymbol{\alpha}_{\mathbf{x}} + \eta_i + \xi_{ij} + \epsilon_{ijk} \tag{1.8}$$

where $\mathbf{x}_{ijk}$ and y_{ijk} represent the observed covariates and the response outcome for subject k nested in sub-cluster ij which is further nested in cluster i, respectively. Cluster effects η_i, ξ_{ij}, and random error ϵ_{ijk} are mutually independent with each other and are assumed to be normally distributed with mean 0. Each level cluster effect accounts for the mean differences amongst clusters of the corresponding clustering level.

To obtain the parameters estimates, maximum likelihood estimate is routinely used. However, in the mixed effects model, besides the fixed effects covariates (i.e. trt_{ij} and $\mathbf{x}_{ij}$) which are observed, there exists random effects terms (i.e. η_i, ξ_{ij}) which are unobserved. Therefore, parameter estimates in mixed effects model can be treated as a missing data problem, and the expectation-maximization (EM) algorithm (Dempster et al. 1977), a well-known algorithm for maximum likelihood estimation based on incomplete data, can be used for parameter estimates.

In addition to the routine normality assumption made on the random effect, it is also assumed that the random effect is independent of other fixed effects terms. All these assumptions are made for the simplicity of statistical analysis and may not hold for many real applications including the medical record data where the treatment assignments are clustered. However, violations of these assumptions can result in severe biased estimates and invalid statistical inferences (Verbeke and Lesaffre 1996).

Generalized Estimating Equations

Another frequently used method for dealing with dependent observations is through what has become generally known as generalized estimating equations (GEEs) (e.g. Liang and Zeger 1986; Hardin and Hilbe 2003). GEEs can be regarded as an extension of quasi-likelihood models for independent measurements. This modeling scheme has often been applied in biomedical research (e.g. Cologne et al. 1993; Hanley et al. 2003).

A notable characteristic of GEE approach is that under mild regularity conditions, the parameter estimates from the GEE are consistent even when the covariance structure is misspecified. The primary interest of GEE is on estimating the average response over the population (i.e. marginal response) rather than the regression parameters that would enable the estimation of the effect of changing one or more covariates on a given individual. The parameter β estimates, i.e. $\hat{\beta}$, of GEE are obtained by solving the following equation:

$$\sum_{i=1}^{n} \mathbf{D}_i^{\prime} \mathbf{V}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = 0 \tag{1.9}$$

where $\mathbf{D}_i$ is the $n_i \times p$ matrix of derivatives of μ_i with respect to β and V_i is treated as known up to certain parameters (referred as working covariance matrix) and μ_i is the vector of mean responses with elements of $\mu_{ij}(\beta) = g^{-1}(x_{ij}\beta)$. The covariance estimates for β estimates, i.e. $\widehat{Cov}(\hat{\beta})$ can be obtained via the sandwich estimator.

Apart from the aforementioned commonly used methods for dealing with clustered data, there exists few research using mixed effects PS model by including the unobserved cluster effect in the logistic regression model to adjust for the treatment assignment heterogeneity confounding factors (e.g. Thoemmes and West 2011) where the normality is assumed for the cluster effects.

1.2 Proposed Research for Medical Record Data

1.2.1 Robust Two-Stage Variance Estimation for PS Regression Models

Propensity score (PS) is commonly used in observational studies for adjusting confounding factors when comparing the effectiveness of different treatments. Among different PS-based methods, the PS covariate adjustment (a.k.a. PS regression) which uses the estimated PS as a covariate in the second stage regression model has been frequently used in clinical research. In practice, researchers tend to make their inference on the treatment effect based on the default variance estimate from the second stage regression model. This variance estimate, however, does not take into consideration of the fact that the propensity score itself is an estimated quantity. Without proper correction, the default variance estimate could be biased. To address this problem, we jointly model the treatment assignment and the response variable under a two-stage regression framework. The asymptotic results for the treatment effect estimator are established, based on which a robust variance estimator is developed.

Specifically, we regard the propensity score regression models as a two-stage procedure shown below:

$$\textit{Stage 1:} \qquad trt_i \mid \mathbf{x}_i \sim f(trt_i \mid \mathbf{x}_i, \boldsymbol{\theta}_1) \tag{1.10}$$

$$\textit{Stage 2:} \qquad (y_i, trt_i) \mid PS(\mathbf{x}_i, \boldsymbol{\theta}_1) \sim f(y_i \mid trt_i, PS(\mathbf{x}_i, \boldsymbol{\theta}_1), \boldsymbol{\theta}_2) f(trt_i \mid PS(\mathbf{x}_i, \boldsymbol{\theta}_1)) \tag{1.11}$$

where y_i is the response variable, trt_i is the treatment assignment (1 for treated and 0 for untreated), and $\mathbf{x}_i$ is the covariate vector for the ith individual ($i = 1, 2 \cdots, n$), respectively. $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are the parameter sets associated with Stage 1 and 2 models. The unknown PS score, $PS(\mathbf{x}, \boldsymbol{\theta}_1)(= Pr(trt = 1 \mid \mathbf{x}, \boldsymbol{\theta}_1))$, is a function of $\mathbf{x}$ and $\boldsymbol{\theta}_1$. Based on this two-stage regression framework, we establish the asymptotic result for $\hat{\boldsymbol{\theta}}_2$, i.e. the parameter estimate for Stage 2 model which includes the treatment effect estimate, our primary interest. From the asymptotic distribution of $\hat{\boldsymbol{\theta}}_2$, we propose the covariance estimator.

1.2.2 A Semi-Nonparametric PS (SNP-PS) Model for Clustered Data

The treatment allocation in medical record data is non-randomly made by physicians according to various factors observed and/or unobserved (e.g. physician's factor, etc.), these factors could also completely or partially impact the disease outcomes. Consequently, samples in medical record data are clustered with respect to physicians. As a result, the mixed effects model is a natural approach for medical record data with the heterogeneity being included as a random effect term. With the attractive features of propensity score adjustment approach for confounding adjustment, we propose a generalized mixed effect propensity score model to deal with the heterogeneity in treatment assignment process. Few current literatures deal with the random effects in logistic regression model (e.g., Thoemmes and West 2011) for propensity score estimation. However, a critical assumption made for the conventional mixed effects model is that the cluster effect term is normally distributed which may be too restrictive for real world medical record data. The validity of this assumption is hard to check for real applications and violation of the assumption will lead to invalid propensity score estimation and result in biased treatment effect estimate. Indeed, the normality assumption often does not hold in practice due to many confounding factors and complex heterogeneity involving in the treatment allocation process. To be more flexible in capturing the heterogeneity structure in the treatment allocation process for medical record data, we relax the normality assumption on the random effect in the generalized mixed effects model with an unspecified random effect term. With the PS regression framework, the estimated PS from the proposed model is incorporated to adjust the heterogeneity confounding and obtain the treatment effect estimate that could be biased if the conventional statistical methods are used where the treatment assignment heterogeneity is incorrectly modeled.

Specifically, we consider the following semi-nonparametric (SNP) logistic regression model for the propensity score estimation:

$$logit(Pr(trt_{ij} = 1 \mid \mathbf{x}_{ij})) = \beta_0 + \mathbf{x}_{ij}\boldsymbol{\beta}_{\mathbf{x}} + \eta_i \tag{1.12}$$

Instead of assuming that the η_is are normally distributed, we make no specific assumption on the form of the distribution of the η_is except that it is a smooth density function $f(\eta)$. We approximate the density function $f(\eta)$ by $K+1$ terms of Hermite polynomial multiplied by a normal density (Gallant and Nychka 1987). Once the PS scores are estimated from the above model, then the rest of the analysis will follow the traditional PS regression procedure.

Estimation of propensity score based on the SNP-PS model is challenging. The primary difficulty results from the fact that the likelihood function of model (1.12) is complicated without closed analytic form since it involves the integration over a nonlinear integrand. To resolve this problem, various approximation techniques (e.g. Laplace approximation) were developed to avoid the integration (Schall 1991; Breslow and Clayton 1993). However, several research (Breslow and Lin 1995; Lin and Breslow 1996) shows that approximations based approaches may yield biased fixed effect estimates, particularly for the binary responses. Alternative approaches (Lee and Nelder 1996; Jiang 1999) were proposed to resolve this difficulty. On the other hand, methods were developed to conduct the integrations via the Markov chain Monte Carlo techniques (Zeger and Karim, 1991) and Monte Carlo EM (MCEM) algorithms (McCulloch 1997; Booth and Hobert 1999). For example, an approach proposed by Booth and Hobert (1999) is based on using rejection sampling technique to generate samples from the appropriate conditional distribution. The advantage of this approach is that it allows to evaluate the Monte Carlo error at each iteration and automatically increase the Monte Carlo sample size accordingly and thus reduce the unnecessary computational workload.

Chen et al. (2002) extend the rejection sampling scheme of Booth and Hobert (1999) with a double rejection Monte Carlo sampling approach for accessing the conditional likelihood for the generalized mixed effects model. However, the Monte Carlo sampling suffers with Monte Carlo errors if the random sample number is not large enough (Booth and Hobert 1999; Chen et al. 2002). Heavy computation workload associated with the large Monte Carlo sample size prohibits the application of the MCEM sampling to the medical record data with hundreds of thousands observations. Furthermore, the second rejection sampling component may suffer a low acceptance rate in some applications, particularly for binary data when the proportions of 0 and 1 responses within a cluster are severely imbalanced (Chen et al. 2002). To avoid these difficulties, alternatively, we propose a numerical integration approach, i.e. adaptive quadrature integration, to calculate the conditional likelihood for the proposed SNP-PS model (1.12). It can be shown that under the assumption of no other unmeasured confounding covariates, the treatment effect estimate from the proposed SNP-PS framework is unbiased and asymptotic normal.

1.2.3 A Flexible Mixed Effects PS (FM-PS) Model for Clustered Data

When medical record data are multilevel clustered, the robust SNP-PS model proposed above is computationally too intensive to be applied to huge multilevel clustered data. More computationally efficient PS models are needed. Note that in the SNP-PS model, the heterogeneity cluster effects are dealt with in the mixed effects PS model which is computationally expensive. Alternatively, we propose a flexible linear mixed effects propensity score (FM-PS) model which does not take into account the heterogeneity cluster effects in the PS model but instead leaves the heterogeneity adjustment to the subsequent PS covariate adjustment model.

Extra covariates are introduced in FM-PS models based on the observed treatment assignment and the estimated PS to model the correlation structures between the random cluster effects and the fixed effect terms such that only the standard simple mixed effect model is needed to fit and obtain the treatment effect estimate. The treatment effect estimate equivalence is established between FM-PS model and the fixed effects PS model (FE-PS) where the cluster effect is treated as the fixed effect by using dummy variables. The equivalence of treatment effect estimate provides the justification that the proposed FM-PS relaxes both the independence and normality assumptions for the random cluster effects that the traditional clustered data modeling frameworks made but are generally not held by the real world medical record data. Therefore, the proposed FM-PS is not only robust for the random effect density structure due to clustering but also flexible to the correlation between the random effects and other fixed effects confounding covariates. More importantly it is applicable for huge dataset like medical record data for practical use.

As all other propensity score based methods, the standard variance estimation for FM-PS models are not valid either. Accordingly, we propose a cluster bootstrapping procedure to obtain the valid variance estimate empirically.

1.3 Innovation of Proposed Research

1.3.1 Significance of Two-Stage Variance Estimation

Even though the treatment effect estimate from the PS regression model (1.6) is unbiased, the variance estimation directly from the PS regression model could be severely biased. In contrast, the proposed two-stage variance estimation scheme for propensity score regression models provides a valid variance estimation approach that corrects the bias of the commonly used default variance estimate from the second stage PS regression model. The default variance estimation ignores the fact that the propensity score is an estimated quantity instead of the observed covariate. Without jointly modeling the response and the treatment assignment will lead to the default variance estimation which could be severely biased as we will see in our simulation studies. The essence of the proposed two-stage variance estimation is the jointly modeling of the response and the treatment assignment given the propensity score. This modeling scheme takes the consideration of the estimation error for the parameters of the first stage model and thus the estimation error of the propensity score.

1.3.2 Significance of SNP-PS

The proposed SNP-PS model for dealing with the heterogeneity of treatment allocation provides a uniform approach to adjust the nonparametric heterogeneity confounding in a robust fashion without the restrictive assumptions. Unlike many existing statistical methods, we extend the prevailing propensity score adjustment approach by including a random effect term without specifying any density distribution in the PS model such that only a simple linear regression is needed to assess the treatment effect. Therefore, SNP-PS is more robust to the distribution misspecification of the random effects.

One challenge with the generalized mixed effects model is that there exists no analytic closed form for the log-likelihood function. The double rejection Monte Carlo sampling EM algorithm of Chen et al. (2002) suffers slow convergence and high Monte Carlo errors, which prevents its practical usage for large medical record datasets. Furthermore, the double-rejection sampling scheme of Chen et al. (2002) may have low acceptance rate for binary data when the cluster cell size is imbalanced. Alternatively, we propose a computationally efficient numerical integration approach, i.e. adaptive quadrature integration, to avoid the large Monte Carlo errors if the sampling scheme is used and the potential low acceptance rate of the second rejection sampling component.

In summary, the innovations of the proposed SNP-PS framework for the treatment effect estimate purpose of medical record data includes: First, the proposed SNP-PS approach robustly models the heterogeneity of treatment allocation process and includes the traditional normality assumption as a special case. Second, the proposed adaptive quadrature integration scheme significantly reduces the computational workload and the potential large Monte Carlo errors in the sampling based method for the parameter estimates in generalized mixed effects model. Third, we propose a valid variance estimator for SNP-PS regression model that corrects the biased variance estimation based on the default variance estimator.

1.3.3 Significance of FM-PS

Medical record data could include hundreds of thousands or even millions of medical records integrated from various resources with the treatments assigned by a number of doctors and patients treated at many different clinics or hospitals. The multilevel clustering feature of medical record data further complicates their analysis when treatment effect is intended since it invalidates the crucial independence assumption for the outcomes that many statistical models make. Mixed effects model is the most widely used statistical tool to deal with the clustered and correlated data. However, two important assumptions for the random effects terms made in the conventional mixed effects models are the independence with respective to other fixed effect terms including the treatment assignment and the normality. These assumptions could be too restrictive for the real world medical record data. We propose a flexible linear mixed effects propensity score model to deal with the complicated correlation structures of medical record data conveniently and relax these two assumptions by establishing the equivalence of treatment effect estimates obtained from FM-PS and the dummy variable fixed effects PS (FE-PS) regression model. This equivalence not only relaxes the assumption on the independence between the random effect terms and the fixed effect terms in linear mixed effects model but also the normality assumption for the random effect term.

The novelties of the proposed FM-PS model include: First, it relaxes the very restrictive independence assumption made in mixed effects model for the random effects terms with respect to other fixed effect terms in modeling the clustered data. Second, it flexibly models the complicated correlation among the confounding covariates of medical record data by incorporating them in the propensity score model and including the proportion of subjects in the treated group and the mean of the estimated propensity score of each cluster in a simple linear mixed effects model without the burdens to model the details of the covariance structure among the random effect and other covariates. Third, FM-PS is computationally efficient as compared to the FE-PS method which is prohibited in the existence of hundreds thousands clusters like many medical record data. Therefore, FM-PS is applicable in practice for the huge clustered dataset. Fourth, a novel cluster bootstrapping procedure to obtain the valid variance estimation for the treatment effect estimate from FM-PS is proposed. Furthermore, a likelihood ratio test statistic is proposed to allow for selecting the efficient and unbiased treatment effect estimate from between FM-PS and SM-PS models.

1.4 Outline of The Remaining of Dissertation

Due to the limitations of existing statistical methods, in this dissertation, new statistical methods are developed for treatment effectiveness inference. The methods focus on addressing some issues of heterogeneity in treatment assignment and multilevel clustering settings for medical record data in comparative effectiveness research. It is the goal to develop robust and efficient statistical methods to deal with these limitations under these scenarios by considering the complicated data structures and relaxing some unrealistic assumptions for real applications. In addition, the statistical properties for the proposed methods and models are extensively explored. Overall, the structure of this dissertation is arranged as the following:

Chapter One: provides the introduction of medical record data, the existing statistical methods for the analysis of these data, and the proposed methods due to the limitations of existing methods.

Chapter Two: describes the details of the proposed two-stage variance estimator for PS regression models and its performance under finite sample settings via simulation studies. The asymptotic results of PS regression models are established in this chapter.

Chapter Three: delineates the details of the proposed semi-nonparametric propensity score (SNP-PS) approach for clustered observational data and its performance under finite sample settings. Practical application of SNP-PS regression method is demonstrated via a real data analysis. In addition, the mathematical derivations for the properties of SNP-PS under large sample scenarios are given in this chapter.

Chapter Four: presents a flexible linear mixed effects propensity score (FM-PS) model for multilevel clustered data. Simulation studies are used to demonstrate FM-PS performance under finite sample setups and different model misspecification settings. Justification for FM-PS approach to model the multilevel clustered data for treatment effectiveness inference is given via cluster bootstrap resampling scheme.

Chapter Five: discuss the future research for the proposed methods and the potential extensions.

Chapter 2

A Robust Two-Stage Variance Estimation for PS Regression Analysis

2.1 Introduction

Randomized controlled trial (RCT) is regarded as the gold standard to establish causal effects of drug efficacy. In completely randomized clinical trials, treatments are randomly assigned and the proper randomization procedure guarantees that there exists no systematic difference among subjects for their baseline features in different treatment groups. However, RCTs have their limitations and can not always be conducted in practice (e.g. Kramer and Shapiro, 1984; Johnston et al, 2006). Data from observational studies or electronic medical records, on the other hand, are readily available and often used as alternatives for the evaluation of therapy effectiveness and drug efficacy.

The key difference between data from the observational study and RCT is that the treatment assignment under the real world observational design is not random. Instead, the treatment allocation of each subject is primarily made by researchers according to the subject's characteristics (e.g. physical condition, disease severity, etc). The non-random treatment assignment could lead to imbalanced baseline characteristics. If unaccounted for, this could bias the treatment effect estimate and result in the problem known as confounding issue that could lead to erroneous scientific conclusions. Among proposed methods for confounding adjustment, multiple regression and various version of propensity score (PS) methods are commonly used in practice.

If the multiple regression model is correctly specified, the treatment effect estimator, i.e. least square estimate (LSE), provides the best linear unbiased estimate (BLUE) for the treatment effect, which can be regarded as the benchmark for evaluating the efficacy of different interventions and procedures in comparative effectiveness research. Alternatively, the PS method (Rosenbaum and Rubin 1983) provides a simple and straightforward way to control confounding factors in non-randomized settings. PS methods are increasingly used as alternatives of multiple regression (e.g. Czajka et al. 1992; Schneeweiss et al. 2002; Bang and Robins 2005) due to its simplicity and robustness. A propensity score is defined as the conditional probability of a subject being assigned to a particular treatment in a study given a set of observed covariates. Rosenbaum and Rubin (1983) had shown that under certain conditions, the distributions of measured baseline covariates for the treated and untreated subjects are similar conditional on any given propensity score. That is, subjects with the same propensity scores, they have the same distributions for the baseline covariates no matter if they come from the treated or untreated groups. They demonstrated that if the treatment assignment is strongly ignorable (i.e. condition (1.3) of Rosenbaum and Rubin, 1983), conditional on propensity scores allows one to obtain an unbiased treatment effect estimate.

In practice, the propensity score is unknown and commonly estimated via the logistic regression model. Once the propensity scores are estimated, the treatment effect can be estimated via matching (e.g. Tanasescu et al. 2002; Neily et al. 2010; Rothberg et al. 2010), stratification (e.g. Rosenbaum and Rubin 1984; He and McDermott 2012), inverse-probability-weighting (IPW) (e.g. Do and Finch 2008), or covariate adjustment (e.g. D'Agostino 1998). PS regression analysis where the estimated propensity score used as a covariate in the second stage regression model is frequently used in clinical research (e.g. Wang and Donnan 2001; Weitzen et al. 2004). Analysis results based on the PS regression constantly appear in top scientific journals such as JAMA and NEJM (e.g. Koch et al. 2008; Shaw et al. 2008; Eklind-Cervenka et al. 2011; Jackson et al. 2012; Bangalore et al. 2012).

While the PS covariate adjustment provides an efficient and robust treatment effect estimate, the variance estimation of the treatment effect estimate from the standard PS regression model is biased. In practice, researchers routinely base their inference on the default variance estimate from the second stage regression model, ignoring the fact that the PS is a estimated quantity. We recognize that the PS regression approach can be viewed as a two-stage procedure used in a wide class of empirical applications where unobserved regressors, such as expectations, are estimated from an auxiliary statistical model. It is well known that the second-step estimated standard errors and related test statistics based on these procedures are incorrect (Murphy and Topel 1985). In this chapter, we jointly model the distribution of the response and treatment assignment given the propensity scores and develop a simple yet general method for calculating asymptotic standard errors in two-stage models for PS regression analysis. The joint modeling scheme resolves the biased variance estimation issue in the standard PS regression model by taking into account the stochastic errors in the parameter estimates when the propensity scores are estimated.

Similar concerns on the variance estimates have been noticed for other PS-based methods and improved variance estimators have been proposed for PS inverse probability weighting (e.g. Lunceford and Davidian 2004; Williamson et al. 2012), and matching (e.g. Abadie and Imbens 2011). This chapter fills in a gap in variance estimation for the PS regression method. Our simulation results further demonstrate the importance of using the proposed two-stage regression model in practical comparative effectiveness research.

The rest of this chapter is organized as follows. In Section 2, we provide some background introductions on existing PS methods. In Section 3, we introduce the PS regression model under the two-stage analysis framework. We then derive the asymptotic result under the proposed two-stage PS regression scheme. Based on the asymptotic result, we propose a robust and improved variance estimator. In Section 4, we conduct simulations to evaluate the finite sample performance of the proposed variance estimator under various confounding settings and model misspecification scenarios. In addition, to appreciate the usefulness of PS regression method in comparative effectiveness research, we compare the performance of our proposed two-stage PS regression method with other existing PS methods, in along with the benchmark method, i.e. multiple regression. We then apply the proposed method to a real data analysis in Section 5 to demonstrate the practical application. We end the chapter with discussions in Section 6.

2.2 Existing Methods

We first introduce some notations. Let y_i be the response variable, trt_i be the binary treatment assignment status (1 for treated and 0 for untreated), and $\mathbf{x}_i$ be other observed covariates with dimension of p for individual i ($i = 1, \cdots, n$). The observation for each subject consists of $(y_i, trt_i, \mathbf{x}_i)$. In this chapter, we focus on the situation where the response variable is continuous and depends on treatment assignment with effect of α_{trt} and other observed covariates $\mathbf{x}$:

$$E(y \mid trt, \mathbf{x}) = \alpha_0 + \alpha_{trt} * trt + \mathbf{x} * \boldsymbol{\alpha}_{\mathbf{x}}$$

Our primary interest is the treatment effect, α_{trt}, estimate and its variance estimation.

Under the counterfactual framework of Rosenbaum and Rubin (1983), the primary interest is the so-call average treatment effect $\Delta \equiv E(y(1)) - E(y(0))$ where $y(1)$ and

$y(0)$ are the potential outcomes if the subject had been assigned to the treated and untreated group, respectively. In reality, $y(1)$ and $y(0)$ can not be observed simultaneously. Instead, they are related to the observed response y as: $y = trt * y(1) + (1 - trt) * y(0)$. Under the strongly ignorable treatment assignment assumption of Rosenbaum and Rubin (1983), i.e. $(y(1), y(0)) \perp trt \mid \mathbf{x}$, there exist equivalence between Δ and α_{trt} which can be easily checked: $\alpha_{trt} = E(y \mid trt = 1, \mathbf{x}) - E(y \mid trt = 0, \mathbf{x}) \Rightarrow \alpha_{trt} = E(\alpha_{trt}) = E(E(y \mid trt = 1, \mathbf{x})) - E(E(y \mid trt = 0, \mathbf{x})) = E(E(y(1) \mid \mathbf{x})) - E(E(y(0) \mid \mathbf{x})) = E(y(1) \mid \mathbf{x}) - E(y(0) \mid \mathbf{x}) = \Delta$.

With the strongly ignorable assumption, the confounding factors can be controlled via a simple statistic, i.e. the propensity score (PS). PS is defined as the conditional probability for a unit to receive the treatment given all other observed covariates, i.e. $PS_i \equiv Pr(trt_i = 1 \mid \mathbf{x}_i)$. That is, under the strongly ignorable assumption, we have: $(y(1), y(0)) \perp trt \mid PS(\mathbf{x})$ as shown by Rosenbaum and Rubin (1983). In practice, the propensity score is unknown and usually estimated by the logistic regression model: $logit(PS) = \mathbf{x}\boldsymbol{\beta}$. With the parameters $\boldsymbol{\beta}$ estimated, the propensity score can be estimated as:

$$\widehat{PS}_i = \frac{\exp{(\mathbf{x}_i\hat{\boldsymbol{\beta}})}}{1 + \exp{(\mathbf{x}_i\hat{\boldsymbol{\beta}})}}$$

and the treatment effect can be estimated via PS matching, stratification, IPW, or covariate adjustment.

Under the standard PS regression framework, the treatment effect estimate and its corresponding variance estimation are obtained by fitting the following simple linear regression model:

$$y_i = \alpha_0 + \alpha_{trt} trt_i + \alpha_1 \widehat{PS}_i + \epsilon_i \tag{2.1}$$

where $\widehat{PS}_i$ is the estimated propensity scores. We denote the treatment effect estimator from the standard PS regression model (2.1) as $\hat{\alpha}_{trt,PSR}$. Even though $\hat{\alpha}_{trt,PSR}$ is unbiased under the strongly ignorable and linearity assumptions, the default variance estimate (denoted as $\widehat{Var}(\hat{\alpha}_{trt,PSR})$) from the standard PS regression model (2.1) is biased as will be demonstrated by the extensive simulation studies in Section 2.4.

2.3 Two-Stage Framework for PS Regression

To delineate the proposed variance estimator, we first express the PS regression analysis as a two-stage regression model shown below:

Stage 1 Model: $trt_i \mid \mathbf{x}_i \sim f(trt_i \mid \mathbf{x}_i, \boldsymbol{\theta}_1)$

Stage 2 Model: $(y_i, trt_i) \mid PS(\mathbf{x}_i, \boldsymbol{\theta}_1) \sim f(y_i \mid trt_i, PS(\mathbf{x}_i, \boldsymbol{\theta}_1), \boldsymbol{\theta}_2) f(trt_i \mid PS(\mathbf{x}_i, \boldsymbol{\theta}_1))$

where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are the parameter sets associated with Stage 1 and 2 models. The unknown PS score, $PS(\mathbf{x}, \boldsymbol{\theta}_1)(= Pr(trt = 1 \mid \mathbf{x}, \boldsymbol{\theta}_1))$, is a function of $\mathbf{x}$ and $\boldsymbol{\theta}_1$.

The log-likelihood of Stage 1 model is

$$\begin{aligned} l_1(\boldsymbol{\theta}_1) &= \sum_{i=1}^{n} l_{1,i}(\boldsymbol{\theta}_1) = \sum_{i=1}^{n} \log f(trt_i \mid \mathbf{x}_i) \\ &= \sum_{i=1}^{n} \{trt_i * \log(PS_i) + (1 - trt_i) * \log(1 - PS_i)\} \end{aligned} \tag{2.2}$$

where $PS_i = Pr(trt_i = 1 \mid \mathbf{x}_i, \boldsymbol{\theta}_1)$. If the logistic regression model is used in Stage 1 model, then $PS_i = \frac{\exp(\beta_0 + \mathbf{x}_i\boldsymbol{\beta})}{1+\exp(\beta_0 + \mathbf{x}_i\boldsymbol{\beta})}$ and $\boldsymbol{\theta}_1 = (\beta_0, \boldsymbol{\beta})'$.

The log-likelihood of Stage 2 model is

$$\begin{aligned} l_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &= \sum_{i=1}^{n} l_{2,i}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sum_{i=1}^{n} \log f(y_i, trt_i \mid PS_i) \\ &= \sum_{i=1}^{n} \log \{f(y_i \mid trt_i, PS_i, \boldsymbol{\theta}_2) f(trt_i \mid PS_i)\} \\ &= \sum_{i=1}^{n} \log f(y_i \mid trt_i, PS_i, \boldsymbol{\theta}_2) + \sum_{i=1}^{n} l_{1,i}(\boldsymbol{\theta}_1). \end{aligned} \tag{2.3}$$

$f(y_i \mid trt_i, PS_i, \boldsymbol{\theta}_2) = \phi(y_i; \alpha_0 + \alpha_{trt} trt_i + \alpha_1 PS_i, \sigma^2)$ where function $\phi(y; \mu, \sigma^2)$ is the normal density function with mean μ and variance σ^2. Thus, $\boldsymbol{\theta}_2 = (\alpha_0, \alpha_{trt}, \alpha_1, \sigma^2)'$.

In Stage 1 model, the parameter $\boldsymbol{\theta}_1$ and PS score, PS_i, are first estimated via equation (2.2) as $\hat{\boldsymbol{\theta}}_1 = argmax_{\boldsymbol{\theta}_1} l_1(\boldsymbol{\theta}_1)$ and $\widehat{PS}_i = PS(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_1)$, respectively. Then the estimated PS scores are plugged into equation (2.3) for estimating $\boldsymbol{\theta}_2$, which contains the treatment effect α_{trt}, our primary interest. Specifically, in the second stage analysis, $\hat{\boldsymbol{\theta}}_2 = argmax_{\boldsymbol{\theta}_2} l_2(\hat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2)$. Note that when PS_is are replaced by $\widehat{PS}_i$s, maximizing the likelihood $l_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ reduces to maximizing the likelihood of the following simple linear regression model:

$$y_i = \alpha_0 + \alpha_{trt} trt_i + \alpha_1 \widehat{PS}_i + \epsilon_i,$$

i.e., the standard PS regression model. We note that the term $l_{1,i}(\boldsymbol{\theta}_1)$ in $l_{2,i}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ plays no role in estimating $\boldsymbol{\theta}_2$ but it is a critical term for the variance estimation. Ignoring this term will result in variance estimate very close to the one given by the standard PS regression model (2.1), i.e. $\widehat{Var}(\hat{\alpha}_{trt,PSR})$, which tends to be biased. We denote the treatment effect estimator for parameter α_{trt} as $\hat{\alpha}_{trt,P}$ under our proposed two-stage framework. Under some regularity conditions, we have the following asymptotic results for the treatment effect estimate $\hat{\alpha}_{trt,P}$:

Theorem 2.3.1. *The treatment effect estimator $\hat{\alpha}_{trt,P}$ is asymptotically normally distributed with,*

$$\sqrt{n}(\hat{\alpha}_{trt,P} - \alpha^*_{trt}) \to N(0, \sigma^2_{22})$$

where σ^2_{22} is the second diagonal element of the covariance matrix $\boldsymbol{\Sigma}$ given as the following:

$$\boldsymbol{\Sigma} = \mathbf{V}_2 + \mathbf{V}_2[\mathbf{C}\mathbf{V}_1\mathbf{C}^T - \mathbf{R}\mathbf{V}_1\mathbf{C}^T - \mathbf{C}\mathbf{V}_1\mathbf{R}^T]\mathbf{V}_2$$

with $\boldsymbol{V}_1^{-1} = E\left\{\left(\frac{\partial l_1}{\partial \boldsymbol{\theta}_1}\right)\left(\frac{\partial l_1}{\partial \boldsymbol{\theta}_1^T}\right)\right\}$, $\boldsymbol{V}_2^{-1} = E\left\{\left(\frac{\partial l_2}{\partial \boldsymbol{\theta}_2}\right)\left(\frac{\partial l_2}{\partial \boldsymbol{\theta}_2^T}\right)\right\}$, $\boldsymbol{C} = E\left\{\left(\frac{\partial l_2}{\partial \boldsymbol{\theta}_2}\right)\left(\frac{\partial l_2}{\partial \boldsymbol{\theta}_1^T}\right)\right\}$, *and* $\boldsymbol{R} = E\left\{\left(\frac{\partial l_2}{\partial \boldsymbol{\theta}_2}\right)\left(\frac{\partial l_1}{\partial \boldsymbol{\theta}_1^T}\right)\right\}$. *Under the strongly ignorable condition (1.3) and the linearity assumption in COROLLARY 4.3 of Rosenbaum and Rubin (1983),* α^*_{trt} *can be replaced by* α_{trt}*, the true marginal treatment effect.*

The proof of Theorem 2.3.1 can be found in Appendix I. The above theorem provides us a basis for a modified variance estimator. To estimate the covariance matrix $\boldsymbol{\Sigma}$, we propose a sample estimate of $\boldsymbol{\Sigma}$ as the following:

$$\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{V}}_2 + \hat{\boldsymbol{V}}_2[\hat{\boldsymbol{C}}\hat{\boldsymbol{V}}_1\hat{\boldsymbol{C}}^T - \hat{\boldsymbol{R}}\hat{\boldsymbol{V}}_1\hat{\boldsymbol{C}}^T - \hat{\boldsymbol{C}}\hat{\boldsymbol{V}}_1\hat{\boldsymbol{R}}^T]\hat{\boldsymbol{V}}_2$$

where $\hat{\boldsymbol{V}}_1^{-1} = \frac{1}{n}\sum_{i=1}^n \frac{\partial l_{1,i}}{\partial \boldsymbol{\theta}_1}\frac{\partial l_{1,i}}{\partial \boldsymbol{\theta}_1^T} \mid_{\hat{\boldsymbol{\theta}}_1}$, $\hat{\boldsymbol{V}}_2^{-1} = \frac{1}{n}\sum_{i=1}^n \frac{\partial l_{2,i}}{\partial \boldsymbol{\theta}_2}\frac{\partial l_{2,i}}{\partial \boldsymbol{\theta}_2^T} \mid_{\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2}$, $\hat{\boldsymbol{C}} = \frac{1}{n}\sum_{i=1}^n \frac{\partial l_{2,i}}{\partial \boldsymbol{\theta}_2}\frac{\partial l_{2,i}}{\partial \boldsymbol{\theta}_1^T} \mid_{\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2}$ and $\hat{\boldsymbol{R}} = \frac{1}{n}\sum_{i=1}^n \frac{\partial l_{2,i}}{\partial \boldsymbol{\theta}_2}\frac{\partial l_{1,i}}{\partial \boldsymbol{\theta}_1^T} \mid_{\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2}$, respectively. Matrix $\hat{\boldsymbol{\Sigma}}$ provides a good estimate to $\boldsymbol{\Sigma}$ as can be demonstrated by the simulation studies below. The proposed variance estimator of the treatment effect estimate $\hat{\alpha}_{trt,P}$ is $\hat{\sigma}^2_{22}$ which is the second diagonal element of covariance matrix $\hat{\boldsymbol{\Sigma}}$.

2.4 Simulation Studies

To evaluate the performance of the proposed two-stage regression estimator versus other existing estimators, we conduct simulation studies under various confounding settings as described in the following. Specifically, we compare the treatment effect estimator using the proposed two-stage model, i.e. $\hat{\alpha}_{trt,P}$, versus the estimates from PS stratification ($\hat{\alpha}_{trt,S}$), IPW ($\hat{\alpha}_{trt,IPW}$), multiple regression ($\hat{\alpha}_{trt,LSE}$), and the standard PS regression ($\hat{\alpha}_{trt,PSR}$).

Treatment allocation model: In our first set of simulations, the treatment assignment is based on the following mechanism:

$$\text{logit}\{PS_i(= P(trt_i = 1 \mid \mathbf{x}_{1i}))\} = \mathbf{x}_{1i}\boldsymbol{\beta}_x, \qquad (2.4)$$

where the confounding covariate vectors $\mathbf{x}_{1i} = (\mathbf{x}_{1i,1}, \cdots, \mathbf{x}_{1i,p})^T$ have p independent covariates. We let half of the confounding covariates be binary and the other half be continuous. For the binary variables, they follow $Bern(0.5)$ and the continuous variables are simulated from $N(0, 1)$. For the ease of description, we let each parameter $\beta_{x,j}$ in $\boldsymbol{\beta}_x = (\beta_{x,1}, \cdots, \beta_{x,p})'$ be randomly generated from $N(0, 1)$ but they are fixed for all the subsequent simulations.

Data generating model: The responses are generated based on the following data generating process:

$$y_i = \alpha_{trt} trt_i + \mathbf{x}_{1i}\boldsymbol{\alpha}_x + \epsilon_i, \quad (i = 1, \cdots, n) \tag{2.5}$$

where $\epsilon_i \sim N(0, 1)$ and is independent of trt_i and $\mathbf{x}_{1i}$. The true treatment effect α_{trt} is fixed at 0.5. The effect of each confounding covariate, $\alpha_{x,j}$ $(j = 1, \cdots, p)$, is also generated from $N(0, 1)$.

In observational studies, there may exist many covariates and we may not know which are the true confounding factors and which are not. It is common to include most if not all of them in the analysis. To mimic the real world observational data, in addition to the (true) confounding covariates $\mathbf{x}_{1i}$, we also simulate an additional set of q $(q \geq 0)$ nuisance variables, $\mathbf{x}_{2i}$. Again, among the q nuisance variables, half of them are generated from $Bern(0.5)$ and the other half follow $N(0, 1)$. These nuisance variables have no effects on either the response variable y or the treatment assignment trt. However, we include the observed covariates $\mathbf{x}_i = (\mathbf{x}_{1i}^T, \mathbf{x}_{2i}^T)^T$ in all analysis to mimic the data analysis in practice. For cases where $q = 0$, only the true confounding covariates are included, the best scenario that could happen in practice.

Table 2.1 summarizes simulation results with varying sample size n, the number of confounding covariates p and the number of nuisance variables q. For each simulation setup, the results are based on 1000 simulations. Column "Mean($\hat{\alpha}_{trt}$)" represents the average treatment effect estimate, column "Monte Carlo SD($\hat{\alpha}_{trt}$)" shows the Monte Carlo standard deviation of the treatment effect estimate, and column "Average SE($\hat{\alpha}_{trt}$)" presents the average standard error for the treatment effect estimate.

Inspecting Table 2.1 reveals that the treatment effect estimate is very close to the true effect size 0.5 in all situations indicating the unbiasedness of treatment effect estimate under different analysis schemes. The value in column "Monte Carlo SD($\hat{\alpha}_{trt}$)" can be viewed as the true error of treatment effect estimate for each estimator. A closer inspection of this column for the first part of Table 2.1 clearly shows that the standard deviations for the LSE, the standard PS regression and the proposed two-stage regression are all close to each other. This indicates that the PS regression based treatment effect estimators can be almost as efficient as the one from the multiple regression method. A further inspection of this column reveals that both PS stratification and IPW methods provide less efficient estimates.

Most importantly, comparing the column of "Average SE($\hat{\alpha}_{trt}$)" with the column of "Monte Carlo SD($\hat{\alpha}_{trt}$)" for the standard PS regression estimator, i.e. $\hat{\alpha}_{trt,PSR}$, we notice that there exist big differences. The former is consistently larger than the latter, indicating the variance estimation from the standard PS regression is upwardly biased. This problem remains unchanged even when the sample size increases. For example, in the scenario of sample size 500, $p = 10$, and $q = 0$, the mean standard error 0.276 is more than doubled from the standard deviation 0.119. When the sample size is increased to 5000, the mean standard error 0.086 is still more than double of the standard deviation 0.037. The invalid variance estimation of the standard PS regression model is also reflected in the 95% confidence interval (CI) coverage which is always far away from 95% in all simulation settings. In contrast, the standard error based on the proposed two-stage framework is always very close to the standard deviation in all simulation settings no matter if the sample size is large or not. Furthermore, the corresponding 95% CI coverage for the proposed two-stage estimator, i.e. $\hat{\alpha}_{trt,P}$ is close to 95% in all simulation settings and further confirms the unbiased variance estimation.

To further investigate the robustness property of the proposed variance estimator, we conducted two additional simulations by varying the data generating model and the treatment allocation model. Specifically, we consider the following two model misspecification scenarios:

Misspecification of PS model: we keep the data generating model (2.5) the same but change the treatment assignment from the logistic regression model (2.4) to the following allocation mechanism:

$$\text{logit}\{PS_i(= P(trt_i = 1 \mid \mathbf{x}_{1i}))\} = \beta_0 + \beta_1 x_{1i,1}^2 + \mathbf{x}_{1i}\boldsymbol{\beta}_x$$

where $\boldsymbol{\beta}_x = (\beta_{x,1}, \beta_{x,2}, \cdots, \beta_{x,p})'$ and $\mathbf{x}_{1i} = (x_{1i,1}, x_{1i,2}, \cdots, x_{1i,p})$, respectively. That is, the first continuous covariates $x_{1i,1}$s affect the treatment assignment trt both linearly and quadratically. However, when estimating the propensity scores, we only include the linear terms of all observed covariates $\mathbf{x}_i$ in our analysis. In this sense, the PS model is misspecified but not the multiple regression model.

Misspecification of multiple regression model: we keep the treatment allocation mechanism (2.4) but change the data generating model as follows:

$$y_i = \alpha_t trt_i + \alpha_1 x_{1i,1}^2 + \mathbf{x}_{1i}\boldsymbol{\alpha}_x + \epsilon_i.$$

Here $\boldsymbol{\alpha}_x = (\alpha_{x,1}, \alpha_{x,2}, \cdots, \alpha_{x,p})'$ and $\mathbf{x}_{1i} = (x_{1i,1}, x_{1i,2}, \cdots, x_{1i,p})$, respectively. In this simulation, the first continuous covariates $x_{1i,1}$s affect the response y_i both linearly and quadratically. When fitting the multiple regression model, we still use the linear term of all observed covariates. That is the multiple regression model is misspecified. However, the PS model is not misspecified. For each of the model misspecification scenario, we consider two sample size scenarios, i.e. 500 and $5,000$. The simulation results are presented in Table 2.2.

From the first half of Table 2.2 where the PS model is misfitted, we note that the treatment effect estimates based on different estimators are quite close to the true effect size except the PS stratification estimator. This observation keeps unchanged when the sample size increases from 500 to 5000. It is evident from Table 2.2 that the variance estimation via the standard PS regression is far away from the true variance. However, the variance estimate based on the proposed two-stage method is always close to the true variance.

In the second half of Table 2.2 where one of the covariates affects the response quadratically, we notice that the treatment effect estimates from PS stratification, IPW, and multiple regression methods all suffer biasness. The biasness from IPW gets reduced when the sample size increases while the biasness does not get improved for PS stratification and multiple regression. On the other hand, the treatment effect estimates from the standard PS regression and the proposed two-stage regression framework are unbiased. However, the variance estimation via the standard PS regression is severely upward biased from the true one. In contrast, our proposed method always provides an accurate variance estimate for the true variance.

With the extensive simulation studies, our overall observation is that the treatment effect estimators based on both the standard PS regression method and the proposed two-stage regression scheme provide very accurate treatment effect estimate even when the model is misspecified. They can be as efficient as that of BLUE based on the multiple regression model when the model is correctly specified. However, the variance estimation of the treatment effect estimate under the standard PS regression scheme is always biased. In contrast, the proposed variance estimation based on the two-stage framework is consistently close to the true variance in all scenarios considered.

2.5 Real Data Analysis

To demonstrate the usage of the proposed method, we applied it to the analysis of a breast cancer study by the German Breast Cancer Study Group (Rauschecker et al. 1995). The study originally was intended as a randomized trial but it had to be changed to an observational study due to the low randomization rate. The primary objective for this study was to compare two breast cancer treatment procedures, i.e. the mastectomy ($trt = 0$) versus lumpectomy (breast conservation, $trt = 1$), on the effect of quality of life (QoL) for the breast cancer patients after the surgery. A subset of the data for this study can be obtained from “nonrandom” R package with 646 subjects. The primary outcome was the performance status (PST) 9 months after surgery, which is quantified as a score between 0 and 100 based on the 25 QoL questionnaire responses, where higher scores reflect better QoL. Covariates other than the therapies (i.e. mastectomy vs. lumpectomy) including patient age (Age : ranges from 23 to 82) and tumor size (ts : 1mm $\sim$ 22mm) are considered as potential confounding factors.

We categorized age as young (age: $\leq$ 55) and old (age: $>$ 55) and tumor size as small (ts: $\leq 10mm$) and large (ts: $> 10mm$), respectively as did in Senn et al (2007). Distribution of baseline characteristics for these two covariates among the two treatment groups and each stratum of age and tumor size combination are given in Table 2.3.

First part of Table 2.3 suggests that age and tumor size are somewhat imbalanced between the two treatment groups (59.4 yr v.s. 52.0 yr for mean age and 14.5mm v.s. 13.5mm for mean tumor size). Closer inspection of the second part of Table 2.3 reveals that older patients with larger tumor size favor mastectomy procedure while younger patients with larger tumor size prefer lumpectomy procedure. This suggests that the interaction of age and tumor size plays a role on the treatment assignment process. Thus, we use the following logistic regression model to estimate the propensity score:

$$\text{logit}\{PS_i(= Pr(trt_i = 1 \mid (age_i, ts_i)))\} = \beta_0 + \beta_1 * age_i + \beta_2 * ts_i + \beta_3 * age_i * ts_i$$

Analysis results using different analysis schemes are presented in Table 2.4.

First row of Table 2.4 presents the crude treatment effect estimate of 1.589 without adjusting any confounding covariate. However, after adjusting the confounding factors, the standard PS covariate adjustment, the proposed two-stage framework, and the multiple regression all end up with a sharply reduced treatment effect estimate as 0.793. The treatment effect estimates using the PS stratification (with 4 strata) and IPW scheme are even more sharply reduced to 0.013. All methods give us positive treatment effect estimate indicating that the QoL for patients in the breast conservation group is better than that of mastectomy group. However, this conclusion is not statistically significant which can be easily checked from the corresponding 95% confidence intervals. A further inspection of the standard error from each confounding adjustment scheme also reveals that the standard error based on the proposed two-stage regression framework is the smallest one among all methods compared.

2.6 Discussion

In this chapter, we proposed a new two-stage variance estimator for the treatment effect estimate under the PS regression framework. We jointly modeled the response and treatment assignment under the two-stage analysis framework which is different from the standard PS regression scheme. We established the asymptotic result for the treatment effect estimator based on the two-stage joint modeling scheme. An improved variance estimator was proposed based on the asymptotic result. As shown by our simulations, the variance estimator from the standard PS regression model is biased in general and this will lead to dramatically reduced power for hypothesis testings. This variance estimator did not take into consideration the fact that the PSs used in the second stage regression were estimated with errors. In contrast, the proposed two-stage regression framework took this into consideration and provided an accurate variance estimate.

Our simulation studies showed that both the standard PS regression and the proposed two-stage methods can provide as efficient estimate as the multiple regression can. We further demonstrated that the proposed variance estimator is robust under various model misspecification settings. Accurate treatment effect estimate can still be obtained using both the standard PS regression and the proposed two-stage framework even when the model is misspecified while other methods including the multiple regression may fail. These simulation results demonstrated the importance of using the proposed variance estimator in practical comparative effectiveness research. The proposed method has been implemented in a simple R function which can be freely available from our website at **http://www.bios.unc.edu/∼bzou/TwoStagePS/**.

Table 2.1: Simulation Results Under Settings (2.4) & (2.5)

# of covariates				Monte Carlo	Average	95% CI
p	q	Estimator	Mean($\hat{\alpha}_{trt}$)	SD($\hat{\alpha}_{trt}$)	SE($\hat{\alpha}_{trt}$)	Coverage
Sample Size n=500						
10	0	$\hat{\alpha}_{trt,S}$	0.501	0.273	0.301	94.7
		$\hat{\alpha}_{trt,IPW}$	0.496	0.341	0.245	92.0
		$\hat{\alpha}_{trt,LSE}$	0.498	0.115	0.114	94.2
		$\hat{\alpha}_{trt,PSR}$	0.500	0.119	0.276	100.0
		$\hat{\alpha}_{trt,P}$	0.497	0.119	0.117	94.6
10	10	$\hat{\alpha}_{trt,S}$	0.499	0.279	0.305	95.4
		$\hat{\alpha}_{trt,IPW}$	0.487	0.319	0.250	92.6
		$\hat{\alpha}_{trt,LSE}$	0.499	0.113	0.115	95.9
		$\hat{\alpha}_{trt,PSR}$	0.499	0.115	0.279	100.0
		$\hat{\alpha}_{trt,P}$	0.499	0.115	0.115	94.0
Sample Size n=5000						
10	0	$\hat{\alpha}_{trt,S}$	0.506	0.070	0.103	99.9
		$\hat{\alpha}_{trt,IPW}$	0.501	0.104	0.095	95.6
		$\hat{\alpha}_{trt,LSE}$	0.500	0.036	0.036	94.5
		$\hat{\alpha}_{trt,PSR}$	0.500	0.037	0.086	100.0
		$\hat{\alpha}_{trt,P}$	0.500	0.037	0.036	93.7
10	10	$\hat{\alpha}_{trt,S}$	0.503	0.076	0.103	99.4
		$\hat{\alpha}_{trt,IPW}$	0.499	0.103	0.093	93.5
		$\hat{\alpha}_{trt,LSE}$	0.498	0.036	0.036	94.4
		$\hat{\alpha}_{trt,PSR}$	0.498	0.037	0.086	100.0
		$\hat{\alpha}_{trt,P}$	0.498	0.037	0.036	94.0

Note: p is the # of confounding covariates
q is the # of nuisance covariates
$\hat{\alpha}_{trt,S}$ is PS stratification estimator with 5 equally divided strata
$\hat{\alpha}_{trt,IPW}$ is IPW estimator
$\hat{\alpha}_{trt,LSE}$ is the least square estimator
$\hat{\alpha}_{trt,PSR}$ is PS regression estimator
$\hat{\alpha}_{trt,P}$ is proposed estimator

Table 2.2: Simulation Results Under Model Misspecification

Estimator	Mean($\hat{\alpha}_{trt}$)	Monte Carlo SD($\hat{\alpha}_{trt}$)	Average SE($\hat{\alpha}_{trt}$)	95% CI Coverage
Quadratic term in treatment assignment but PS misfitted with linear term only Sample Size n=500				
$\hat{\alpha}_{trt,S}$	0.462	0.268	0.293	93.5
$\hat{\alpha}_{trt,IPW}$	0.491	0.305	0.244	91.7
$\hat{\alpha}_{trt,LSE}$	0.497	0.112	0.115	95.3
$\hat{\alpha}_{trt,PSR}$	0.497	0.114	0.276	100.0
$\hat{\alpha}_{trt,P}$	0.497	0.114	0.126	96.4
Sample Size n=5000				
$\hat{\alpha}_{trt,S}$	0.460	0.075	0.103	99.1
$\hat{\alpha}_{trt,IPW}$	0.493	0.105	0.090	93.2
$\hat{\alpha}_{trt,LSE}$	0.499	0.035	0.036	95.4
$\hat{\alpha}_{trt,PSR}$	0.498	0.036	0.086	100.0
$\hat{\alpha}_{trt,P}$	0.498	0.036	0.039	96.6
Quadratic term in response but multiple regression misfitted with linear term only Sample Size n=500				
$\hat{\alpha}_{trt,S}$	0.449	0.374	0.341	92.3
$\hat{\alpha}_{trt,IPW}$	0.468	0.454	0.342	91.2
$\hat{\alpha}_{trt,LSE}$	0.449	0.239	0.247	95.2
$\hat{\alpha}_{trt,PSR}$	0.483	0.246	0.316	98.6
$\hat{\alpha}_{trt,P}$	0.483	0.246	0.252	95.1
Sample Size n=5000				
$\hat{\alpha}_{trt,S}$	0.459	0.106	0.118	97.0
$\hat{\alpha}_{trt,IPW}$	0.501	0.147	0.134	95.8
$\hat{\alpha}_{trt,LSE}$	0.467	0.076	0.078	93.1
$\hat{\alpha}_{trt,PSR}$	0.501	0.077	0.098	99.0
$\hat{\alpha}_{trt,P}$	0.501	0.077	0.079	95.5

Note: see Table 2.1 for the table legends

Table 2.3: Baseline Characteristics for German Breast Cancer Study Data

	Mastectomy			Lumpectomy		
	N	Mean	SD	N	Mean	SD
Age (year)	167	59.4	10.4	479	52.0	11.5
Tumor Size (mm)	167	14.5	3.6	479	13.5	4.4

		Mastectomy		Lumpectomy	
		N	Proportion	N	Proportion
Young ($\leq$ 55 yr)	Small ($\leq$ 10 mm)	7	0.042	88	0.184
Young ($\leq$ 55 yr)	Large ($>$ 10 mm)	49	0.293	206	0.430
Old ($>$ 55 yr)	Small ($\leq$ 10 mm)	23	0.138	42	0.088
Old ($>$ 55 yr)	Large ($>$ 10 mm)	88	0.527	143	0.298
Total		167	1.000	479	1.000

Table 2.4: Analysis Results for German Breast Cancer Study Data

Estimator	$\hat{\alpha}_{trt}$	SE($\hat{\alpha}_{trt}$)	95% CI
$\hat{\alpha}_{trt,Unadj}$	1.589	1.261	(-0.883,4.061)
$\hat{\alpha}_{trt,S}$	0.013	1.408	(-2.747,2.773)
$\hat{\alpha}_{trt,IPW}$	0.013	1.417	(-2.764,2.790)
$\hat{\alpha}_{trt,LSE}$	0.793	1.299	(-1.753,3.339)
$\hat{\alpha}_{trt,PSR}$	0.793	1.302	(-1.759,3.345)
$\hat{\alpha}_{trt,P}$	0.793	1.248	(-1.653,3.239)

Note: $\hat{\alpha}_{trt,Unadj}$ is the estimator without adjusting any confounding factor see Table 2.1 for the table legends

Chapter 3

A Semi-Nonparametric PS Model for Clustered Observational Data

3.1 Introduction

In completely randomized clinical trials (RCT), treatments are randomly assigned. Proper randomization procedures guarantee that there exists no systematic baseline difference among subjects from different treatment groups. Therefore, RCT is regarded as the scientific standard to establish the causal effect for a treatment and routinely used in clinical trials for assessing drug efficacy. However, RCTs have their limitations (e.g. Kramer and Shapiro 1984; Johnston et al. 2006) and can not always be conducted in practice. With the ever readily availability of large clinical datasets, especially the electronic medical record data, various efforts have been made to look into those datasets in comparing the effectiveness of different treatments. Real world clinical data offer a broader population spectrum as well as longer time-intervals than a typical RCT data (e.g. Benson and Hartz 2000). They tend to reflect daily clinical practice more closely and provide more clinically relevant information than RCTs (e.g. Yang et al. 2010). The cost of using observational clinical data for comparative effectiveness research (CER) is often much lower than that of using RCTs. Furthermore, large medical record database provide an important resource to detect rare adverse events that can not be detected during randomized controlled trials. Proper use of medical record data for the comparative effectiveness research provides investigators an effective way to compare various interventions for the inferiority, equivalence, or superiority (e.g. Mitka 2010).

However, treatment allocation in practical clinical data tends to be not random. Rather, the treatment assignment could be influenced by several factors that could include patient's, physician's, and health care system's factors. This nonrandomness in treatment assignment could create imbalanced baseline covariates, i.e. confounding variables, and result in severe biased treatment effect estimate if the confounding factors are not appropriately adjusted. Several methods have been proposed to address the confounding problems in observational studies, such as matching (e.g. Miettinen 1968), stratification (e.g. Cochran 1968), the instrumental variable approach (e.g. Angrist et al. 1996) and the propensity score method (e.g. Rosenbaum and Rubin 1983). Among these methods, the propensity score (PS) approach is one of the leading ones commonly used in practice.

For real world medical record data, the treatment assignment is clustered by physicians, clinics and so on. They come from various sources that include insurance claim data, hospital and clinic prescription records, etc. Generally, the particular treatment a patient assigned depends on many factors which include but are not limited to: (1) physician's factors which include physician's professional training, experience, practice style and his/her determination on each patient's reception to a particular treatment; (2) patient's factors which include patient's prognostic status, his/her personal preference on different treatments; and (3) system's factors which include insurance policy, hospital policy, etc. All these factors may reflect samples heterogeneity, for example, in terms of patients' health condition, social economical status which may affect both treatment allocation process and disease outcomes.

In current statistical literature, there are a few methods dealing with the heterogeneity cluster effects in PS models based on a parametric mixed effects model where the cluster effects are assumed normally distributed (e.g. Thoemmes and West 2011). This heterogeneity could be complex enough to be fully tracked, and existing simple PS regression models may not be flexible enough to address the underlying heterogeneity. For electronic medical data, the normality assumption may not hold and the heterogeneity caused by the incomplete measuring of the underlying dynamics in the real world treatment assignment process could be too complex to be described by a parametric model. Even worse, the heterogeneity is hidden and cannot be observed directly, making model misspecification hard to be diagnosed. Such cluster effect misspecification could lead to invalid estimation of PS scores and subsequently result in biased estimate of the treatment effect and erroneous conclusions.

To better capture the heterogeneity in treatment allocation process and reflect the clustering feature of treatment assignment for real world medical data, we propose a generalized mixed effect propensity score model to deal with the heterogeneity of treatment allocations. We relax the normality assumption on the cluster effect in the generalized mixed effects model with an unspecified cluster effect term. We refer the proposed PS model as "SNP-PS" model (a.k.a. semi-nonparametric PS model).

The proposed SNP-PS model deals with the heterogeneity of treatment allocation in a robust fashion without the parametric assumptions on the random effects. It jointly models the heterogeneity structure in treatment allocation process and the outcome under non-randomized designs. We establish the asymptotic results for the treatment effect estimator under semi-nonparametric propensity score regression framework that allows us to develop an improved variance estimator for the treatment effect estimate. Simulation results reveal that our proposed variance estimator for the treatment effect estimate is unbiased while the commonly used default variance estimator by PS regression method is biased. A new and efficient adaptive integration algorithm is developed to avoid the potential large Monte Carlo errors in assessing the non-closed form log-likelihood function of SNP-PS model for parameters estimation.

The rest of the chapter is arranged as follows. We present the mathematical description of the proposed SNP-PS approach in Section 2. Detailed procedures for the parameter estimates and propensity score estimations are given in Section 3. Asymptotic results for the proposed SNP-PS estimator are outlined in Section 4. We present the simulation studies under different heterogeneity structures in Section 5. We demonstrate the real application of the proposed SNP-PS method in Section 6. Final remarks are given in Section 7. In Appendix II, we present the mathematical derivation for the asymptotic results of SNP-PS estimator. Furthermore, since the sampling from SNP density is not readily available in the existing statistical software package, we also provide the sampling scheme from SNP density in Appendix III.

3.2 Semi-Nonparametric PS Model

Before introducing our semi-nonparametric propensity score model, we first briefly describe the standard propensity score model below.

3.2.1 Standard PS Model

Let y_i, trt_i and $\mathbf{x}_i$ represent the outcome, treatment assignment, and observed baseline covariates of subject i $(i = 1, \ldots, n)$, respectively. Furthermore, y_is are independently identically distributed (iid).

The validity of the PS approach was established by Rosenbaum and Rubin (1983) who showed that under the following conditions,

$$(y(1), y(0)) \perp trt|\mathbf{x} \qquad \text{and} \qquad 0 < Pr(trt = 1|\mathbf{x}) < 1, \tag{3.1}$$

the measured baseline covariates will be similar between the treated and untreated subjects with similar propensity scores ($PS \equiv Pr(trt = 1|\mathbf{x})$). Here $\mathbf{x}$ are the observed covariates. $y(1)$ and $y(0)$ are the potential outcomes of a particular unit if it had been assigned to the treated and untreated group, respectively. They are never observed simultaneously in reality. Their relationship with the observed outcome y and the treatment assignment trt (1 and 0 for treated and untreated assignment, respectively) can be expressed as $y = trt \times y(1) + (1 - trt) \times y(0)$. The first condition in (3.1) says that the treatment assignment is independent of the potential outcomes conditional on the observed baseline covariates. Based on PS, the treatment effect can be estimated via matching (e.g. Tanasescu et al. 2002; Dehejia and Wahba 2002), stratification (e.g. Rosenbaum and Rubin 1984; He and McDermott 2012), inverse-probability-weighting (IPW) (e.g. Do and Finch 2008), or covariate adjustment (e.g. Koch et al. 2008; Shaw et al. 2008). In practice, the propensity score is not observed and the following logistic regression model:

$$logit(Pr(trt = 1 \mid \mathbf{x})) = \beta_0 + \mathbf{x}\boldsymbol{\beta}, \tag{3.2}$$

is often used to estimate the PS as the following:

$$\widehat{PS} = \exp(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}})/[1 + \exp(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}})].$$

where $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ are parameters estimates of β_0 and $\boldsymbol{\beta}$ in model (3.2).

In the above standard PS model, the treatment assignment heterogeneity is not taken into consideration. As described earlier, in practical medical record data, the treatments are non-randomly assigned and clustered, for example, by physician, and/or clinics, etc. To capture the complex heterogeneity structures of treatment assignment process in practical medical record data, we propose the following semi-nonparametric propensity score model.

3.2.2 First Stage Semi-Nonparametric PS Model

Let y_{ij}, trt_{ij} and $\mathbf{x}_{ij}$ represent the outcome, the treatment assignment, and the observed covariates of subject j nested in cluster i ($j = 1, \ldots, n_i$; $i = 1, \ldots, n$), respectively.

We extend the standard PS model (3.2) to the following mixed effects logistic regression model:

$$logit(Pr(trt_{ij} = 1 \mid \mathbf{x}_{ij})) = \beta_0 + \mathbf{x}_{ij}\boldsymbol{\beta} + \eta_i \tag{3.3}$$

where η_i is a random cluster effect. Instead of restricting η to follow a normal distribution, we assume that it follows an unknown smooth density function $f(\eta)$.

Estimation of $f(\eta)$, can be done through various density estimation methods (e.g. Rosenblatt 1956; Parzen 1962; Wahba 1975; Scott 1979; Sheather and Jones 1991; Botev et al. 2010). We adopt Gallant and Nychka (1987) method that uses a truncated Hermite series multiplied by a normal density to estimate $f(\eta)$.

SNP representation: Specifically, we approximate $f(\eta)$ by the following truncated $K + 1$ terms of Hermite polynomial multiplied by a normal density (e.g. Gallant and Nychka 1987; Davidian and Gallant 1993; Zhang and Davidian 2001; Chen et al. 2002) as shown below:

$$f(\eta; \psi, \mu, \sigma^2) \approx f_K(\eta; \psi, \mu, \sigma^2) = H_K^2(\frac{\eta - \mu}{\sigma}; \psi)\phi(\eta; \mu, \sigma^2)$$

where $\phi(\eta; \mu, \sigma^2)$ is the density function of the normal distribution with mean μ and variance σ^2. $H_K(z; \psi)$ is a Hermite polynomial with $K+1$ terms. In practice, most data including skewed, fat, and t-like tail densities can be approximated well by $f_K(\eta; \psi, \mu, \sigma^2)$ with $K \leq 2$ (Gallant and Nychka 1987). The first three Hermite polynomials are given below (details on higher order Hermite polynomials can be found in Gallant and Nychka

1987).

$$H_K(z;\psi) = \begin{cases} 1 & \text{if } K=0 \\ cos(\psi) + zsin(\psi) & \text{if } K=1 \\ (cos(\psi_1) - \frac{sin(\psi_1)sin(\psi_2)}{\sqrt{2}}) + zsin(\psi_1)cos(\psi_2) + z^2\frac{sin(\psi_1)sin(\psi_2)}{\sqrt{2}} & \text{if } K=2 \end{cases}$$

where parameters $(\psi, \psi_1, \psi_2) \in (-\frac{\pi}{2}, \frac{\pi}{2}]$. This approximation provides a fully parametric representation for the completely nonparametric specifications and is termed by Gallant and Nychka (1987) as semi-nonparametric (SNP) representation. Let $\omega = (\mu, \sigma^2)$ (for $K = 0$), or $\omega = (\psi, \mu, \sigma^2)$ (for $K = 1$) or $\omega = (\psi_1, \psi_2, \mu, \sigma^2)$ (for $K = 2$) be the set of parameters used for characterizing the density of η under SNP representation. When $K = 0$, $f_K(\eta;\omega)$ reduces to a normal distribution.

SNP-PS log-likelihood: With the use of SNP representation for the density function, the log-likelihood of the semi-nonparametric logistic regression model (3.3) can be written as:

$$l(\boldsymbol{\beta}, \boldsymbol{\omega}) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta}, \boldsymbol{\omega}) = \sum_{i=1}^{n} \log \left[\int_{-\infty}^{\infty} \prod_{j=1}^{n_i} f(trt_{ij} \mid \mathbf{x}_{ij}, \eta_i; \boldsymbol{\beta}) f(\eta_i; \boldsymbol{\omega}) d\eta_i \right] \tag{3.4}$$

where $f(trt \mid \mathbf{x}, \eta; \boldsymbol{\beta})$ is the logistic density function given by the following:

$$f(trt \mid \mathbf{x}, \eta; \boldsymbol{\beta}) = \frac{\exp(trt * [\mathbf{x}\boldsymbol{\beta} + \eta])}{1 + \exp(\mathbf{x}\boldsymbol{\beta} + \eta)}$$

Note that in the proposed SNP-PS model (3.3), we do not restrict the cluster effect term η to have mean 0. Thus, parameters β_0 and $E(\eta)$ cann't be separated and only $\beta_0 + E(\eta)$ is estimable. For model identifiability purpose, we set $\beta_0 = 0$. No analytic closed form for this log-likelihood is available. To get the maximum likelihood estimate (MLE) of the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\omega})$, we adopt EM algorithm (Dempster et al. 1977) based on an adaptive integration mechanism to evaluate the non-analytic closed form log-likelihood function (detailed in Section 3).

3.2.3 Second Stage Semi-Nonparametric PS Regression Model

After the parameters of SNP-PS model (3.3) are estimated, we estimate the propensity score for each subject used for the subsequent treatment effect estimate. Specifically, we predict the cluster effect of cluster i, i.e. η_i as:

$$\hat{\eta}_i = E_{\eta_i|\{trt_{ij}\}_j}(\eta_i \mid \{trt_{ij}\}_j; \hat{\boldsymbol{\theta}}) = \int_{-\infty}^{\infty} \eta_i f(\eta_i \mid \{trt_{ij}\}_j; \hat{\boldsymbol{\theta}}) d\eta_i$$

Based on it, we get the estimated propensity score of subject j in cluster i as:

$$\widehat{PS}_{ij} = Pr(trt_{ij} = 1 \mid \mathbf{x}_{ij}; \hat{\boldsymbol{\theta}}) = \frac{\exp(\mathbf{x}_{ij}\hat{\boldsymbol{\beta}} + \hat{\eta}_i)}{1 + \exp(\mathbf{x}_{ij}\hat{\boldsymbol{\beta}} + \hat{\eta}_i)} \tag{3.5}$$

With the propensity scores estimated (i.e. $\widehat{PS}_{ij}$), we estimate the treatment effect via the following regression model:

$$y_{ij} = \alpha_0 + \alpha_{trt} trt_{ij} + \alpha_1 \widehat{PS}_{ij} + \epsilon_{ij} \tag{3.6}$$

with $\epsilon_{ij} \sim N(0, \sigma^2)$ being the random measurement error.

3.3 An EM Procedure for Parameter Estimation

To obtain the estimation of the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\omega})$ in (3.4), we adopt EM algorithm by treating the cluster effect η as missing, resulting in the full data as (trt, $\mathbf{x}$, η). The full data likelihood is then given by:

$$\begin{aligned} f(\{trt_{ij}\}_j, \{\mathbf{x}_{ij}\}_j, \eta_i; \boldsymbol{\theta}) &= f(\{trt_{ij}\}_j, \{\mathbf{x}_{ij}\}_j \mid \eta_i; \boldsymbol{\beta}) f(\eta_i; \boldsymbol{\omega}) \\ &= [\prod_{j=1}^{n_i} f(trt_{ij} \mid \mathbf{x}_{ij}, \eta_i; \boldsymbol{\beta})] f(\eta_i; \boldsymbol{\omega}) \end{aligned}$$

where $f(\eta_i;\boldsymbol{\omega})$ is the unspecified heterogeneity density that can be represented via semi-nonparametric method given in the previous section. This leads to the full data log-likelihood given below:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta}) = \sum_{i=1}^{n}\{[\sum_{j=1}^{n_i} \log f(trt_{ij} \mid \mathbf{x}_{ij}, \eta_i; \boldsymbol{\beta})] + \log f(\eta_i;\boldsymbol{\omega})\}.$$

E-step: Along with the above full data log-likelihood, the $(r+1)th$ E-step of EM algorithm for SNP-PS is to evaluate the following conditional expectation of full data log-likelihood:

$$\begin{aligned} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(r)}) &= E_{\eta \mid trt}(l(\boldsymbol{\theta})) = \sum_{i=1}^{n} E_{\eta_i \mid \{trt_{ij}\}_j}(l_i(\boldsymbol{\theta})) \\ &= \sum_{i=1}^{n} \int_{-\infty}^{\infty} \{[\sum_{j=1}^{n_i} \log f(trt_{ij} \mid \mathbf{x}_{ij}, \eta_i; \boldsymbol{\beta})] + \log f(\eta_i;\boldsymbol{\omega})\} \times \\ & \quad f(\eta_i \mid \{trt_{ij}\}_j, \{\mathbf{x}_{ij}\}_j; \boldsymbol{\theta}^{(r)}) d\eta_i \end{aligned}$$

where $\boldsymbol{\theta}^{(r)}$ is the estimates from the $(r)th$ M-step or the initial values of the parameters.

The evaluation of the above conditional expectation is not trivial since there is no analytic closed form. Chen et al. (2002) has proposed a double-rejection Monte Carlo sampling scheme to replace the integration over the cluster effect by the summation over the samples from the conditional density $f(\eta_i \mid \{trt_{ij}\}_j, \{\mathbf{x}_{ij}\}_j; \boldsymbol{\theta}^{(r)})$. However, the double-rejection sampling approach is not efficient for the SNP-PS model due to its potentially large Monte Carlo error and low acceptance rate for binary data. Alternatively, we adopt the numerical integration scheme. By Bayes' law, we can rewrite the conditional density $f(\eta_i \mid \{trt_{ij}\}_j, \{\mathbf{x}_{ij}\}_j; \theta^{(r)})$ as the following:

$$f(\eta_i \mid \{trt_{ij}\}_j, \{\mathbf{x}_{ij}\}_j; \boldsymbol{\theta}^{(r)}) = \frac{[\prod_{j=1}^{n_i} f(trt_{ij} \mid \mathbf{x}_{ij}, \eta_i; \boldsymbol{\beta}^{(r)})] f(\eta_i; \boldsymbol{\omega}^{(r)})}{\int_{-\infty}^{\infty} [\prod_{j=1}^{n_i} f(trt_{ij} \mid \mathbf{x}_{ij}, \eta_i; \boldsymbol{\beta}^{(r)})] f(\eta_i; \boldsymbol{\omega}^{(r)}) d\eta_i}.$$

Based on this new expression of $f(\eta_i \mid \{trt_{ij}\}_j, \{\mathbf{x}_{ij}\}_j; \boldsymbol{\theta}^{(r)})$, we can obtain the numerical evaluation for quantity $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(r)})$ since the analytic expression for $[\prod_{j=1}^{n_i} f(trt_{ij} \mid \mathbf{x}_{ij}, \eta_i; \boldsymbol{\beta}^{(r)})] f(\eta_i; \boldsymbol{\omega}^{(r)})$ is available with a given Hermite expansion term K.

M-step: Once $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(r)})$ is estimated, in M-step, we maximize it to obtain the $(r+1)th$ step parameter estimate of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\omega}) = (\boldsymbol{\beta}, \psi, \mu, \sigma^2)$. Notice that the parameters related to the fixed effects part, i.e. $\boldsymbol{\beta}$, and the cluster effects part, i.e. ψ, μ, σ^2, can be separated and optimized separately. That is, we perform the optimization on the following two functions evaluated in the E-step separately:

$$Q_F(\boldsymbol{\beta} \mid \boldsymbol{\theta}^{(r)}) = \sum_{i=1}^{n} \int_{-\infty}^{\infty} [\sum_{j=1}^{n_i} \log f(trt_{ij} \mid \mathbf{x}_{ij}, \eta_i; \boldsymbol{\beta})] f(\eta_i \mid \{trt_{ij}\}_j, \{\mathbf{x}_{ij}\}_j; \boldsymbol{\theta}^{(r)}) d\eta_i$$

$$Q_R(\boldsymbol{\omega} \mid \boldsymbol{\theta}^{(r)}) = \sum_{i=1}^{n} \int_{-\infty}^{\infty} [\log P_K(z_i; \boldsymbol{\omega}) + \log \phi(\eta_i; \mu, \sigma^2)] f(\eta_i \mid \{trt_{ij}\}_j, \{\mathbf{x}_{ij}\}_j; \boldsymbol{\theta}^{(r)}) d\eta_i$$

where $z_i = \frac{\eta_i - \mu}{\sigma}$ and $P_K(z_i; \boldsymbol{\omega}) = H_K^2(z_i; \boldsymbol{\omega})$.

Specifically, we optimize $Q_F(\boldsymbol{\beta} \mid \boldsymbol{\theta}^{(r)})$ using the L-BFGS-B algorithm (Byrd et al. 1995). The L-BFGS-B algorithm is an extension of the L-BFGS algorithm to handle simple bounds on the model (Zhu et al. 1997). The L-BFGS algorithm is an efficient algorithm and particularly suited for optimization problems with a large number of variables. L-BFGS-B uses the ideas from the trust region methods while keeping the L-BFGS update of the Hessian and line search algorithms.

However, there exists singularity for the derivatives of the cluster effect with respect to $\boldsymbol{\omega}$ when $K > 0$. Therefore, we use Nelder-Mead algorithm (Nelder and Mead 1965) to maximize $Q_R(\boldsymbol{\omega} \mid \boldsymbol{\theta}^{(r)})$. Nelder-Mead algorithm uses only function values and works reasonably well for non-differentiable functions.

In practice, the Hermit expansion term K for the SNP representation is unknown and needs to be determined. Commonly used model selection criteria, for example, *AIC* (Akaike 1974) or *BIC* (Schwarz 1978) can be used for this purpose (Zhang and

Davidian 2001). Specifically, we select the model that minimizes $-\log L(\hat{\theta}; \mathbf{trt}, \mathbf{x}) + c_K$, where

$$-\log L(\hat{\boldsymbol{\theta}}; \mathbf{trt}, \mathbf{x}) = -\sum_{i=1}^{n} \{ \int_{-\infty}^{\infty} [\prod_{j=1}^{n_i} f(trt_{ij} \mid \mathbf{x}_{ij}, \eta_i; \hat{\boldsymbol{\beta}})] f(\eta_i; \hat{\boldsymbol{\omega}}) d\eta_i \},$$

and $c_K = D_K$ or $\frac{1}{2}(\log N) D_K$ (for AIC and BIC, respectively) is a penalization term on the number of parameters used in the model with D_K denoting the dimension of $\boldsymbol{\theta}$ and $N = \sum_{i=1}^{n} n_i$ representing the total number of subjects in the dataset.

In summary, the following procedure summarizes the proposed SNP-PS regression model to obtain the treatment effect estimate for the clustered data:
Step 1: Fit a set of SNP-PS models with different Hermite expansion terms (e.g. $K = 0, 1$, and 2) via the above EM algorithm.
Step 2: Use the model selection criteria AIC or BIC to select the optimum SNP-PS model and estimate the PS for the selected model.
Step 3: Plug in the estimated PS scores into PS regression model (3.6) to obtain the treatment effect estimate $\hat{\alpha}_{trt}$.

To obtain the valid variance estimator for $\hat{\alpha}_{trt}$, we first establish the asymptotic results for $\hat{\alpha}_{trt}$ in the below section. From which we propose an improved variance estimator for the treatment effect estimate under SNP-PS regression framework.

3.4 Asymptotic Results

Before presenting the details of the theoretical results for SNP-PS model, we first introduce the following assumptions:

For samples within each cluster, $(y(1), y(0)) \perp trt|\mathbf{x}$ and $0 < Pr(trt = 1|\mathbf{x}) < 1$

These assumptions are similar to the strongly ignorable treatment assignment assumptions of Rosenbaum and Rubin (1983). The strongly ignorable assumptions in Rosenbaum and Rubin (1983) are conditional on $\mathbf{x}$ for all samples while ours are conditional on $\mathbf{x}$ restricted for samples within each cluster and refered as the modified strongly ignorable treatment assignment assumption. In order words, the cluster effect is treated as an unobserved confounding covariate.

Following the similar argument of Rosenbaum and Rubin (1983), under the above modified strongly ignorable treatment assignment assumptions, $\hat{\alpha}_{trt}$ in model (3.6) can be shown to approximate the true treatment effect well.

However, the default variance estimate of $\hat{\alpha}_{trt}$ obtained directly from model (3.6), on the other hand, is a biased estimate for the (unknown true) variance of $\hat{\alpha}_{trt}$ since it ignores the fact that the PS used in (3.6) are estimated quantities with random errors instead of the observed covariates. It turns out that the PS regression can be viewed as a two-stage procedure that commonly used in a wide class of empirical applications where unobserved regressors, such as expectations, are estimated from auxiliary statistical models. It is well-known that the estimated standard errors and related test statistics directly from the second-step regression are incorrect (Murphy and Topel 1985). With the joint modeling scheme, under the SNP-PS regression framework, we develop the asymptotic result for $\hat{\alpha}_{trt}$ as described below. Based on it, we propose a variance estimator with good finite sample performance as shown in our subsequent simulation studies.

First, we write the log-likelihood for the first stage SNP-PS model (3.3) as the following:

$$l_1(\boldsymbol{\theta}_1) = \sum_{i=1}^{n} l_{1i}(\boldsymbol{\theta}_1) = \sum_{i=1}^{n} \log \left\{ \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} f(trt_{ij} \mid \mathbf{x}_{ij}, \eta_i; \boldsymbol{\beta}) f(\eta_i; \boldsymbol{\omega}) d\eta_i \right\} \text{ with } \boldsymbol{\theta}_1 = (\boldsymbol{\beta}, \boldsymbol{\omega}).$$

For the second stage analysis, we have

$$\begin{aligned}
l_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &= \sum_{i=1}^{n} l_{2i}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \equiv \sum_{i=1}^{n} \log f(\{y_{ij}\}_j; \{trt_{ij}\}_j \mid \{PS_{ij}\}_j) \\
&= \sum_{i=1}^{n} \log \{f(\{y_{ij}\}_j \mid \{PS_{ij}\}_j; \{trt_{ij}\}_j) f(\{trt_{ij}\}_j \mid \{PS_{ij}\}_j)\} \\
&= \sum_{i=1}^{n} \log \{f(\{y_{ij}\}_j \mid \{PS_{ij}\}_j; \{trt_{ij}\}_j)\} + \sum_{i=1}^{n} \log \{f(\{trt_{ij}\}_j \mid \{PS_{ij}\}_j)\}
\end{aligned}$$

with $\boldsymbol{\theta}_2 = (\alpha_0, \alpha_{trt}, \alpha_1, \sigma^2)^T$.

Given the (unknown) true PS score PS_{ij} of each sample and the strong ignorable assumptions, the above log-likelihood can be approximated as

$$l_{2i}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = l_{2i,a}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) + l_{2i,b}(\boldsymbol{\theta}_1)$$

where $l_{2i,a}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = -\frac{n_i}{2}\log(2\pi) - \frac{n_i}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n_i}(y_{ij} - \alpha_0 - \alpha_{trt}trt_{ij} - \alpha_1 PS_{ij})^2$ and $l_{2i,b}(\boldsymbol{\theta}_1) = \sum_{j=1}^{n_i} \log\left(PS_{ij}^{trt_{ij}}(1 - PS_{ij})^{1-trt_{ij}}\right)$

In PS regression model, the PS scores are first estimated via model (3.5), then the estimated PS scores are plugged into model (3.6) to obtain $\hat{\alpha}_{trt}$, the parameter of the primary interest. Note that when PS_{ij}s are replaced by $\widehat{PS}_{ij}$s, maximizing the likelihood $l_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ reduces to maximizing the log-likelihood of the simple linear regression model (3.6) for the purpose of the treatment effect estimates. The term $l_{2i,b}(\boldsymbol{\theta}_1)$ in the above expression for $l_{2i}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ plays a critical role for the variance estimation. Ignoring this term will lead to the biased variance estimate for the parameter estimators. With these preparations, under some regularity conditions, we have the following asymptotic result for $\hat{\alpha}_{trt}$ within the SNP-PS regression framework:

Theorem 3.4.1. (Asymptoticness) *The treatment effect estimator $\hat{\alpha}_{trt}$ is asymptotically normally distributed with:*

$$\sqrt{n}(\hat{\alpha}_{trt} - \alpha^*_{trt}) \to N(0, \sigma^2_{22})$$

where σ^2_{22} is the second diagonal element of the covariance matrix $\mathbf{\Sigma} = \mathbf{V}_2 + \mathbf{V}_2[\mathbf{C}\mathbf{V}_1\mathbf{C}^T - \mathbf{R}\mathbf{V}_1\mathbf{C}^T - \mathbf{C}\mathbf{V}_1\mathbf{R}^T]\mathbf{V}_2$ with $\boldsymbol{V}_1^{-1} = E\left\{\left(\frac{\partial l_1}{\partial \boldsymbol{\theta}_1}\right)\left(\frac{\partial l_1}{\partial \boldsymbol{\theta}_1^T}\right)\right\}$, $\boldsymbol{V}_2^{-1} = E\left\{\left(\frac{\partial l_2}{\partial \boldsymbol{\theta}_2}\right)\left(\frac{\partial l_2}{\partial \boldsymbol{\theta}_2^T}\right)\right\}$, $\boldsymbol{C} = E\left\{\left(\frac{\partial l_2}{\partial \boldsymbol{\theta}_2}\right)\left(\frac{\partial l_2}{\partial \boldsymbol{\theta}_1^T}\right)\right\}$, and $\boldsymbol{R} = E\left\{\left(\frac{\partial l_2}{\partial \boldsymbol{\theta}_2}\right)\left(\frac{\partial l_1}{\partial \boldsymbol{\theta}_1^T}\right)\right\}$. Under the strongly ignorable condition (3.7) and the linearity assumption in COROLLARY 4.3 of Rosenbaum and Rubin (1983), α^*_{trt} can be replaced by α_{trt}, the true marginal treatment effect.

The detailed proof is outlined in Appendix II. The above theorem provides us a basis for a modified variance estimator. To estimate the covariance matrix $\mathbf{\Sigma}$, we propose a sample estimate of $\mathbf{\Sigma}$ as follows:

$$\hat{\mathbf{\Sigma}} = \hat{\boldsymbol{V}}_2 + \hat{\boldsymbol{V}}_2[\hat{\boldsymbol{C}}\hat{\boldsymbol{V}}_1\hat{\boldsymbol{C}}^T - \hat{\boldsymbol{R}}\hat{\boldsymbol{V}}_1\hat{\boldsymbol{C}}^T - \hat{\boldsymbol{C}}\hat{\boldsymbol{V}}_1\hat{\boldsymbol{R}}^T]\hat{\boldsymbol{V}}_2$$

where $\hat{\boldsymbol{V}}_1^{-1} = \frac{1}{n}\sum_{i=1}^n \frac{\partial l_{1,i}}{\partial \boldsymbol{\theta}_1}\frac{\partial l_{1,i}}{\partial \boldsymbol{\theta}_1^T} \ |_{\hat{\boldsymbol{\theta}}_1}$, $\hat{\boldsymbol{V}}_2^{-1} = \frac{1}{n}\sum_{i=1}^n \frac{\partial l_{2,i}}{\partial \boldsymbol{\theta}_2}\frac{\partial l_{2,i}}{\partial \boldsymbol{\theta}_2^T} \ |_{\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2}$, $\hat{\boldsymbol{C}} = \frac{1}{n}\sum_{i=1}^n \frac{\partial l_{2,i}}{\partial \boldsymbol{\theta}_2}\frac{\partial l_{2,i}}{\partial \boldsymbol{\theta}_1^T} \ |_{\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2}$ and $\hat{\boldsymbol{R}} = \frac{1}{n}\sum_{i=1}^n \frac{\partial l_{2,i}}{\partial \boldsymbol{\theta}_2}\frac{\partial l_{1,i}}{\partial \boldsymbol{\theta}_1^T} \ |_{\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2}$, respectively. Matrix $\hat{\mathbf{\Sigma}}$ provides a good estimate to $\mathbf{\Sigma}$ as can be demonstrated by the simulation studies below.

3.5 Simulation Studies

We evaluate the finite sample performance of the proposed SNP-PS estimator by conducting extensive Monte Carlo simulation studies under various heterogeneity structures. The simulation studies are conducted according to the following treatment assignment mechanism:

$$logit(PS_{ij} = Pr(trt_{ij} = 1 \mid \mathbf{x}_{ij})) = \beta_1 x_{1,ij} + \beta_2 x_{2,ij} + \eta_i, \tag{3.7}$$

where $j = 1, \cdots, n_i$ indexes subjects nested in cluster i, $(\beta_1, \beta_2) = (-0.5, 0.15)$, $x_{1,ij} \sim Bern(0.5)$, and $x_{2,ij} \sim N(0,1)$. We let the cluster effect η to follow the following distribution:

$$\eta \sim f_K(\eta; \psi, \mu, \sigma^2) = f_1(\eta; -\frac{\pi}{3}, 0.3, 1.5).$$

This is a SNP density with Hermite expansion terms truncated at K=1 corresponding to a non-normal distribution (i.e. bio-mode distribution).

The responses are generated based on the following data generating process:

$$y_{ij} = \alpha_0 + \alpha_{trt} trt_{ij} + \alpha_{x_1} x_{1,ij} + \alpha_{x_2} x_{2,ij} + b\xi_i + \epsilon_{ij}, \tag{3.8}$$

where $(\alpha_0, \alpha_{trt}, \alpha_{x_1}, \alpha_{x_2}) = (0.3, 0.5, 1.5, 0.5)$. $\epsilon_{ij} \sim N(0, 1.0)$ is independent of trt_{ij}, $\mathbf{x}_{ij}$, and ξ_i. Distribution of ξ_i will be discussed in next paragraph. The overall cluster effect on the response is set as $b = 0$ and -0.3, respectively. For $b = 0$, it corresponds to the scenario of no cluster effect on the response.

To mimic the real world data, we also generate two additional covariates $\mathbf{x}_3 \sim Unif(-1, 1)$ and $\mathbf{x}_4 \sim Exp(1)$ which do not affect the treatment assignment and the response. We refer these covariates as nuisance variables. We include both the confounding covariates and nuisance variables in the SNP-PS model to estimate the propensity scores.

In addition to the above various configurations for the parameter values, we also investigate the performance of the proposed estimator $\hat{\alpha}_{trt}$ under different cluster effects settings for ξ in the response with respect to the cluster effects η in the treatment assignment process including no cluster effects, i.e. $b = 0$ (Case 1), $\xi = \eta$ (Case 2), and $\xi = 0.65 * \eta + 0.35 * N(0, 1)$ (Case 3).

Under each setting, we compare the proposed SNP-PS regression method with three other methods. The first method is the unadjusted model where the treatment effect estimate is obtained by comparing the differences between the two treatment groups without adjusting any other confounding factors. The other two methods are PS regression based methods where the propensity scores are estimated using the following logistic regression models:

$$logit(PS_{ij} = Pr(trt_{ij} = 1 \mid \mathbf{x}_{ij})) = \gamma_0 + \gamma_1 x_{1,ij} + \gamma_2 x_{2,ij} + \gamma_3 x_{3,ij} + \gamma_4 x_{4,ij}, \tag{3.9}$$

where the cluster effect is completely ignored in the treatment assignment process and we refer this method as Naive-PS method.

Instead of ignoring the cluster effect, a normal cluster effect is taken into consideration in the following mixed effects logistic regression model:

$$logit(PS_{ij} = Pr(trt_{ij} = 1 \mid \mathbf{x}_{ij})) = \gamma_0 + \gamma_1 x_{1,ij} + \gamma_2 x_{2,ij} + \gamma_3 x_{3,ij} + \gamma_4 x_{4,ij} + \zeta_i, \tag{3.10}$$

where the mixed effects logistic regression model with a normally distributed cluster effects $\zeta_i \sim N(0, \sigma_\zeta^2)$ is used to estimate the PS. We refer this PS estimation method as Normal-PS method.

For all methods compared, the treatment effect estimate $\hat{\alpha}_{trt}$ is obtained from the PS regression model (3.6) by plugging in the estimated PS. We first start with sample size of 1000 and the cluster cell size fixed at 5. Results for this simulation setting are presented in Table 3.1. We then increase the sample size to 5000 and results for this setting are given in Table 3.2. Column "Mean($\hat{\alpha}_{trt}$)" represents the average of the estimated treatment effect, $\hat{\alpha}_{trt}$. Column "Monte Carlo $SD(\hat{\alpha}_{trt})$" represents the Monte Carlo standard deviation of $\hat{\alpha}_{trt}$ based on 500 simulations which can be viewed as the true error of the treatment effect estimate. Column "Average $\widehat{SE}_D(\hat{\alpha}_{trt})$" is the average of the default standard errors of $\hat{\alpha}_{trt}$ output from the second stage PS regression model (3.6). Column "Average $\widehat{SE}_P(\hat{\alpha}_{trt})$" is the average of the standard errors of $\hat{\alpha}_{trt}$

calculated based on our proposed variance estimator, i.e. $\widehat{SE}_P(\hat{\alpha}_{trt}) = \sqrt{\hat{\mathbf{\Sigma}}_{[2,2]}}$, with $\hat{\mathbf{\Sigma}}_{[2,2]}$ being the second diagonal element of covariance matrix $\hat{\mathbf{\Sigma}}$ given in Theorem 3.4.1.

Comparing Tables 3.1 and 3.2, we notice that the unadjusted model always produces severely biased treatment effect estimates in all simulation settings. Naive-PS regression approach which ignores the treatment assignment heterogeneity also produces severely biased treatment effect estimates when there exists the cluster effects in the response (Case 2 & 3). Naive-PS is only acceptable when there is no cluster effects in the response (Case 1). A further examining Tables 3.1 and 3.2 also show that the normal-PS estimator also exhibited consistent bias in estimating α_{trt}. The proposed SNP-PS estimator is the only estimator that unbiasedly estimates α_{trt}. This demonstrated the robustness of the proposed SNP-PS model in both the point estimate and variance estimate when data is not following a normal heterogeneity distribution.

Tables 3.1 and 3.2 also show that the default standard error $\widehat{SE}_D(\hat{\alpha}_{trt})$ output directly from the PS regression model is biased in all simulation settings no matter which propensity score regression scheme is used. In contrast, the proposed variance estimator $\widehat{SE}_P(\hat{\alpha}_{trt})$ performs very well in all situations considered regardless if normal or SNP-PS regression approach is used.

To investigate the efficiency of SNP-PS regression in the scenario of normal cluster effect in the treatment assignment process, we conduct another set of simulation by setting the cluster effect η in the treatment allocation model (3.7) to follow $N(-0.2, 1)$ and the setting of the cluster effect ξ in response model (3.8) for Case 3 is set as $\xi = 0.65 * \eta + 0.35 * N(0, 1)$. Results for this setting of sample size 5000 are presented in Table 3.3.

Similar observations can be obtained from Table 3.3: there exists a large difference between the treatment effect estimate based on the unadjusted model and the true effect size. Naive-PS regression model results in biased treatment effect estimates in the situations of clustered data. Normal-PS and SNP-PS regression models provide unbiased treatment effect estimates regardless whether there are the cluster effects in response or not. Again, the proposed variance estimator provides very accurate variance estimates in all simulation settings. Also, the SNP-PS regression estimator and the normal PS regression estimator have identical efficiency in all simulation settings. This is expected since SNP-PS model includes the normal-PS model as a special case. That is the proposed SNP-PS method won't lose efficiency when the true model is normal. The price to pay is the extra computational workload.

In summary, the proposed SNP-PS regression model provides unbiased point estimate for α_{trt} regardless if there is cluster effect in response, or if heterogeneity in cluster effect is normally distributed. Furthermore, the proposed variance estimator provides accurate variance estimate for the variance of $\hat{\alpha}_{trt}$ under the SNP-PS regression framework in all simulation settings. In contrast, the default variance estimator provides biased variance estimate.

3.6 Real Data Analysis

To demonstrate the practical use of the proposed SNP-PS regression model, we applied it to a multi-center breast cancer study conducted by German Breast Cancer Study Group (Rauschecker et al. 1995). The study originally was intended as a randomized trial but had to be changed to an observational study due to the low randomization rate. The primary objective of this study was to compare the simple mastectomy ($trt = 0$) versus lumpectomy (BC, breast conservation, $trt = 1$) on the effect of quality of life (QoL) for 1036 breast cancer patients. The primary outcome was the performance status 9 months after surgery, which is scored between 0 and 100 based on the 25 QoL questionnaire responses, with higher scores reflecting better QoL. Covariates other than the therapies (i.e. mastectomy vs. lumpectomy) including patient age (ranges from 23 to 82) and tumor size (1mm $\sim$ 22mm) were considered as the potential confounding factors.

Our analysis was based on a sub-data set of this study from "nonrandom" R package with 646 patients and 63 clinics. We categorized age as young (age: ≤ 55) and old (age: > 55) and tumor size as small (ts: $\leq 10mm$) and large (ts: $> 10mm$), respectively as done by Senn et al. (2007). Distribution of baseline characteristics for these two covariates among the two treatment groups and each stratum of age and tumor size combination are presented in Table 3.4.

First part of Table 3.4 suggests that both age and tumor size are somewhat imbalanced between the two treatment groups. The mean age for the two groups are 59.4 yr and 52.0 yr with the corresponding standard deviation 10.4 and 11.5, respectively. The mean tumor size for the two groups are 14.5 mm and $13.5mm$ with standard deviation of 3.6 and 4.4. Taking a closer inspection of the second part of Table 3.4 reveals that older patients with larger tumor size favor mastectomy procedure while younger patients with larger tumor size prefer lumpectomy procedure. This suggests that the interaction of age and tumor size plays a role on the treatment assignment process. Furthermore, our likelihood ratio test showed that the cluster effect due to the clinic is significant ($p < 10^{-10}$) in determining the treatment assignment. Based on it, we include *age*, *ts*, and their interactions as the observed confounding covariates in the corresponding logistic regression models considered. Analysis results based on different analysis schemes are summarized in Table 3.5.

An evident observation from Table 3.5 is that the treatment effect estimate based on the unadjusted model is much larger than that based on all other methods compared. However, after adjusting the confounding factors via the propensity score regression method based on different versions of propensity score models, the treatment effect estimates are appreciably reduced. Different estimation schemes give us quite different but all positive treatment effect estimates. The positive number indicates that the QoL for the breast conservation patients is on average better than that of mastectomy patients. However, this conclusion is not statistically significant. This can be confirmed by checking the corresponding 95% confidence interval. Per our model selection criteria, SNP-PS regression model with Hermite expansion term $K = 1$ best fit this multicenter observational dataset and results based on it should be used.

Overall, based on the data collected for this multi-center study, there exists no statistically significant difference on the QoLs of breast cancer patients between the two treatment procedures after adjusting the patient age, tumor size, their interaction and the cluster effect.

3.7 Discussions

In this chapter, we proposed a semi-nonparametric propensity score model to deal with the treatment allocation heterogeneity that is commonly observed in real world medical data. The proposed SNP-PS model is robust to the parametric assumption for the heterogeneity distribution without making any specific parametric distribution assumption. Instead, a truncated Hermite polynomial along with the normal density is used to approximate the unspecified heterogeneity density. The Hermite expansion term can be determined via the frequently used model selection criteria, i.e. AIC or BIC etc. Numerically, we propose an adaptive quadrature integration algorithm to assess the non-closed form log-likelihood function for the proposed SNP-PS parameter estimates to avoid the large Monte Carlo errors of existing sampling based methods and reduce the computational workload. Furthermore, the biased variance estimate for the commonly used PS regression model is identified and corrected by our proposed robust variance estimator. Our proposed variance estimator for the SNP-PS regression model is critical in practice, as it will allow us to conduct valid statistical inference and lead to correct scientific conclusions that could be erroneous if the default variance estimator was used.

In addition to the propensity score estimation method described in equation (3.5), we have also studied another estimation method, the posterior mode method. In this method, $\widehat{PS}_{ij} = \frac{exp(\mathbf{x}_{ij}\hat{\boldsymbol{\beta}}+\hat{\eta}_{Mi})}{1+exp(\mathbf{x}_{ij}\hat{\boldsymbol{\beta}}+\hat{\eta}_{Mi})}$ with $\hat{\eta}_{Mi} = \arg\max_{\eta_i} l_i(\eta_i \mid \mathbf{trt}_i, \mathbf{x}_i; \hat{\boldsymbol{\beta}})$. This prediction method is the default one in the generalized linear mixed model packages in both SAS and R. However, this prediction does not perform nearly as good as the empirical Bayesian based prediction method (3.5).

The proposed SNP-PS for treatment allocation heterogeneity and their performances are all based on continuous responses. In real applications, there exists other type of responses commonly observed in CER studies, such as the binary (e.g. cure vs non-cure) events, and time to event (e.g. cure or death) data. Extending the proposed SNP-PS model to these data types under heterogeneous treatment assignment deserves further research and investigations. For example, longitudinal data where patients are followed up over years of medical interventions frequently exist in real world medical data. Such longitudinal resources would pose additional challenges in dealing with the time-dependent confounding.

Table 3.1: Non-Normal Cluster Effects in Treatment Allocation (Sample Size 1000)

Method	Average $\hat{\alpha}_{trt}$	Monte Carlo $SD(\hat{\alpha}_{trt})$	Average $\widehat{SE}_D(\hat{\alpha}_{trt})$	Average $\widehat{SE}_P(\hat{\alpha}_{trt})$
Case 1: $\xi = 0$				
Unadjusted	0.426	0.095	0.093	
Naive-PS	0.498	0.073	0.087	0.073
Normal-PS	0.539	0.107	0.127	0.100
SNP-PS	0.524	0.098	0.124	0.096
Case 2: $\xi = \eta$				
Unadjusted	-0.165	0.108	0.096	
Naive-PS	-0.098	0.093	0.092	0.078
Normal-PS	0.536	0.106	0.128	0.105
SNP-PS	0.503	0.098	0.125	0.102
Case 3: $\xi = 0.65 * \eta + 0.35 * N(0, 1)$				
Unadjusted	0.041	0.108	0.096	
Naive-PS	0.110	0.093	0.092	0.078
Normal-PS	0.537	0.106	0.131	0.107
SNP-PS	0.509	0.098	0.127	0.103

Note: Results are based on 500 simulations

Table 3.2: Non-Normal Cluster Effects in Treatment Allocation (Sample Size 5000)

Method	Average $\hat{\alpha}_{trt}$	Monte Carlo $SD(\hat{\alpha}_{trt})$	Average $\widehat{SE}_D(\hat{\alpha}_{trt})$	Average $\widehat{SE}_P(\hat{\alpha}_{trt})$
Case 1: $\xi = 0$				
Unadjusted	0.426	0.041	0.041	
Naive-PS	0.502	0.030	0.040	0.031
Normal-PS	0.546	0.048	0.057	0.044
SNP-PS	0.529	0.043	0.055	0.042
Case 2: $\xi = \eta$				
Unadjusted	-0.168	0.050	0.043	
Naive-PS	-0.097	0.041	0.041	0.033
Normal-PS	0.545	0.047	0.057	0.046
SNP-PS	0.508	0.043	0.055	0.044
Case 3: $\xi = 0.65 * \eta + 0.35 * N(0,1)$				
Unadjusted	0.040	0.047	0.043	
Naive-PS	0.113	0.038	0.042	0.033
Normal-PS	0.546	0.047	0.058	0.047
SNP-PS	0.515	0.044	0.057	0.045

Note: Results are based on 500 simulations

Table 3.3: Normal Cluster Effects in Treatment Allocation (Sample Size 5000)

Method	Average $\hat{\alpha}_{trt}$	Monte Carlo $SD(\hat{\alpha}_{trt})$	Average $\widehat{SE}_D(\hat{\alpha}_{trt})$	Average $\widehat{SE}_P(\hat{\alpha}_{trt})$
Case 1: $\xi = 0$				
Unadjusted	0.424	0.037	0.040	
Naive-PS	0.501	0.030	0.039	0.030
Normal-PS	0.505	0.037	0.049	0.037
SNP-PS	0.505	0.037	0.049	0.037
Case 2: $\xi = \eta$				
Unadjusted	-0.108	0.069	0.042	
Naive-PS	-0.036	0.068	0.041	0.033
Normal-PS	0.502	0.037	0.049	0.039
SNP-PS	0.502	0.037	0.049	0.039
Case 3: $\xi = 0.65 * \eta + 0.35 * N(0,1)$				
Unadjusted	0.078	0.067	0.042	
Naive-PS	0.152	0.065	0.041	0.033
Normal-PS	0.503	0.037	0.050	0.040
SNP-PS	0.503	0.037	0.050	0.040

Note: Results are based on 500 simulations

Table 3.4: Baseline Distribution for German Breast Cancer Study Data

	Mastectomy			Lumpectomy		
	N	Mean	SD	N	Mean	SD
Age (year)	167	59.4	10.4	479	52.0	11.5
Tumor Size (mm)	167	14.5	3.6	479	13.5	4.4

		Mastectomy		Lumpectomy	
		N	Proportion	N	Proportion
Young ($\leq$ 55 yr)	Small ($\leq$ 10 mm)	7	0.042	88	0.184
Young ($\leq$ 55 yr)	Large ($>$ 10 mm)	49	0.293	206	0.430
Old ($>$ 55 yr)	Small ($\leq$ 10 mm)	23	0.138	42	0.088
Old ($>$ 55 yr)	Large ($>$ 10 mm)	88	0.527	143	0.298
Total		167	1.000	479	1.000

Table 3.5: Analysis of German Breast Cancer Study Data with Cluster Effects

Method	$\hat{\alpha}_{trt}$	$SE(\hat{\alpha}_{trt})$	95% CI
Unadjusted	1.589	1.261	(-0.883,4.061)
Naive-PS	0.793	1.248	(-1.653,3.239)
Normal-PS	0.560	1.645	(-2.664,3.784)
SNP-PS	0.715	1.636	(-2.492,3.922)

Note: model selection criteria indicates $K = 1$ for SNP-PS best fit this dataset

Chapter 4

A Flexible Mixed Effects PS Model for Clustered Data

4.1 Introduction

With the ever readily availability of large clinical datasets, especially the electronic medical record data, various efforts have been made to look into those datasets in comparing the effectiveness of different treatments. Real world clinical data offer a broader population spectrum as well as longer time-intervals than a typical RCT data (e.g. Benson and Hartz 2000). They tend to reflect daily clinical practice more closely and provide more clinically relevant information than RCTs (e.g. Yang et al. 2010). The cost of collecting observational clinical data for conducting comparative effectiveness research (CER) is often much lower than that of conducting RCTs. Proper use of medical record data for the comparative effectiveness research provides investigators an effective way to compare various interventions for the inferiority, equivalence, or superiority (e.g. Mitka 2010).

As a kind of observational data, a key feature of medical record data is that the treatment allocation is not random. Rather, the treatment assignment could be influenced by many factors related to patients, physicians, and health care systems etc. This non-randomness in treatment assignment could create imbalanced baseline covariates, i.e. confounding variables, and result in severe biased treatment effect estimate if the confounding factors are not appropriately adjusted. Several methods have been proposed to address the confounding problems in observational studies, such as matching (e.g. Miettinen 1968), stratification (e.g. Cochran 1968), the instrumental variable approach (e.g. Angrist et al. 1996) and various versions of propensity score methods (e.g. Rosenbaum and Rubin 1983). Among them, the propensity score (PS) is one of the most commonly used in practice.

Apart from its distinctive observational feature, another key feature of medical record data is that treatment assignments are clustered. For example, the medical record data are highly clustered by physicians, clinics, hospitals, and insurance agencies. Real world medical data come from various sources that include insurance claim data, hospital, and clinic prescription records, etc. Generally, a treatment that a patient eventually receives depends on many factors which include but are not limited to: (1) physician's factors which include physician's professional trainings, practice styles etc; (2) patient's factors which include patient's age, gender, social economic status etc; (3) system's factors which include insurance policies, hospital policies , etc. All these factors can create potential biases toward certain type of treatments. The clustering feature of medical record data may reflect sample heterogeneity due to these factors that can influence both the treatment allocation process as well as the disease outcomes.

Practical medical record data is not only clustered but also multilevel or hierarchically clustered. For example, patients are clustered within physicians who are also further clustered within hospitals or clinics. As shown in the previous chapter, for medical record data, both the observational and clustering features should be taken into consideration for unbiased treatment effect estimates. In Chapter 3, we have developed a robust propensity score model, i.e. SNP-PS, for single layer clustered data. Though the proposed method is computationally more efficient than the existing Monte Carlo sampling based semi-nonparametric approaches (e.g. Chen et al. 2002), it is not scaled to handle large clustered medical record data. Typical medical record data contains medical records for a large number of patients, ranging from personal information like age, demographics, social status to medical information such as blood pressure, family disease history, medical history, medication and allergies, to immunization status for each patient. Digitization of these records promotes the integration of hundreds of thousands or even millions of patients into a large database from different resources (e.g. Tinetti and Studenski 2011; Tannen et al. 2008; Weiner et al. 2008) such as insurance claim data, hospital records, prescription records, and observational studies, etc. It would be very computationally challenging to apply the method developed in Chapter 3 to large clustered data directly. The computational challenge arises from the facts that 1) we model the cluster effect in the logistic PS model where the MLEs have no analytic closed form; and 2) the random effect in the logistic PS model is non-parametrically modeled, making maximization on the likelihood even more complicated. Developing computationally efficient statistical methods to properly handle the heterogeneity structure of medical record data is more important than ever for making valid statistical inferences.

To ease the computational workload, in this chapter, we propose to not model the heterogeneity structure in the first stage logistic PS model and instead try to resolve the heterogeneity problem in the second stage PS regression model. Specifically, we propose two PS regression approaches, one based on the multiple regression model and the other on the mixed effects model. The first one is a flexible multilevel propensity score (FM-PS) regression approach based on mixed effects models. In traditional mixed effects models, there is a critical independence assumption between fixed effects covariates and random effects. However, this assumption is unrealistic and rarely holds for real medical record data. For example, when patients make decisions where and which clinic to visit, the physicians' specialties, and insurance coverages play important roles. These factors don't act independently but jointly. They affect the treatment allocation process and disease outcomes as well. Furthermore, the independence assumption is not easy to check since the cluster effects are unobserved. If the assumption does not hold, the estimates for the fixed effects parameters could be biased (Verbeke and Lesaffre 1996) and the degree of biasness depends on the degree of the correlation between the fixed effect covariates and the cluster effects. We proposed FM-PS models to relax the independence assumption between the treatment assignment and the heterogeneity cluster effects for large clustered datasets. Under this modeling scheme, the cluster effect is not necessarily independent of the fixed effects covariates or the cluster effects from other clustering levels.

The remaining of this chapter is organized as follows. In Section 2, we describe the details of FM-PS model. The theoretical properties of the treatment effect estimator based on FM-PS are presented in this section also. A cluster bootstrapping procedure is described in Section 3 for variance estimation of treatment effect estimate. We provide a statistics to test the hypothesis of independence of the cluster effect terms (with respect to other fixed effects terms) in Section 4 to select the optimal model for treatment effect estimate. Extensive simulation studies and results are presented in Section 5. The robustness property of FM-PS for dealing with model mis-specifications and the capability/limitation to handle unobserved confounding covariates are also investigated and presented in Section 5. The practical use of FM-PS is demonstrated via a real multi-center observational dataset in Section 6. This chapter ends with discussions and future research for FM-PS in Section 7.

4.2 Proposed Methods

4.2.1 Data Types

Before presenting the details of our proposed models, we first introduce three types of clustered data (i.e. single level, multilevel, and hierarchical) that we consider in this chapter.

Single level clustered data (**Type 1**): for the jth sample nested in the cluster unit $i(j = 1, \cdots, n_i; i = 1, \cdots, n)$, we define its outcome as y_{ij}, treatment assignment trt_{ij}, and all other observed covariates as $\mathbf{x}_{ij}$.

Multilevel clustered data (**Type 2**): for the kth sample in the ith level of cluster 1 (e.g. physicians) and the jth level of cluster 2 (e.g. insurance policies) $(k = 1, \cdots, n_{ij}; i = 1, \cdots, n_i; j = 1, \cdots, n_j)$, we denote outcome as y_{ijk}, treatment assignment trt_{ijk} and observed covariates $\mathbf{x}_{ijk}$. Here the two clusters, i.e. 1 and 2, are not nested one in another and potentially overlap. For example, the same physician may see patients with all sorts of health insurances, and the same policy holders could visit different physicians.

Hierarchical clustered data (**Type 3**): for the kth sample in the ith level of cluster 1 (such as clinics/hospitals) and the jth level of cluster 2 (such as physicians) $(k = 1, \cdots, n_{ij}; j = 1, \cdots, n_i; i = 1, \cdots, n)$, we again have outcome y_{ijk}, treatment assignment trt_{ijk} and observed covariates $\mathbf{x}_{ijk}$. In contrast to Type 2 data, here units of cluster 2 (e.g. physicians) are completely nested in units of cluster 1 (e.g. clinics).

Even though we only consider clustered data with single level or two levels for ease of presentations, the proposed methods below can be straightforwardly extended to higher level clustered data.

PS Model: For all types of data, we first fit the following logistic regression model with only all the observed covariates included without any cluster effect terms

$$logit(Pr(trt = 1 \mid \mathbf{x})) = \beta_0 + \mathbf{x}\boldsymbol{\beta}_{\mathbf{x}}, \tag{4.1}$$

from which we get the estimated PS score given by the following:

$$\widehat{PS} = \frac{\exp(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}}_{\mathbf{x}})}{1 + \exp(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}}_{\mathbf{x}})}$$

where $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}_{\mathbf{x}}$ are the MLE from model (4.1). That is we don't deal with the cluster effects in (4.1) and instead we take care of the clustering heterogeneity of treatment assignment in the downstream PS regression analysis. Several random and fixed effects PS regression models can be used for the above three types of clustered data to obtain the treatment effect estimate as we describe in the following.

Standard Mixed Effects PS (SM-PS) Models: once PS are estimated from (4.1), it is natural to fit the following mixed effects PS regression models:

$$\begin{aligned}
\text{Type 1}: \quad y_{ij} &= \alpha_0 + \alpha_{trt} trt_{ij} + \alpha_{PS}\widehat{PS}_{ij} + \eta_i + \epsilon_{ij} && (4.2)\\
\text{Type 2}: \quad y_{ijk} &= \alpha_0 + \alpha_{trt} trt_{ijk} + \alpha_{PS}\widehat{PS}_{ijk} + \eta_i + \xi_j + \epsilon_{ijk} && (4.3)\\
\text{Type 3}: \quad y_{ijk} &= \alpha_0 + \alpha_{trt} trt_{ijk} + \alpha_{PS}\widehat{PS}_{ijk} + \eta_{ij} + \epsilon_{ijk} && (4.4)
\end{aligned}$$

and obtain the corresponding treatment effect estimate, $\hat{\alpha}_{trt}$. However, the above models have a strong independence assumption imposed on the random effects. That is, all cluster effect terms, i.e. $\eta_{i(j)}$ and ξ_j, are independent of the fixed effects covariates, i.e. $trt_{ij(k)}$ and $\widehat{PS}_{ij(k)}$. Specifically, the SM-PS models assume the following:

$$\begin{cases}
\text{Type 1: } \eta_i \mid (trt_{ij}, \widehat{PS}_{ij}) \sim N(0, \sigma_\eta^2) \\
\text{Type 2: } \eta_i \mid (trt_{ijk}, \widehat{PS}_{ijk}) \sim N(0, \sigma_\eta^2) \text{ and } \xi_j \mid (trt_{ijk}, \widehat{PS}_{ijk}) \sim N(0, \sigma_\xi^2) \\
\text{Type 3: } \eta_{ij} \mid (trt_{ijk}, \widehat{PS}_{ijk}) \sim N(0, \sigma_\eta^2)
\end{cases}$$

However, the above independence assumptions between the random cluster effects and other fixed effects covariates including the treatment assignment seldom hold for many real world medical record data due to various confounding factors and complexity of cluster heterogeneity structures in the treatment assignment process. If the independence assumptions are violated, treatment effect estimate could be severely biased. To relax the stringent independence assumptions made in the SM-PS models (4.2) $\sim$ (4.4), we propose a set of novel models as given in the following subsection. Specifically, we propose the following set of mixed effects propensity score regression models for the above three types of data under the above corresponding assumptions. However, it should be noted that the proposed method is not necessary restricted to the three types of clustering. It is straightforward to extend the method to other higher level clustering scenarios.

4.2.2 Flexible Mixed Effects PS (FM-PS) Models

For each of the three types of clustered data, we impose a dependence condition between the random effects and fixed effects covariates:

Type 1: $\eta_i \mid (trt_{ij}, \widehat{PS}_{ij}) \sim N(\alpha_{\overline{trt}}\overline{trt}_i + \alpha_{\overline{PS}}\overline{PS}_i, \sigma_\eta^2)$

Type 2:

$$\begin{cases} \eta_i \mid (trt_{ijk}, \widehat{PS}_{ijk}) \sim N(\alpha_{(\overline{trt},1)}\overline{trt}_i + \alpha_{(\overline{PS},1)}\overline{PS}_i, \sigma_\eta^2) \\ \xi_j \mid (trt_{ijk}, \widehat{PS}_{ijk}) \sim N(\alpha_{(\overline{trt},2)}\overline{trt}_j + \alpha_{(\overline{PS},2)}\overline{PS}_j, \sigma_\xi^2) \end{cases}$$

Type 3: $\eta_{ij} \mid (trt_{ijk}, \widehat{PS}_{ijk}) \sim N(\alpha_{(\overline{trt},2)}\overline{trt}_{ij} + \alpha_{(\overline{PS},2)}\overline{PS}_{ij}, \sigma_\eta^2)$

where the $\overline{trt}$s and $\overline{PS}$s are the proportion of subjects receiving one given treatment and the average propensity scores within each cluster/sub-cluster unit indicated by the subscript, respectively. These conditions relieve the unrealistic independence assumptions and lead us to the following three flexible mixed effects PS (FM-PS) models, one for each data type.

$$
\begin{aligned}
\text{Type 1}: \quad y_{ij} &= \alpha_0 + \alpha_{trt} trt_{ij} + \alpha_{PS}\widehat{PS}_{ij} + \alpha_{\overline{trt}}\overline{trt}_i + \alpha_{\overline{PS}}\overline{PS}_i + \eta_i + \epsilon_{ij} && (4.5)\\
\text{Type 2}: \quad y_{ijk} &= \alpha_0 + \alpha_{trt} trt_{ijk} + \alpha_{PS}\widehat{PS}_{ijk} + \alpha_{(\overline{trt},1)}\overline{trt}_i + \alpha_{(\overline{PS},1)}\overline{PS}_i + \alpha_{(\overline{trt},2)}\overline{trt}_j \\
&\quad + \alpha_{(\overline{PS},2)}\overline{PS}_j + \eta_i + \xi_j + \epsilon_{ijk} && (4.6)\\
\text{Type 3}: \quad y_{ijk} &= \alpha_0 + \alpha_{trt} trt_{ijk} + \alpha_{PS}\widehat{PS}_{ijk} + \alpha_{\overline{trt}}\overline{trt}_{ij} + \alpha_{\overline{PS}}\overline{PS}_{ij} + \eta_{ij} + \epsilon_{ijk} && (4.7)
\end{aligned}
$$

where $\eta_{i(j)} \sim N(0, \sigma_\eta^2)$, $\xi_j \sim N(0, \sigma_\xi^2)$, and $\epsilon \sim N(0, \sigma^2)$, respectively. We refer the new procedures as FM-PS which stands for *F*lexible *M*ixed effects *P*ropensity *S*core approaches. The new cluster effect terms η and ξ now follow the regular assumptions in the standard mixed effects models (i.e. normality and independence, e.g. Laird and Ware 1982; Schall 1991; Zeger and Karim 1991; Breslow and Clayton 1993).

The proposed approach is computationally efficient as will be shown by our simulations. It can handle large medical record datasets with almost no constrains if memory is of no concern. Modeling the single layer clustered or multilevel clustered data via mixed effects model is not new (Diggle et al. 2002; Fitzmaurice et al. 2004). In practice, it is also common to use fixed effect models alternatively to analyze such kinds of data with cluster effects modeled with dummy variables (Suits 1957). Alternatively, we propose the following fixed effect models using dummy variables for the three types of clustered data.

Fixed Effect PS (FE-PS) Models:

$$\text{Type 1:} \quad y_{ij} = \alpha_0 + \alpha_{trt} trt_{ij} + \alpha_{PS}\widehat{PS}_{ij} + \mathbf{DV}_i \boldsymbol{\alpha}_{DV} + \epsilon_{ij} \qquad (4.8)$$

$$\text{Type 2:} \quad y_{ijk} = \alpha_0 + \alpha_{trt} trt_{ijk} + \alpha_{PS}\widehat{PS}_{ijk} + \mathbf{DV}_i \boldsymbol{\alpha}_{DV_1} + \mathbf{DV}_j \boldsymbol{\alpha}_{DV_2} + \epsilon_{ijk} \qquad (4.9)$$

$$\text{Type 3:} \quad y_{ijk} = \alpha_0 + \alpha_{trt} trt_{ijk} + \alpha_{PS}\widehat{PS}_{ijk} + \mathbf{DV}_{ij} \boldsymbol{\alpha}_{DV} + \epsilon_{ijk} \qquad (4.10)$$

where the $\mathbf{DV}_i$s and $\mathbf{DV}_j$ are dummy variables created to represent units in cluster

1 and cluster 2, respectively. In (4.10), for hierarchical clustered data, the dummy variables $\mathbf{DV}_{ij}$ are created for units in cluster 2, which are nested in cluster 1.

The relationship between fixed effect models with dummy variables and mixed effects models have been well studied and the equivalence between the fixed effect estimates from the two model strategies have been established by Mundlak (1978). The following theorem summarizes the equivalence between the treatment effect estimates from the FE-PS and FM-PS models. The proof can be found in Mundlak (1978).

Theorem 4.2.1. (Equivalence) *The treatment effect estimate, denoted as $\hat{\alpha}_{trt}$, for parameter α_{trt} in FM-PS models (4.5) and (4.7) are identical to the OLS estimate of α_{trt} in models (4.8) and (4.10), respectively.*

Though the above theorem has established the equivalence of the two PS procedures, in practice, fitting the FE-PS models with ordinary linear regression using standard software packages, e.g. R or SAS, might be computationally prohibited for large medical data. However, FM-PS models have no such issues. Furthermore, the mixed effects models are more flexible in extending the conclusions to patients from, for example clinics and physicians not in the data analyzed. In contrast, conclusions based on fixed effect models are restricted to the clinics and physicians included in the original data analysis.

4.3 Variance Estimate of $\hat{\alpha}_{trt}$

As noted in the previous chapters, PS is an estimated quantity instead of an observed covariate. Ignoring this fact will result in biased variance estimation for $\hat{\alpha}_{trt}$. In chapters 2 and 3, we proposed a two-stage variance estimator for PS regression models. The validity of the two-stage variance estimator is based on the consistency of parameters estimated from the first stage PS models. In the proposed FM-PS models, the first stage PS models ignore the cluster effects and thus the parameter estimates may not be consistent and the two-stage variance estimators may not be applicable for the FM-PS models. In this section, we propose a cluster bootstrapping procedure to estimate the variance of $\hat{\alpha}_{trt}$ for the proposed FM-PS models as we describe in the following.

The bootstrap technique by Efron (1979) has been often used in practice for inferencing on a population distribution based on the sample data (sample → population) by resampling the sample data and performing inference on the resampled data (resample → sample). However, for the clustered data, the standard bootstrap resampling or residual resampling procedure will not work (e.g. Liu and Chen 1998; Whitley 1994), since it is not able to replicate the correlation structure in the data. Alternatively, cluster bootstrapping (e.g. Davison and Hinkley 1997; McCullagh 2000; Andersson and Karlsson 2001; Ukoumunne et al. 2003; Carpenter et al. 2003; Butar and Lahiri 2003; Monaco et al. 2005; Field and Welsh 2007) procedures have been proposed which we adopt here for our purpose.

The idea of cluster bootstrapping is to treat each cluster unit as an independent sample and randomly draw the n cluster units with replacement. All subjects within a selected cluster unit will be included as bootstrapped samples and used for the bootstrapping analysis. The cluster bootstrapping replicates the correlation structure by resampling based on the "basic" cluster units. For single level clustered data, the "basic" sampling units are the original cluster units. Similarly, for hierarchical clustered data, the sampling units are the cluster units in cluster 2, i.e. the lowest (finest) level of clustering units (e.g. physicians in the clinics versus physicians hierarchical clustering relationship). For multilevel clustered data, we construct new cluster units as the "basic" bootstrapping units. Specifically, each new cluster unit, ij, is the joint sub-unit of ith level in cluster 1 and jth level in cluster 2. Cluster bootstrapping for multilevel FM-PS model (4.6) is based on these reconstructed new cluster units. Simulations will be used to investigate the performance of the proposed cluster bootstrapping approach.

4.4 Testing of Cluster Effects

If the cluster effect is independent of the fixed effect covariates, the SM-PS models are expected to be more efficient than the FM-PS models. We propose the following test statistic for comparing the SM-PS and FM-PS models. For the ease of illustration, we use $\boldsymbol{\zeta}$ to represent all the cluster effects (e.g. $\boldsymbol{\zeta} = (\{\eta_i\}, \{\xi_j\})$ in the outcome for the multilevel cluster data) in each of the three FM-PS models. We refer the log-likelihood from each of the FM-PS models as l_{FM-PS}. Similarly, we refer the log-likelihood from each of the SM-PS models as l_{SM-PS}. Then, testing for independence is equivalent to test

$$H_0 : \boldsymbol{\zeta} \perp trt \; H_a : \boldsymbol{\zeta} \not\perp trt \tag{4.11}$$

To test the hypothesis (4.11), we introduce the following likelihood ratio test statistic:

$$\begin{aligned} LRT_{indep} &= -2\log\frac{\text{Likelihood of SM-PS model}}{\text{Likelihood of FM-PS model}} \\ &= 2(l_{FM-PS} - l_{SM-PS}) \qquad (4.12) \\ &\overset{\mathrm{H_0}}{\sim} \chi^2(df) \end{aligned}$$

where df, the degree of freedoms of the χ^2 distribution, equals the difference between the numbers of fixed covariates in the FM-PS and SM-PS models. More specifically, $df = 2, 4,$ and 2, respectively for the single level, multilevel, and hierarchical cluster data. Statistics LRT_{indep} provides a tool for us to select between the SM-PS and FM-PS models.

In summary, the following are the procedures proposed to estimate treatment effects for the clustered observational data:

Step 1: Fit the data with both SM-PS and FM-PS models.

Step 2: Use the test statistics LRT_{indep} to select between SM-PS and FM-PS models.

Step 3: Based on the selected model, estimate α_{trt}.

Step 4: Conduct the cluster bootstrapping procedure to obtain a valid variance estimate of $\hat{\alpha}_{trt}$.

4.5 Monte Carlo Simulations

To evaluate the finite sample performance of the proposed FM-PS models and the cluster bootstrapping procedure, we conduct intensive simulation studies under various settings with varying sample size and heterogeneity correlation structures. We start the simulation studies with the following settings for the three types of clustered data:

Treatment allocation model: In our first set of simulations, the treatment assignments are generated for single, multilevel, and hierarchical clustered data through the following mechanisms

$$\text{Type 1: } trt_{ij} = \begin{cases} 1 & \text{if } 0.2 + \mathbf{x}_{(1,ij)}\beta_{\mathbf{x}} + w_i + \epsilon_{ij} > 0 \\ 0 & \text{if } 0.2 + \mathbf{x}_{(1,ij)}\beta_{\mathbf{x}} + w_i + \epsilon_{ij} \leq 0 \end{cases}$$

$$\text{Type 2: } trt_{ijk} = \begin{cases} 1 & \text{if } 0.2 + \mathbf{x}_{(1,ijk)}\beta_{\mathbf{x}} + w_i + \epsilon_{ijk} > 0 \\ 0 & \text{if } 0.2 + \mathbf{x}_{(1,ijk)}\beta_{\mathbf{x}} + w_i + \epsilon_{ijk} \leq 0 \end{cases}$$

$$\text{Type 3: } trt_{ijk} = \begin{cases} 1 & \text{if } 0.2 + \mathbf{x}_{(1,ijk)}\beta_{\mathbf{x}} + w_{ij} + \epsilon_{ijk} > 0 \\ 0 & \text{if } 0.2 + \mathbf{x}_{(1,ijk)}\beta_{\mathbf{x}} + w_{ij} + \epsilon_{ijk} \leq 0 \end{cases}$$

respectively, where confounding covariate vectors $\mathbf{x}_{(1,ij(k))} = (\mathbf{x}_{(1,ij(k),1)}, \cdots, \mathbf{x}_{(1,ij(k),8)})'$ consists 8 covariates. Pairs of confounding covariates are generated from binary, normal, exponential, and uniform distributions as $Bern(0.5)$, $N(0,1)$, $Exp(1)$, and $Unif(-1,1)$, respectively. Parameter $\beta_{x,j}$s in $\boldsymbol{\beta}_x = (\beta_{x,1}, \cdots, \beta_{x,8})'$ are randomly generated from $Unif(-1,1)$ first and then fixed for all the subsequent simulations. The random measurement error $\epsilon \sim N(0,1)$. The distribution of cluster effects $w_{i(j)}$ will be described later.

Data generating model: The responses are generated based on the following mechanisms for the three types of clustered data:

$$\text{Type 1:} \quad y_{ij} = 0.3 + \alpha_{trt} trt_{ij} + \mathbf{x}_{(1,ij)} \alpha_{\mathbf{x}} + \eta_i + \varepsilon_{ij} \tag{4.13}$$

$$\text{Type 2:} \quad y_{ijk} = 0.3 + \alpha_{trt} trt_{ijk} + \mathbf{x}_{(1,ijk)} \alpha_{\mathbf{x}} + \eta_i + \xi_j + \varepsilon_{ijk} \tag{4.14}$$

$$\text{Type 3:} \quad y_{ijk} = 0.3 + \alpha_{trt} trt_{ijk} + \mathbf{x}_{(1,ijk)} \alpha_{\mathbf{x}} + \eta_{ij} + \varepsilon_{ijk} \tag{4.15}$$

where the random measurement error $\varepsilon_{ij(k)} \sim N(0,1)$ which is independent of $trt_{ij(k)}$, $\mathbf{x}_{(1,ij(k)}$ and cluster effects $\eta_{i(j)}$ and ξ_j. The true treatment effect α_{trt} is fixed at 0.5. The effect of each confounding covariate, $\alpha_{x,j}$ $(j = 1, \cdots, 8)$, is also randomly generated from $Unif(-1, 1)$ and fixed for all subsequent simulations.

The cluster effects in the treatment assignment models and the response generating models are sampled from non-normal distributions, and are correlated or independent of each other. Specifically, we let the cluster effect w_i in the treatment assignments and the cluster effect η_i in outcome responses (4.13) and (4.14) be correlated as follows:

$$\eta_i \sim \begin{cases} w_i & \text{if } w_i \geq 0 \\ -Exp(1) & \text{if } w_i < 0 \end{cases}$$

where $w_i \sim \frac{1}{2}Exp(2) + \frac{1}{2}N(-0.5, 1)$.

The cluster effects w_{ij}s and η_{ij}s are simulated similarly as above where the subscript i is replaced by the subscript ij. For the case where $w_{i(j)}$s are independent of $\eta_{i(j)}$s, $\eta_{i(j)}$ are simulated from $\frac{1}{2}Exp(1) + \frac{1}{2}N(-1, 1)$ while $w_{i(j)}$ is generated the same as above. For multilevel clustering, we further simulate the cluster effects ξ_j from $\frac{1}{2}N(2, 1) +$

$\frac{1}{2}N(-2, 1)$.

To mimic the real world observational data, in addition to the true confounding covariates $\mathbf{x}_{(1,ij(k))}$, we also simulate an additional set of 8 nuisance variables, $\mathbf{x}_{(2,ij(k))}$. Again, pairs of the 8 nuisance variables are generated from $Bern(0.5)$, $N(0, 1)$, $Exp(1)$, and $Unif(-1, 1)$, respectively. These nuisance variables have no effects on either the response variable y or the treatment assignment trt. However, we include them in our analysis. That is we include all of the 16 covariates $\mathbf{x}_{ij(k)} = (\mathbf{x}'_{(1,ij(k))}, \mathbf{x}'_{(2,ij(k))})'$ in our analysis to obtain the estimation for PS.

To evaluate the performance of the proposed FM-PS models, we compare them with several other competing models, which include 1) unadjusted model where the treatment effect estimate equals the response difference between the two treatment groups without any confounding covariate adjustment; 2) non-random PS (NR-PS) model where the treatment effect estimate is obtained via the simple PS regression in which propensity scores are estimated by the PS model (4.1) without considering unobserved heterogeneity; 3) the SM-PS models; and 4) the FE-PS models.

For the single level clustering data, we simulate samples with size n of 1000 or 5000. The cluster cell sizes vary from $5, 10$, or 15 with total cluster units of 100 and 500 respectively for the two sample sizes. Simulation results for these two sample sizes are summarized in Table (4.1) and (4.2), respectively for correlated and independent cluster effects between treatment assignments and outcomes. For each simulation setup, the results are based on 500 simulations and the cluster bootstrapping is conducted with 500 resampling. Column "Independence" indicates if the cluster effect in treatment assignment and the cluster effect in outcome are independent or not. Column "Average($\hat{\alpha}_{trt}$)" represents the average treatment effect estimate, column "Monte Carlo SD($\hat{\alpha}_{trt}$)" shows the Monte Carlo standard deviation of the treatment effect estimate, column "Average $\widehat{SE}_D(\hat{\alpha}_{trt})$" presents the average standard error for the treatment effect estimate output directly from the corresponding model fitting, and column "Average $\widehat{SE}_B(\hat{\alpha}_{trt})$" displays the average standard error for the treatment effect estimate via the cluster bootstrapping procedure.

Tables (4.1) and (4.2) show that when cluster effect w_is in treatment assignment and that in response η_is are independent of each other, the treatment effect estimated by all methods except the unadjusted model are close to the true treatment effect size, indicating the effectiveness of all the PS based methods in adjusting the confounding factors. However, a closer examining the second part of both tables reveals that when w_is and η_is are correlated, only the FE-PS model and the proposed FM-PS model can provide unbiased treatment effect estimates. Furthermore, the estimates from the two models are identical which is consistent with the conclusion of Theorem 4.2.1. In contrast, the treatment effect estimates from the NR-PS model and the SM-PS model are biased.

Comparing the columns Monte Carlo $SD(\hat{\alpha}_{trt})$ and $\widehat{SE}_D(\hat{\alpha}_{trt})$, we conclude that for the proposed FM-PS model, the default variance estimates are biased. For example, when sample size $n = 1000$, the Monte Carlo standard deviation from the FM-PS model is 0.117 while the default standard error is 0.159 where the later number is more than 30% larger than the former one. This observation holds for $n = 5000$ (i.e. 0.054 vs 0.070). In contrast, numbers in column Monte Carlo $SD(\hat{\alpha}_{trt})$ and column $\widehat{SE}_B(\hat{\alpha}_{trt})$ are close to each other for all models and sample sizes, demonstrating the superior performance of the cluster bootstrapping procedure.

For multilevel clustering data, we fix cluster cell size of cluster 1 as 5 with total 1000 cluster units while we let the cluster cell size for cluster 2 varies from 100 to 900 among total 10 cluster units. For hierarchical clustering data, again we set cluster cell size for cluster 1 to 5 with total 1000 cluster units which are nested within 500 units in cluster 2 with cell size 10. Simulation results for multilevel and hierarchical clustering scenarios with sample size $n = 5000$ are presented in Tables (4.3) and (4.4) respectively.

Table (4.3) and (4.4) demonstrate that when w_i (or w_{ij}) and η_i (or η_{ij}) are independent of each other for multilevel and hierarchical clustered data, we again can obtain unbiased treatment effect estimates from all the four PS procedures, though the unadjusted model fails. For data with correlated cluster effects, only estimates from the FE-PS and FM-PS models are unbiased. Similarly, the default variance estimates are biased which can be drastically improved by the cluster bootstrapping procedure.

All four tables consistently demonstrate that when the cluster effects in treatment assignments are independent of the cluster effects in the responses, the treatment effect estimate from SM-PS model is more efficient than the proposed FM-PS (or FE-PS) models. Thus, the SM-PS model is preferred for data where the cluster effects are independent of the fixed effect covariates. The proposed likelihood ratio test statistic, i.e. LRT_{indep}, serves the purpose to select the optimum model by testing the independence assumption. To investigate the performance of LRT_{indep}, we generate several Q-Q plots of the empirical LRT_{indep} against the theoretical χ^2 distribution for the data with independent cluster effects for the three types of clustered data with sample size $n = 5000$ in Figures (4.1) $\sim$ (4.3).

All Q-Q plots clearly indicate that the empirical test statistics follow the asymptotic χ^2 distribution well. For all data simulated with correlated cluster effects, we also plot the observed LRT_{indep} across the 500 simulations. In addition, we add the 500 LRT_{indep} observed values for the data with independent cluster effects. Clearly, the empirical LRT_{indep} from the data under the null (i.e. independent situations) and under the alternative (i.e. correlated situations) are well separated, indicating that for the simulated correlated data, the proposed LRT_{indep} has very good power to detect the correlation.

In summary, the proposed FM-PS models (or equivalent FE-PS models) provide unbiased treatment effect estimates no matter if the cluster effects in treatment assignments are independent of the cluster effects in the responses or not. Traditional SM-PS models provide more efficient and unbiased treatment effect estimates than the proposed models when the cluster effects in treatment allocations are independent of the cluster effects in the responses. However, when the cluster effects are correlated, the treatment effect estimates from SM-PS models are biased. The empirical cluster bootstrapping procedure provides good variance estimation to $\hat{\alpha}_{trt}$.

The above simulations have shown that the proposed FM-PS models are robust to the departure of cluster effects η and ξ from normality distributions, because in all our simulations, the cluster effects are non-normally distributed. Next, we further investigate the robustness property of the FM-PS models with other model misspecification scenarios under the single level clustering setting. The conclusions hold for other two types of data and are omitted.

Link function misspecification: we first design the following simulation setting where the treatment allocation does not follow linear logit link functions. Specifically, we simulate trt from

$$logit(Pr(trt_{ij} = 1 \mid \mathbf{x}_{ij})) \sim \begin{cases} \frac{(\tau_{ij}+1)^{1.5}-1}{1.5} & \text{if } \tau_{ij} \geq 0 \\ -\frac{(\tau_{ij}+1)^{1.2}-1}{1.2} & \text{if } \tau_{ij} < 0 \end{cases}$$

with $\tau_{ij} = -0.5x_{(1,ij)} + 0.3x_{(2,ij)} + w_i$, $x_{(1,ij)} \sim Bern(0.5)$ and $x_{(2,ij)} \sim N(0,1)$.

The responses are generated from the following model:

$$y_{ij} = 0.3 + 0.5trt_{ij} + 0.3x_{(1,ij)} + 0.2x_{(2,ij)} + \eta_i + \epsilon_{ij}$$

where $\epsilon_{ij} \sim N(0,1)$. $\eta_i = b\kappa_i$ where $\kappa_i \sim N(0.5, 1)$ and we set b to 0 or 0.3. The cluster effect w_i in treatment assignments is correlated with κ_i in responses as the following:

$$w_i \sim \begin{cases} N(0.55, 1) & \text{if } \kappa_i \geq 0 \\ N(-0.45, 1) & \text{if } \kappa_i < 0. \end{cases}$$

Covariate functional form misspecification: To further demonstrate the robustness of FM-PS model, we design the following simulation setting:

$$y_{ij} = 0.3 + 0.5trt_{ij} + 0.3x_{(1,ij)} + 0.2x^2_{(2,ij)} + \eta_i + \epsilon_{ij}$$

where ϵ_{ij}, $x_{(1,ij)}$, $x_{(2,ij)}$, and η_i are simulated the same way as the above link function misspecification scenario. Treatment is allocated via the following thresholding mechanism:

$$trt_{ij} = \begin{cases} 1 & \text{if } -0.5x_{(1,ij)} + 0.3x^3_{(2,ij)} + w_i + \delta_{ij} > 0 \\ 0 & \text{if } -0.5x_{(1,ij)} + 0.3x^3_{(2,ij)} + w_i + \delta_{ij} \leq 0 \end{cases}$$

with $\delta_{ij} \sim N(0,1)$ and w_i is set the same as in the above link function misspecification scenario. Note, in this simulation setting, covariate $x_{2,ij}$ affects both the response and the treatment allocation nonlinearly.

In all four propensity score based models we compared (i.e. NR-PS, SM-PS, FE-PS and FM-PS models), the propensity scores are estimated by the propensity score model (4.1) where only the linear covariate terms are included. Simulation results with sample size $n = 10000$ are summarized in Tables (4.5) and (4.6). Both tables show that the proposed FM-PS models are robust to data that are generated from the non-linear logit link function or the non-linear confounding covariate functional form. Furthermore, the cluster bootstrapping procedure gives valid variance estimation for $\hat{\alpha}_{trt}$ in these model misspecification scenarios.

The existence of unobserved confounding covariatesis likely to result in biased parameter estimates. Various methods have been proposed to deal with the unobserved confounding covariate from different aspects (e.g. Rosenbaum and Rubin 1983b; Lin et al. 1998; Sturmer et al. 2005). We investigate the performance of the proposed FM-PS models in handling hidden confounding covariates that are at cluster level, or at sample individual level under the single level clustering scenario.

Cluster level unobserved (hidden) confounding covariate: The treatment assignment trt_{ij} is generated via the following mechanism:

$$trt_{ij} = \begin{cases} 1 & \text{if } 0.2 - 0.7x_{(1,ij)} - 0.4x_{(2,ij)} + 0.7x_{(3,ij)} + 0.4x_{(4,i)} + w_i + \upsilon_{ij} > 0 \\ 0 & \text{if } 0.2 - 0.7x_{(1,ij)} - 0.4x_{(2,ij)} + 0.7x_{(3,ij)} + 0.4x_{(4,i)} + w_i + \upsilon_{ij} \leq 0 \end{cases}$$

where $\upsilon_{ij} \sim N(0,1)$ is a random measurement error. The response y_{ij} is generated from the following model:

$$y_{ij} = 0.3 + 0.5trt_{ij} - 0.6x_{(1,ij)} - 0.3x_{(2,ij)} + 0.6x_{(3,ij)} + 0.3x_{(4,i)} + \eta_i + \epsilon_{ij}$$

where $\epsilon_{ij} \sim N(0,1)$, $x_{(1,ij)} \sim Bern(0.5)$, $x_{(2,ij)} \sim N(0,1)$ and $x_{(3,ij)} \sim Exp(1)$. Cluster effect w_i is correlated with η_i as follows:

$$w_i \sim \begin{cases} \eta_i & \text{if } \eta_i \geq 0 \\ -Exp(1) & \text{if } \eta_i < 0 \end{cases}$$

where $\eta_i \sim \frac{1}{2}Exp(2) + \frac{1}{2}N(-0.5,1)$.

When w_i and η_i are independent, we have $w_i \sim \frac{1}{2}Exp(1) + \frac{1}{2}N(-1,1)$ and $\eta_i \sim \frac{1}{2}(-Exp(2)) + \frac{1}{2}Exp(1.5)$.

The confounding variable $x_{(4,i)} \sim \frac{1}{2}Unif(-1,1) + \frac{1}{2}N(0,1)$ is assumed to be unobserved and not included as a covariate in the PS models investigated.

Subject level unobserved confounding covariate: The simulation settings under this scenario for treatment assignment and response are nearly the same as above except that the unobserved confounding covariate, $x_{4,i}$ now varies at the subject level, and we indexed this variable as $x_{(4,ij)}$ which follows $Unif(-0.5, 1)$.

Again, four nuisance covariates, from $Bern(0.5), N(0,1), Exp(1)$, and $Unif(-1,1)$ are generated. When estimating propensity scores, only covariates, $x_{(1,ij)}, \cdots, x_{(3,ij)}$, and the nuisance covariates are included in Model (4.1). Simulation results for the unobserved confounding covariate at the cluster level and subject level are given in Table (4.7) and (4.8), respectively.

Table (4.7) demonstrates that FM-PS and FE-PS can provide unbiased treatment effect estimates in the existence of cluster level unobserved confounding covariate. Under this setting, no matter if the cluster effect in treatment assignment is independent of or correlated with the cluster effect in response, the treatment effects estimated by other PS based models, i.e. NR-PS and SM-PS, are severely biased. Checking this table also reveals that the cluster bootstrapping procedure provides valid variance estimate for $\hat{\alpha}_{trt}$. This observation indeed matches what we observed earlier. This makes sense since the random cluster effects in the FM-PS and FE-PS automatically account for the effects of the cluster level unobserved confounding covariates. In contrast, the treatment effects estimated by NR-PS and SM-PS models are all biased in all situations. However, Table (4.8) indicates that the treatment effects estimated by all models including FM-PS and FE-PS are biased in the existence of subject level unobserved confounding covariates since the strongly ignorable treatment assignment assumption is violated and no any PS approach is powerful enough to handle this confounding problem.

In summary, for unobserved confounding covariates, if they are highly correlated with other observed covariates and/or cluster effects then the treatment effect estimated by our proposed FM-PS model will be unbiased. But this conclusion may not hold for

NR-PS and SM-PS models. However, if the unobserved confounding covariate(s) is at the subject level and not highly correlated with any other observed covariates and cluster effects, the strongly ignorable treatment assignment assumption can not hold and none of the methods gives unbiased treatment effect estimates.

4.6 Real Data Analysis

To demonstrate the practical use of the proposed FM-PS regression model, we applied it to a multi-center breast cancer study conducted by the German Breast Cancer Study Group (Rauschecker et al. 1995) that we described in details in Section 3.6 of Chapter 3. Based on the preliminary analysis there, the observed covariates include age (i.e. Age), tumor size (i.e. TS), and their interactions. Therefore, we use the following logistic regression model to estimate the propensity scores for the downstream PS regression analysis with analysis results presented in Table 4.9:

$$\text{logit}\{PS_i(= Pr(trt_i = 1 \mid (Age_i, TS_i)))\} = \gamma_0 + \gamma_1 * Age_i + \gamma_2 * TS_i + \gamma_3 * Age_i * TS_i$$

From Table 4.9, an evident observation is that the treatment effect estimate based on the unadjusted model is larger than that based on other methods. After adjusting the confounding factors based on different versions of PS methods, the treatment effect estimates are reduced. Different estimation schemes give us quite different but positive treatment effect estimates. The positive number indicates that the QoL for the breast conservation group patients is on average better than that of mastectomy group patients. However, this conclusion is not statistically significant. This can be confirmed by checking the corresponding 95% confidence interval (i.e. column 95% CI).

Overall, based on the data collected for this multi-center study, there exists no statistically significant difference on the QoLs of breast cancer patients between the two treatment procedures after adjusting the patient age, tumor size, their interaction and the cluster effect.

4.7 Discussions

In this chapter, we have developed a flexible mixed effects propensity score (FM-PS) approach to model the hidden heterogeneity structure of medical record data. The FM-PS models relax the unrealistic independence assumption made by traditional mixed effects models between the random effects and the fixed effect covariates. Due to treatment allocation dynamics in real world data, this assumption rarely holds, especially for clustered medical record data. The heterogeneity across cluster units may influence both treatment allocations as well as disease outcomes. The proposed FM-PS framework relaxes the independence assumption by incorporating the proportion of patients assigned to one of the two treatment groups and the average of propensity scores for each cluster unit as additional covariates into traditional mixed effects models. Including these additional covariates effectively captures complicated correlation structures between the cluster effects and fixed effect covariates. We further show that there exists equivalence between the FM-PS and the FE-PS models for estimating treatment effects. However, our investigations (not shown) have indicated that the FM-PS models are more computationally efficient for handling large medical data. We have repeated another set of simulations with the simulation setting of Table (4.1) where sample size n is increased to 500,000 with the number of cluster units 100,000. This number is not unrealistic for electronic medical data. Interestingly, the FM-PS models run well with R function of **lmer** in package **lme4** for this simulation but the FE-PS models fail in R with the **lm** function due to a large number of dummy variables that need to be created. The FM-PS models require much less computational workload for data with a large number of cluster units and thus is more practically useful for large data like electronic medical record data where there exist hundreds and thousands of cluster units (i.e. physicians and/or clinics).

Our simulation results demonstrated that when the independence assumption for the cluster effects does not hold, the treatment effect estimated from SM-PS model and NR-PS model can be severely biased. However, both the proposed FM-PS and the FE-PS approach provide unbiased treatment effect estimates. Additionally, we demonstrate the robustness of FM-PS approach under various model misspecification. The proposed FM-PS approach can be extended to higher level clustered data straightforwardly but further research is needed to check the performance of extended methods in this regard.

Even though the proposed FM-PS approach can provide unbiased treatment effect estimate no matter if the independence assumption for the cluster effect term holds or not, the estimate is not as efficient as that from the SM-PS model when the independence assumption does hold. We propose a likelihood ratio test statistic to test the independence, which provides us a guidance on selecting an appropriate PS approach for each data.

Under the propensity score adjustment framework, Rosenbaum and Rubin (1983) have shown that the treatment effect estimate is unbiased if there is no unobserved confounding (i.e. strongly ignorable treatment assignment assumption). In real world observational data, due to the complexity in treatment allocation and factors influencing the outcome, the strong ignorable treatment assignment assumption may not hold. Our simulation studies further show that the proposed FM-PS models can handle unobserved confounding covariates reasonably well under certain simple circumstances. The performance of FM-PS with more complicated unobserved confounding factors is very critical and deserve further investigations.

Figure 4.1: Single Level Clustering with Sample Size=5000

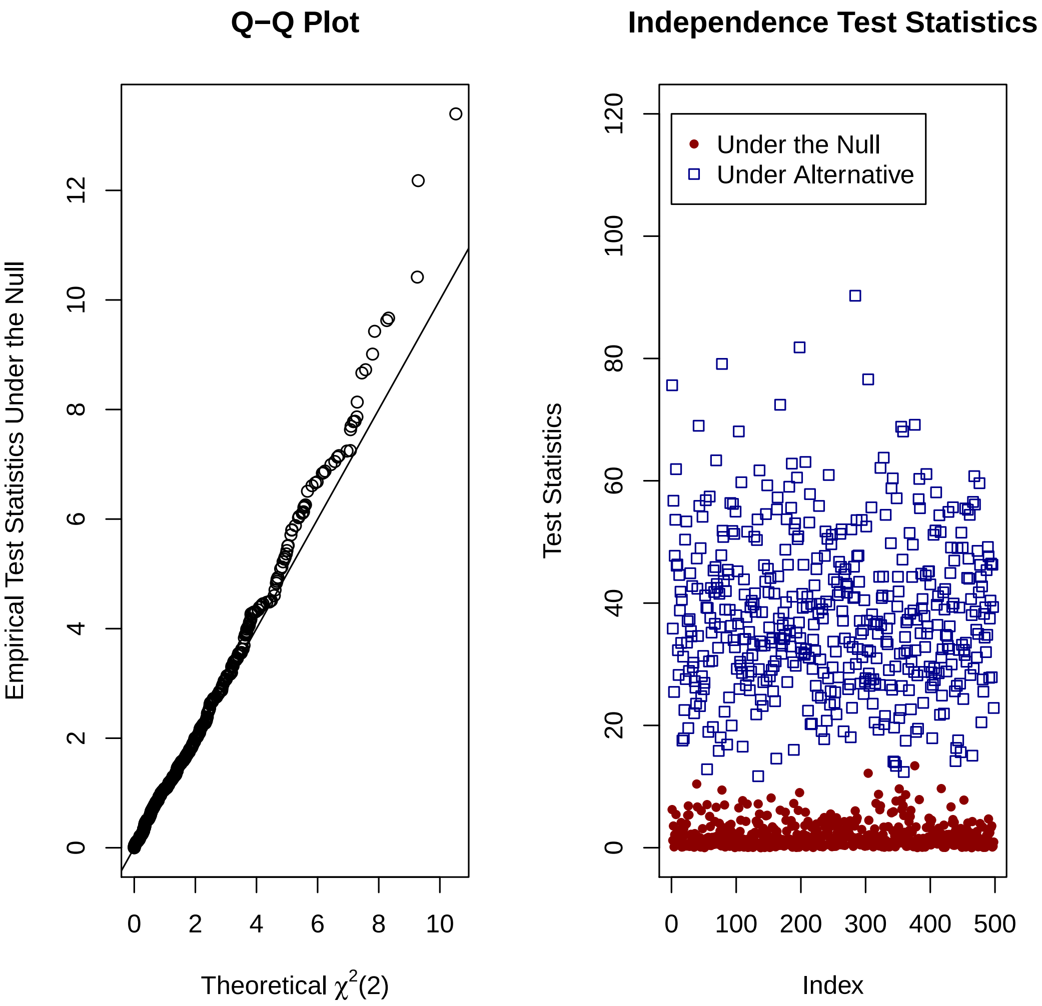

Figure 4.2: Multilevel Clustering with Sample Size=5000

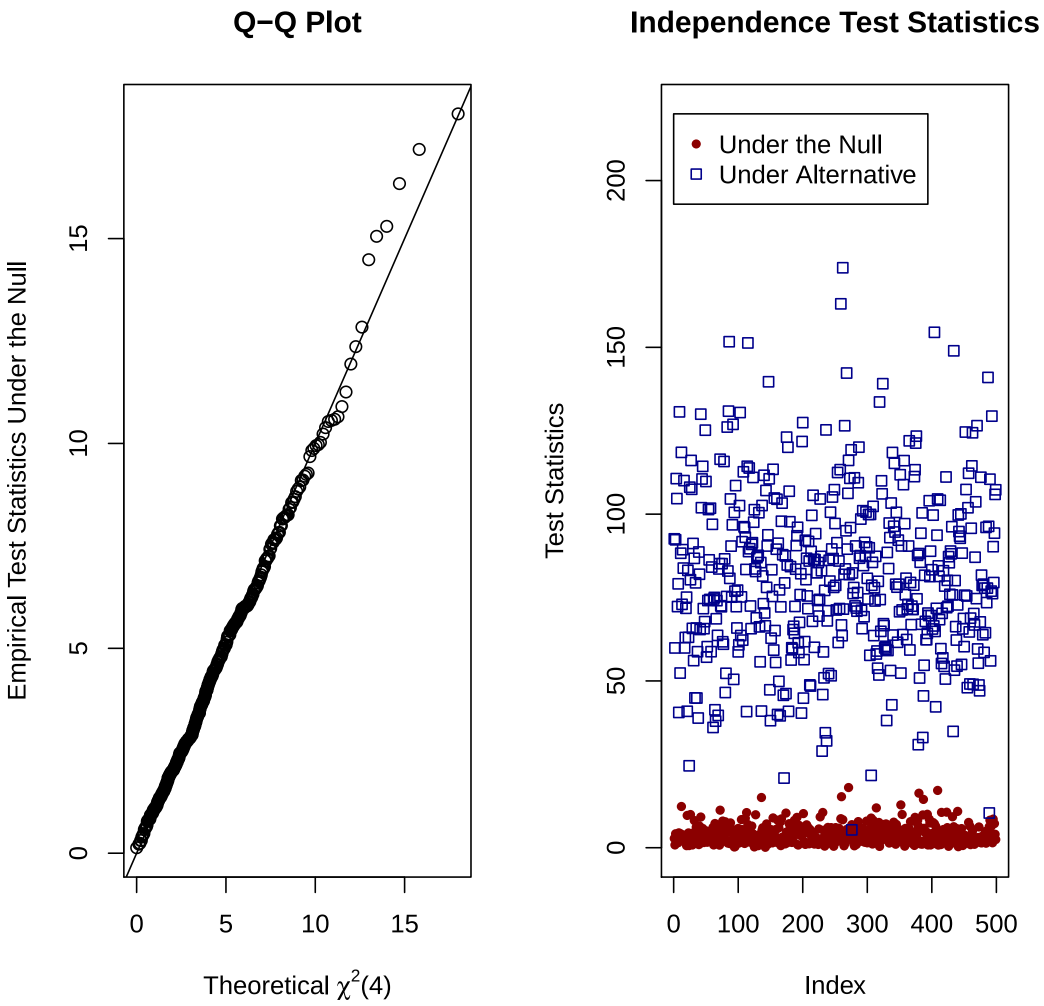

Figure 4.3: Hierarchical Clustering with Sample Size=5000

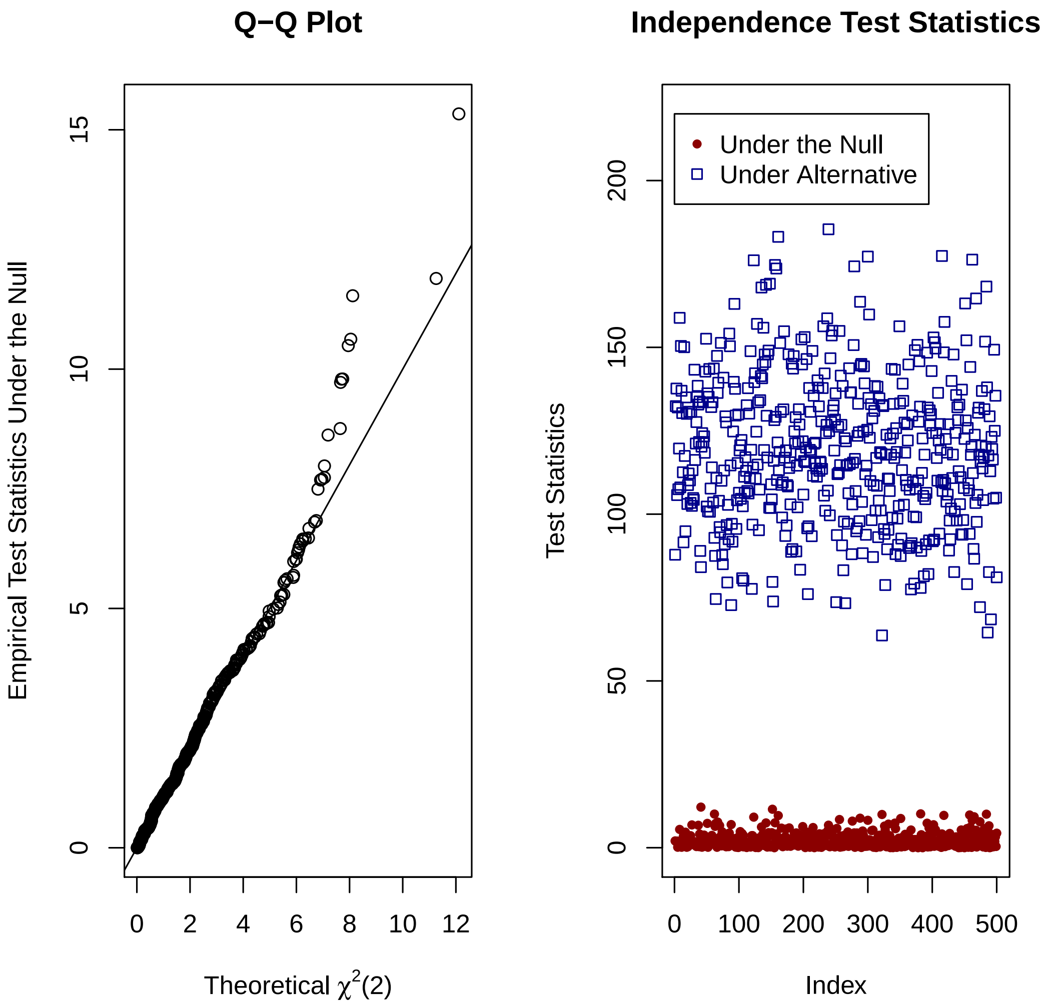

Table 4.1: Single Level Clustering with Sample Size=1000

Model	Independence	Average($\hat{\alpha}_{trt}$)	Monte Carlo $SD(\hat{\alpha}_{trt})$	Average $\widehat{SE}_D(\hat{\alpha}_{trt})$	Average $\widehat{SE}_B(\hat{\alpha}_{trt})$
Unadjusted	Yes	0.754	0.111	0.112	0.111
NR-PS		0.498	0.083	0.137	0.085
SM-PS		0.500	0.087	0.144	0.093
FE-PS		0.503	0.117	0.159	0.120
FM-PS		0.503	0.117	0.159	0.120
Unadjusted	No	0.754	0.111	0.112	0.111
NR-PS		0.747	0.085	0.137	0.087
SM-PS		0.636	0.094	0.144	0.098
FE-PS		0.503	0.117	0.159	0.120
FM-PS		0.503	0.117	0.159	0.120

Note: Results are based on 500 simulations & true $\alpha_{trt} = 0.5$

Table 4.2: Single Level Clustering with Sample Size=5000

Model	Independence	Average($\hat{\alpha}_{trt}$)	Monte Carlo $SD(\hat{\alpha}_{trt})$	Average $\widehat{SE}_D(\hat{\alpha}_{trt})$	Average $\widehat{SE}_B(\hat{\alpha}_{trt})$
Unadjusted		0.759	0.047	0.050	0.050
NR-PS		0.501	0.037	0.060	0.038
SM-PS	Yes	0.502	0.040	0.064	0.042
FE-PS		0.500	0.054	0.070	0.054
FM-PS		0.500	0.054	0.070	0.054
Unadjusted		0.759	0.047	0.050	0.050
NR-PS		0.754	0.038	0.060	0.039
SM-PS	No	0.636	0.042	0.064	0.044
FE-PS		0.500	0.054	0.070	0.054
FM-PS		0.500	0.054	0.070	0.054

Note: Results are based on 500 simulations & true $\alpha_{trt} = 0.5$

Table 4.3: Multilevel Clustering with Sample Size=5000

Model	Independence	Average($\hat{\alpha}_{trt}$)	Monte Carlo $SD(\hat{\alpha}_{trt})$	Average $\widehat{SE}_D(\hat{\alpha}_{trt})$	Average $\widehat{SE}_B(\hat{\alpha}_{trt})$
Unadjusted	Yes	1.263	0.067	0.052	0.049
NR-PS		0.509	0.074	0.081	0.064
SM-PS		0.503	0.036	0.053	0.038
FE-PS		0.505	0.061	0.066	0.058
FM-PS		0.505	0.061	0.066	0.058
Unadjusted	No	1.265	0.082	0.075	0.068
NR-PS		0.937	0.079	0.082	0.064
SM-PS		0.801	0.059	0.056	0.058
FE-PS		0.505	0.061	0.066	0.058
FM-PS		0.505	0.061	0.066	0.058

Note: Results are based on 500 simulations & true $\alpha_{trt} = 0.5$

Table 4.4: Hierarchical Clustering with Sample Size=5000

Model	Independence	Average($\hat{\alpha}_{trt}$)	Monte Carlo $SD(\hat{\alpha}_{trt})$	Average $\widehat{SE}_D(\hat{\alpha}_{trt})$	Average $\widehat{SE}_B(\hat{\alpha}_{trt})$
Unadjusted	Yes	1.173	0.060	0.056	0.059
NR-PS		0.503	0.057	0.064	0.058
SM-PS		0.502	0.042	0.062	0.044
FE-PS		0.502	0.067	0.074	0.067
FM-PS		0.502	0.067	0.074	0.067
Unadjusted	No	1.259	0.053	0.064	0.052
NR-PS		0.803	0.046	0.062	0.046
SM-PS		0.636	0.042	0.064	0.044
FE-PS		0.502	0.067	0.074	0.067
FM-PS		0.502	0.067	0.074	0.067

Note: Results are based on 500 simulations & true $\alpha_{trt} = 0.5$

Table 4.5: Link Function Misspecification with Sample Size=10000

Model	Independence	Average($\hat{\alpha}_{trt}$)	Monte Carlo $SD(\hat{\alpha}_{trt})$	Average $\widehat{SE}_D(\hat{\alpha}_{trt})$	Average $\widehat{SE}_B(\hat{\alpha}_{trt})$
Unadjusted	Yes (i.e. $b = 0$)	0.519	0.020	0.021	0.021
NR-PS		0.500	0.020	0.021	0.020
SM-PS		0.500	0.021	0.022	0.021
FE-PS		0.499	0.027	0.026	0.027
FM-PS		0.499	0.027	0.026	0.027
Unadjusted	No (i.e. $b \neq 0$)	0.612	0.022	0.021	0.022
NR-PS		0.596	0.022	0.022	0.022
SM-PS		0.549	0.022	0.023	0.022
FE-PS		0.499	0.027	0.026	0.027
FM-PS		0.499	0.027	0.026	0.027

Note: Results are based on 500 simulations & true $\alpha_{trt} = 0.5$

Table 4.6: Covariate Functional Form Misspecification with Sample Size=10000

Model	Independence	Average($\hat{\alpha}_{trt}$)	Monte Carlo $SD(\hat{\alpha}_{trt})$	Average $\widehat{SE}_D(\hat{\alpha}_{trt})$	Average $\widehat{SE}_B(\hat{\alpha}_{trt})$
Unadjusted	Yes (i.e. $b = 0$)	1.173	0.060	0.056	0.059
NR-PS		0.503	0.057	0.064	0.058
SM-PS		0.502	0.042	0.062	0.044
FE-PS		0.502	0.067	0.074	0.067
FM-PS		0.502	0.067	0.074	0.067
Unadjusted	No (i.e. $b \neq 0$)	1.259	0.053	0.064	0.052
NR-PS		0.803	0.046	0.062	0.046
SM-PS		0.636	0.042	0.064	0.044
FE-PS		0.502	0.067	0.074	0.067
FM-PS		0.502	0.067	0.074	0.067

Note: Results are based on 500 simulations & true $\alpha_{trt} = 0.5$

Table 4.7: Cluster Level Unobserved Confounding with Sample Size=5000

Model	Independence	Average($\hat{\alpha}_{trt}$)	Monte Carlo $SD(\hat{\alpha}_{trt})$	Average $\widehat{SE}_D(\hat{\alpha}_{trt})$	Average $\widehat{SE}_B(\hat{\alpha}_{trt})$
Unadjusted	Yes	1.131	0.042	0.036	0.039
NR-PS		0.593	0.039	0.034	0.038
SM-PS		0.552	0.037	0.036	0.036
FE-PS		0.501	0.042	0.040	0.040
FM-PS		0.501	0.042	0.040	0.040
Unadjusted	No	1.360	0.040	0.036	0.039
NR-PS		0.862	0.041	0.034	0.040
SM-PS		0.687	0.039	0.036	0.038
FE-PS		0.501	0.042	0.040	0.040
FM-PS		0.501	0.042	0.040	0.040

Note: Results are based on 500 simulations & true $\alpha_{trt} = 0.5$

Table 4.8: Subject Level Unobserved Confounding with Sample Size=5000

Model	Independence	Average($\hat{\alpha}_{trt}$)	Monte Carlo $SD(\hat{\alpha}_{trt})$	Average $\widehat{SE}_D(\hat{\alpha}_{trt})$	Average $\widehat{SE}_B(\hat{\alpha}_{trt})$
Unadjusted	Yes	1.074	0.039	0.037	0.038
NR-PS		0.522	0.036	0.034	0.036
SM-PS		0.526	0.034	0.035	0.035
FE-PS		0.534	0.040	0.041	0.040
FM-PS		0.534	0.040	0.041	0.040
Unadjusted	No	1.311	0.037	0.036	0.037
NR-PS		0.800	0.037	0.034	0.037
SM-PS		0.701	0.036	0.036	0.037
FE-PS		0.534	0.040	0.041	0.040
FM-PS		0.534	0.040	0.041	0.040

Note: Results are based on 500 simulations & true $\alpha_{trt} = 0.5$

Table 4.9: Multilevel Analysis for German Breast Cancer Study Data

Model	$\hat{\alpha}_{trt}$	Average $\widehat{SE}_D(\hat{\alpha}_{trt})$	Average $\widehat{SE}_B(\hat{\alpha}_{trt})$	95% CI
Unadjusted	1.533	1.279	1.527	(-1.460,4.526)
NR-PS	0.725	1.322	1.507	(-2.229,3.679)
SM-PS	1.076	1.374	1.326	(-1.523,3.675)
FE-PS	1.213	1.436	1.402	(-1.535,3.961)
FM-PS	1.213	1.436	1.402	(-1.535,3.961)

Note: Results based on 500 cluster bootstrap resampling

Chapter 5

Future Research

In this dissertation, we have proposed a two-stage variance estimation scheme for PS regression models and different PS methods for dealing with the heterogeneity in treatment assignment process for clustered data. Even though we have conducted research under various settings for the proposed methods, they are all based on the continuous response data type. Potential extensions for the proposed methods and other related research include the following:

5.1 Extending the Proposed Methods to Other Types of Data

In the future research, we plan to extend our PS models from continuous response variables to other types of response variables commonly observed in CER studies, such as the binary cure versus non-cure events, and time to event data (i.e. cure or death). We plan to combine techniques such as multinomial or ordered logit, parametric and non-parametric survival analysis with the proposed PS models where the treatment assignment is heterogeneous. Longitudinal resources where patients are followed up over years of medical interventions, which frequently exist in observational data, would provide us with great opportunities to assess true treatment effects yet at the same time pose more challenges in dealing with time-dependent confounding. Developing robust PS models that adjust time-dependent confounding will be one of our future research focus.

5.2 Developing New Methods with Missing Confounding Factors

Missing data is a commonly occurring complication in scientific investigations. In CER studies, missing important confounding covariates could have a significant impact on the validity of the estimation of the true treatment effect. Determining the appropriate analytic approach in the presence of incomplete observations is a major problem for data analysts. The development of statistical methods to address missing data has been an active area of research. For heterogeneous observational data, we plan to combine nonparametric kernel regression methods for missing important variable with the proposed PS models under various data missing mechanisms.

In addition, for CER studies with missing important confounding covariates, often there exist some readily available auxiliary covariates information. Auxiliary information can also be obtained from inaccurate measures of the confounding variables. For example, it is found that the self reported height is often inflated in males. Other situations where auxiliary confounding information arises are where the exposure assessment relies on the subjects responses to questionnaires, or the exposure of interest may be too difficult or expensive to obtain. In such situations, a related variable may be used as an auxiliary variable for the exposure of interest. There are some existing methods in the statistical literatures on using auxiliary variables, however, there is few research on using the auxiliary variable in sharpening the estimation of the treatment effects, not to mention under the intractable heterogeneity found in the treatment assignment. We plan to extend the proposed PS methods to data with auxiliary covariates.

APPENDIX I: PROOF OF THEOREM 2.3.1

Proof. Denote $\boldsymbol{\theta}_1^\star$ and $\boldsymbol{\theta}_2^\star$ as the true parameter values of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ under the models (2.2) and (2.3). Following similar procedures of Murphy and Topel (1985), with the notations defined in Section 2.3, the MLE of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, i.e. $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ derived from Stage 1 and 2 models satisfy the following score equations:

$$\begin{aligned} \sum_{i=1}^{n} \frac{\partial l_{1,i}(\hat{\boldsymbol{\theta}}_1)}{\partial \boldsymbol{\theta}_1} &= 0 \\ \sum_{i=1}^{n} \frac{\partial l_{2,i}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)}{\partial \boldsymbol{\theta}_2} &= 0 \end{aligned}$$

Under the standard regularity conditions, $\hat{\boldsymbol{\theta}}_1$ is consistent. Therefore, the maximization of quantity $\frac{1}{n}\sum_{i=1}^{n} l_{2,i}(\hat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2)$ is asymptotically equivalent to the maximization of $\frac{1}{n}\sum_{i=1}^{n} l_{2,i}(\boldsymbol{\theta}_1^\star, \boldsymbol{\theta}_2)$. Thus, $\hat{\boldsymbol{\theta}}_2$ is consistent.

Taking Taylor expansions on $\frac{\partial l_{1,i}(\hat{\boldsymbol{\theta}}_1)}{\partial \boldsymbol{\theta}_1}$ and $\frac{\partial l_{2,i}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)}{\partial \boldsymbol{\theta}_2}$ at $\boldsymbol{\theta}^\star = (\boldsymbol{\theta}_1^\star, \boldsymbol{\theta}_2^\star)$, we obtain the following approximations:

$$\begin{aligned} &\frac{\partial l_{1,i}(\hat{\boldsymbol{\theta}}_1)}{\partial \boldsymbol{\theta}_1} \approx \frac{\partial l_{1,i}(\boldsymbol{\theta}_1^\star)}{\partial \boldsymbol{\theta}_1} + \frac{\partial^2 l_{1,i}(\boldsymbol{\theta}_1^\star)}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1^T}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^\star) \\ &\frac{\partial l_{2,i}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)}{\partial \boldsymbol{\theta}_2} \approx \frac{\partial l_{2,i}(\boldsymbol{\theta}_1^\star, \boldsymbol{\theta}_2^\star)}{\partial \boldsymbol{\theta}_2} + \frac{\partial^2 l_{2,i}(\boldsymbol{\theta}_1^\star, \boldsymbol{\theta}_2^\star)}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_1^T}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^\star) + \frac{\partial^2 l_{2,i}(\boldsymbol{\theta}_1^\star, \boldsymbol{\theta}_2^\star)}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_2^T}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2^\star) \end{aligned}$$

Plugging the above two terms into the score equations, we immediately obtain the following:

$$\begin{aligned}
-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_{1,i}(\boldsymbol{\theta}_1^\star)}{\partial \boldsymbol{\theta}_1} &\approx \frac{1}{n}\{\sum_{i=1}^{n}\frac{\partial^2 l_{1,i}(\boldsymbol{\theta}_1^\star)}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1^T}\}\sqrt{n}(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^\star) \qquad (5.1)\\
-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_{2,i}(\boldsymbol{\theta}_1^\star,\boldsymbol{\theta}_2^\star)}{\partial \boldsymbol{\theta}_2} &\approx \frac{1}{n}\{\sum_{i=1}^{n}\frac{\partial^2 l_{2,i}(\boldsymbol{\theta}_1^\star,\boldsymbol{\theta}_2^\star)}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_1^T}\}\sqrt{n}(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^\star)\\
&+ \frac{1}{n}\{\sum_{i=1}^{n}\frac{\partial^2 l_{2,i}(\boldsymbol{\theta}_1^\star,\boldsymbol{\theta}_2^\star)}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_2^T}\}\sqrt{n}(\hat{\boldsymbol{\theta}}_2-\boldsymbol{\theta}_2^\star) \qquad (5.2)
\end{aligned}$$

By the central limit theorem, we conclude that the joint distribution of statistics $-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_{1,i}(\boldsymbol{\theta}_1^\star)}{\partial \boldsymbol{\theta}_1}$ and $-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_{2,i}(\boldsymbol{\theta}_1^\star,\boldsymbol{\theta}_2^\star)}{\partial \boldsymbol{\theta}_2}$ is normal and given by the following:

$$\begin{bmatrix} -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_{1,i}(\boldsymbol{\theta}_1^\star)}{\partial \boldsymbol{\theta}_1} \\ -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_{2,i}(\boldsymbol{\theta}_1^\star,\boldsymbol{\theta}_2^\star)}{\partial \boldsymbol{\theta}_2} \end{bmatrix} \to N(0,\mathbf{V})$$

where $\mathbf{V} = \left(\begin{smallmatrix} \mathbf{V}_1^{-1}(\boldsymbol{\theta}_1) & \mathbf{R}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) \\ \mathbf{R}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) & \mathbf{V}_2^{-1}(\boldsymbol{\theta}_2) \end{smallmatrix}\right)$.

By the law of large number theorem, the asymptotic equivalence of (5.1) can be written as the following:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1^\star) \approx \mathbf{V}_1(\boldsymbol{\theta}_1)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_{1,i}(\boldsymbol{\theta}_1^\star)}{\partial \boldsymbol{\theta}_1} \qquad (5.3)$$

Plugging (5.3) into (5.2) and applying the law of large number theorem again, we have:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_2-\boldsymbol{\theta}_2^\star) \approx \mathbf{V}_2(\boldsymbol{\theta}_2)\mathbf{C}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\mathbf{V}_1(\boldsymbol{\theta}_1)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_{1,i}(\boldsymbol{\theta}_1^\star)}{\partial \boldsymbol{\theta}_1} + \mathbf{V}_2(\boldsymbol{\theta}_2)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_{2,i}(\boldsymbol{\theta}_1^\star,\boldsymbol{\theta}_2^\star)}{\partial \boldsymbol{\theta}_2}$$

By the joint distribution of $-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_{1,i}(\boldsymbol{\theta}_1^\star)}{\partial \boldsymbol{\theta}_1}$ & $-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_{2,i}(\boldsymbol{\theta}_1^\star,\boldsymbol{\theta}_2^\star)}{\partial \boldsymbol{\theta}_2}$, we obtain the asymptotic distribution of $\hat{\boldsymbol{\theta}}_2$:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2^{\star}) \to N(0, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma} = \mathbf{V}_2 + \mathbf{V}_2[\mathbf{C}\mathbf{V}_1\mathbf{C}^T - \mathbf{R}\mathbf{V}_1\mathbf{C}^T - \mathbf{C}\mathbf{V}_1\mathbf{R}^T]\mathbf{V}_2$ and conclusions follow immediately.

□

APPENDIX II: PROOF OF THEOREM 3.4.1

Proof. Denote $\boldsymbol{\theta}_1^\star$ and $\boldsymbol{\theta}_2^\star$ as the true parameter values of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ under the models (3.3) and (3.6). Following similar procedures of Murphy and Topel (1985), with the notations defined in Section 3.4, the MLE of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, i.e. $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ derived from models (3.3) and (3.6) satisfy the following score equations:

$$\begin{aligned}\sum_{i=1}^{n} \frac{\partial l_{1,i}(\hat{\boldsymbol{\theta}}_1)}{\partial \boldsymbol{\theta}_1} &= 0 \\ \sum_{i=1}^{n} \frac{\partial l_{2,i}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)}{\partial \boldsymbol{\theta}_2} &= 0\end{aligned}$$

Under the standard regularity conditions, $\hat{\boldsymbol{\theta}}_1$ is consistent. Therefore, the maximization of quantity $\frac{1}{n}\sum_{i=1}^{n} l_{2,i}(\hat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2)$ is asymptotically equivalent to the maximization of $\frac{1}{n}\sum_{i=1}^{n} l_{2,i}(\boldsymbol{\theta}_1^\star, \boldsymbol{\theta}_2)$. Thus, $\hat{\boldsymbol{\theta}}_2$ is consistent.

Taking Taylor expansions on $\frac{\partial l_{1,i}(\hat{\boldsymbol{\theta}}_1)}{\partial \boldsymbol{\theta}_1}$ and $\frac{\partial l_{2,i}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)}{\partial \boldsymbol{\theta}_2}$ at $\boldsymbol{\theta}^\star = (\boldsymbol{\theta}_1^\star, \boldsymbol{\theta}_2^\star)$, we obtain the following approximations:

$$\begin{aligned}&\frac{\partial l_{1,i}(\hat{\boldsymbol{\theta}}_1)}{\partial \boldsymbol{\theta}_1} \approx \frac{\partial l_{1,i}(\boldsymbol{\theta}_1^\star)}{\partial \boldsymbol{\theta}_1} + \frac{\partial^2 l_{1,i}(\boldsymbol{\theta}_1^\star)}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1^T}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^\star) \\ &\frac{\partial l_{2,i}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)}{\partial \boldsymbol{\theta}_2} \approx \frac{\partial l_{2,i}(\boldsymbol{\theta}_1^\star, \boldsymbol{\theta}_2^\star)}{\partial \boldsymbol{\theta}_2} + \frac{\partial^2 l_{2,i}(\boldsymbol{\theta}_1^\star, \boldsymbol{\theta}_2^\star)}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_1^T}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^\star) + \frac{\partial^2 l_{2,i}(\boldsymbol{\theta}_1^\star, \boldsymbol{\theta}_2^\star)}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_2^T}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2^\star)\end{aligned}$$

Plugging the above two terms into the score equations, we immediately obtain the following:

$$
\begin{aligned}
-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_{1,i}(\boldsymbol{\theta}_1^\star)}{\partial \boldsymbol{\theta}_1} &\approx \frac{1}{n}\{\sum_{i=1}^{n}\frac{\partial^2 l_{1,i}(\boldsymbol{\theta}_1^\star)}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1^T}\}\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^\star) \qquad (5.4)\\
-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_{2,i}(\boldsymbol{\theta}_1^\star, \boldsymbol{\theta}_2^\star)}{\partial \boldsymbol{\theta}_2} &\approx \frac{1}{n}\{\sum_{i=1}^{n}\frac{\partial^2 l_{2,i}(\boldsymbol{\theta}_1^\star, \boldsymbol{\theta}_2^\star)}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_1^T}\}\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^\star)\\
&+ \frac{1}{n}\{\sum_{i=1}^{n}\frac{\partial^2 l_{2,i}(\boldsymbol{\theta}_1^\star, \boldsymbol{\theta}_2^\star)}{\partial \boldsymbol{\theta}_2 \partial \boldsymbol{\theta}_2^T}\}\sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2^\star) \qquad (5.5)
\end{aligned}
$$

Following the central limit theorem, we conclude that the joint distribution of statistics $-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_{1,i}(\boldsymbol{\theta}_1^\star)}{\partial \boldsymbol{\theta}_1}$ and $-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_{2,i}(\boldsymbol{\theta}_1^\star,\boldsymbol{\theta}_2^\star)}{\partial \boldsymbol{\theta}_2}$ is normal and given by:

$$
\begin{bmatrix} -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_{1,i}(\boldsymbol{\theta}_1^\star)}{\partial \boldsymbol{\theta}_1} \\ -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_{2,i}(\boldsymbol{\theta}_1^\star,\boldsymbol{\theta}_2^\star)}{\partial \boldsymbol{\theta}_2} \end{bmatrix} \rightarrow N(0, \mathbf{V})
$$

where $\mathbf{V} = \left(\begin{smallmatrix} \mathbf{V}_1^{-1}(\boldsymbol{\theta}_1) & \mathbf{R}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) \\ \mathbf{R}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) & \mathbf{V}_2^{-1}(\boldsymbol{\theta}_2) \end{smallmatrix}\right)$, $\boldsymbol{V}_1^{-1} = E\left\{\left(\frac{\partial l_1}{\partial \boldsymbol{\theta}_1}\right)\left(\frac{\partial l_1}{\partial \boldsymbol{\theta}_1^T}\right)\right\}$, $\boldsymbol{V}_2^{-1} = E\left\{\left(\frac{\partial l_2}{\partial \boldsymbol{\theta}_2}\right)\left(\frac{\partial l_2}{\partial \boldsymbol{\theta}_2^T}\right)\right\}$, and $\boldsymbol{R} = E\left\{\left(\frac{\partial l_2}{\partial \boldsymbol{\theta}_2}\right)\left(\frac{\partial l_1}{\partial \boldsymbol{\theta}_1^T}\right)\right\}$.

Furthermore, by the law of large number theorem, the asymptotic equivalence of (5.4) can be written as the following:

$$
\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^\star) \approx \mathbf{V}_1(\boldsymbol{\theta}_1)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_{1,i}(\boldsymbol{\theta}_1^\star)}{\partial \boldsymbol{\theta}_1} \qquad (5.6)
$$

Plugging (5.6) into (5.5) and applying the law of large number theorem again, we have:

$$
\sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2^\star) \approx \mathbf{V}_2(\boldsymbol{\theta}_2)\mathbf{C}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\mathbf{V}_1(\boldsymbol{\theta}_1)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_{1,i}(\boldsymbol{\theta}_1^\star)}{\partial \boldsymbol{\theta}_1} + \mathbf{V}_2(\boldsymbol{\theta}_2)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial l_{2,i}(\boldsymbol{\theta}_1^\star, \boldsymbol{\theta}_2^\star)}{\partial \boldsymbol{\theta}_2}
$$

where $\boldsymbol{C} = E\left\{\left(\frac{\partial l_2}{\partial \boldsymbol{\theta}_2}\right)\left(\frac{\partial l_2}{\partial \boldsymbol{\theta}_1^T}\right)\right\}$.

By the joint distribution of $-\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\partial l_{1,i}(\boldsymbol{\theta}_1^\star)}{\partial \boldsymbol{\theta}_1}$ & $-\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\partial l_{2,i}(\boldsymbol{\theta}_1^\star,\boldsymbol{\theta}_2^\star)}{\partial \boldsymbol{\theta}_2}$, we obtain the asymptotic distribution of $\hat{\boldsymbol{\theta}}_2$:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2^\star) \to N(0, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma} = \mathbf{V}_2 + \mathbf{V}_2[\mathbf{C}\mathbf{V}_1\mathbf{C}^T - \mathbf{R}\mathbf{V}_1\mathbf{C}^T - \mathbf{C}\mathbf{V}_1\mathbf{R}^T]\mathbf{V}_2$ and conclusions follow immediately.

□

APPENDIX III: SAMPLING UNDER SNP DENSITY

Since the density under SNP representation is not standard, the sampling from SNP density, $f_K(z;\psi)$, is not straightforward and cannot be called within the existing statistical software packages. Below we provide detailed sampling procedures to sample from SNP density for $K \leq 2$, and more details can be found in Gallant and Tauchen (1992). Obviously, for the case of $K = 0$, the distribution reduces to a normal distribution and thus can be easily sampled. For $K > 0$ we need to use rejection sampling procedure which requires an envelope (say $d_K(z;\psi)$) that is easy to sample and dominates $f_K(z;\psi)$.

SNP sampling (K=1): For $K = 1$, we have the density function as the following:

$$f_1(z;\psi) = (a + zb)^2\phi(z) \leq (|a| + |zb|)^2\phi(z)$$

where $a = cos(\psi)$, $b = sin(\psi)$ and $\phi(z)$ is the standard normal density. Therefore, we construct the envelope as $d_1(z;\psi) = a^2\phi(z) + b^2z^2\phi(z) + 2|ab||z|\phi(z)$. Before moving on to obtain the samples from density of $d_1(z;\psi)$, we establish a useful equation that connects the density of random variable Z, i.e. $f_Z(z)$ and that of the random variable $U = Z^2$, i.e. $f_U(u)$, given as the following:

$$f_U(u) = \frac{1}{2}u^{-\frac{1}{2}}[f_Z(\sqrt{u}) + f_Z(-\sqrt{u})]$$

If $f_Z(z)$ is a symmetric function, the above equation can be further simplified as

$$f_U(u) = u^{-\frac{1}{2}}f_Z(\sqrt{u})$$

Therefore, if random variable Z follows $f(z) = c|z|\phi(z)$, then $U = Z^2 \sim f_U(u) = \frac{c}{\sqrt{2\pi}}e^{-\frac{u}{2}}$. However, $\frac{1}{2}e^{-\frac{u}{2}}$ is recognized as a χ^2_2 density. Thus, random variable $Z \sim$

$f(z) = \frac{\sqrt{2\pi}}{2}|z|\phi(z)$ follows χ_2 distribution. Following similar derivations, we conclude that $Z \sim f(z) = z^2\phi(z)$ follows a χ_3 distribution. It is evident that $\phi(z)$ is a χ_1 density. Therefore, we conclude $d_1(z;\psi) = a^2\phi(z) + b^2z^2\phi(z) + 2|ab||z|\phi(z)$ is a weighted mixture of χ density with weight $\frac{a^2}{w}$ from χ_1, $\frac{b^2}{w}$ from χ_3 and $\frac{4|ab|}{w\sqrt{2\pi}}$ from χ_2 where $w = a^2 + b^2 + \frac{4|ab|}{\sqrt{2\pi}}$. A χ distribution can be sampled from χ^2 distribution with 50% chance to be positive and negative, respectively. In summary, the procedures to draw samples from $f_1(z;\psi)$ density are:

Step 1: Draw samples from density $d_1(z;\psi) = a^2\phi(z) + b^2z^2\phi(z) + 2|ab||z|\phi(z)$ which is a weighted mixture of χ density with weights given above.

Step 2: Use rejection sampling to determine to accept or reject the sample obtained from Step 1.

Step 3: If sample is accepted, then the sample is from $f_1(z;\psi)$. Otherwise, start from Step 1 again. Repeat these procedures till the desired number of samples are drawn.

The envelope is composed of three parts and each follows a χ distribution. To see this clearly, Let $f(z) = c|z|\phi(z)$ be the density of random variable Z. Furthermore, let $U = Z^2$ then the density of U, i.e. $f_U(u)$, and Z, i.e. $f_Z(z)$, is given by $f_U(u) = \frac{1}{2}u^{-\frac{1}{2}}[f_Z(\sqrt{u}) + f_Z(-\sqrt{u})]$. If $f_Z(z)$ is symmetric, then it can be simplified as $f_U(u) = u^{-\frac{1}{2}}f_Z(\sqrt{u})$. Therefore, if $Z \sim f(z) = c|z|\phi(z)$, then $U = Z^2 \sim f_U(u) = \frac{c}{\sqrt{2\pi}}e^{-\frac{u}{2}}$ where $\frac{1}{2}e^{-\frac{u}{2}}$ is recognized as a χ^2_2 density and this leads to $c = \frac{\sqrt{2\pi}}{2}$. That is random variable $Z \sim f(z) = \frac{\sqrt{2\pi}}{2}|z|\phi(z)$ follows χ_2 distribution which can be sampled from χ^2_2 and taking square root with probability of 0.5 being positive and negative due to the symmetry about 0. Similarly, for random variable $Z \sim f(z) = cz^2\phi(z)$, then $U = Z^2 \sim f_U(u) = \frac{c}{\sqrt{2\pi}}u^{\frac{1}{2}}e^{-\frac{u}{2}}$ and $\frac{1}{\sqrt{2\pi}}u^{\frac{1}{2}}e^{-\frac{u}{2}}$ is a χ^2_3 density. Thus, $Z \sim f(z) = z^2\phi(z)$ follows a χ_3 distribution which can be drawn from a χ^2_3 and taking square root with probability 0.5 being positive and negative due to symmetric about 0. Therefore, $d_1(z;\psi)$ is a weighted χ mixture with weight $\frac{a^2}{w}$ from χ_1, $\frac{b^2}{w}$ from χ_3 and $\frac{4|ab|}{w\sqrt{2\pi}}$ from χ_2 where $w = a^2+b^2+\frac{4|ab|}{\sqrt{2\pi}}$.

SNP sampling (K=2): For $K = 2$, we have the following:

$$f_2(z;\psi) = (a + zb + z^2c)^2\phi(z) \leq (|a| + |zb| + z^2|c|)^2\phi(z)$$

where $a = cos(\psi_1) - \frac{sin(\psi_1)sin(\psi_2)}{\sqrt{2}}$, $b = sin(\psi_1)cos(\psi_2)$ and $c = \frac{sin(\psi_1)sin(\psi_2)}{\sqrt{2}}$. Therefore, we construct the envelope as:

$$d_2(z;\psi) = a^2\phi(z) + 2|ab||z|\phi(z) + (b^2 + |2ac|)z^2\phi(z) + |2bc||z^3|\phi(z) + c^2z^4\phi(z)$$

Following similar tedious derivations as above, we conclude $d_2(z;\psi)$ is also a mixture of χ density with weight $\frac{a^2}{w}$ of χ_1, $\frac{4|ab|}{w\sqrt{2\pi}}$ being χ_2, $\frac{b^2+|2ac|}{w}$ from χ_3, $\frac{8|bc|}{w\sqrt{2\pi}}$ being χ_4 and weight $\frac{3c^2}{w}$ from χ_5 with $w = a^2 + \frac{4|ab|}{\sqrt{2\pi}} + b^2 + |2ac| + \frac{8|bc|}{\sqrt{2\pi}} + 3c^2$. Therefore, similar procedures as above can be used to draw samples from density $f_2(z;\psi)$.

To see this clearly, we let $f(z) = c|z^3|\phi(z)$ be the density of random variable Z and $U = Z^2$. Thus, $f_U(u) = u^{-\frac{1}{2}}[\frac{c}{\sqrt{2\pi}}u^{\frac{3}{2}}e^{-\frac{u}{2}}]$ and $\frac{u}{4}e^{-\frac{u}{2}}$ is the density of χ_4^2. Thus, $Z \sim \frac{\sqrt{2\pi}}{4}|z^3|\phi(z)$ can be sampled from χ_4 distribution by taking the square root of a random sample from χ_4^2 distribution with the probability of 0.5 being positive and negative, respectively. Similarly, $Z \sim f(z) = cz^4\phi(z)$ and $U = Z^2$ lead to $f_U(u) = \frac{c}{\sqrt{2\pi}}u^{\frac{3}{2}}e^{-\frac{u}{2}}$. If $c = \frac{1}{3}$, then $f_U(u)$ is the χ_5^2 density and we conclude that $Z \sim f(z) = \frac{1}{3}z^4\phi(z)$ can be drawn from χ_5^2 distribution and take square root with 0.5 probability being positive and negative. Therefore, $d_2(z;\psi)$ is a weighted χ density with weight $\frac{a^2}{w}$ of χ_1, $\frac{4|ab|}{w\sqrt{2\pi}}$ being χ_2, $\frac{b^2+|2ac|}{w}$ from χ_3, $\frac{8|bc|}{w\sqrt{2\pi}}$ being χ_4 and weight $\frac{3c^2}{w}$ from χ_5 with $w = a^2 + \frac{4|ab|}{\sqrt{2\pi}} + b^2 + |2ac| + \frac{8|bc|}{\sqrt{2\pi}} + 3c^2$.

BIBLIOGRAPHY

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] M.K. Andersson and S. Karlsson. Bootstrapping error component models. *Computnl. Statist.*, 16:221–231, 2001.

[3] J. Angrist and A. Krueger. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15(4):69–85, 2001.

[4] J.D. Angrist, G.W. Imbens, and D.B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

[5] P.C. Austin. An introduction to propensity score methods for reducing the effectsof confounding in observational studies. *Multivariate Behav Res.*, 46(3):399–424, 2011.

[6] H. Bang and J. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–972, 2005.

[7] S. Bangalore, G. Steg, P. Deedwania, K. Crowley, K.A. Eagle, S. Goto, E.M. Ohman, C.P. Cannon, S.C. Smith, U. Zeymer, E.B. Hoffman, F.H. Messerli, and D.L. Bhatt. Beta-blocker use and clinical outcomes in stable outpatients with and without coronary artery disease. *Journal of the American Medical Association*, 308:1340–1349, 2012.

[8] H. Becher. The concept of residual confounding in regression models and some applications. *Statistics in Medicine*, 11:1747–1758, 1992.

[9] K. Benson and A.J. Hartz. A comparison of observational studies and randomized, controlled trials. *New England Journal of Medcine*, 342(25):1878–1886, 2000.

[10] M.L. Berger, M. Mamdani, D. Atkins, and M.L. Johnson. Good research practices for comparative effectiveness researh: Defining, reporting and interpreting nonrandomized studies of treatment effects using secondary data sources: The ispor good research practices for retrospective database analysis task force report—part i. *Value In Health*, 12(8):1044–1052, 2009.

[11] J.G. Booth and J.P. Hobert. Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society, Series B*, 61:265–285, 1999.

[12] Z.I. Botev, J.F. Grotowski, and D.P. Kroese. Kernel density estimation via diffusion. *Annals of Statistics*, 38(5):2916–2957, 2010.

[13] N.E. Breslow and D.G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25, 1993.

[14] N.E. Breslow and X. Lin. Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika*, 82:81–91, 1995.

[15] M.A. Brookhart, S. Schneeweiss, K.J. Rothman, R.J. Glynn, J. Avorn, and T. Sturmer. Variable selection for propensity score models. *American Journal of Epidemiology*, 141(12):1–8, 2006.

[16] M.A. Brookhart, P.S. Wang, D.H. Solomon, and S. Schneeweiss. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology*, 17(3):268–275, 2006.

[17] F.B. Butar and P. Lahiri. On measures of uncertainty of empirical bayes small area estimators. *J. Statist. Planng. Inf.*, 112:63–76, 2003.

[18] R.H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[19] J.R. Carpenter, H. Goldstein, and J. Rasbash. A novel bootstrap procedure for assessing the relationship between class size and achievement. *Appl. Statist.*, 52:431–443, 2003.

[20] R.D. Cebul, T.E. Love, A.K. Jain, and C.J. Hebert. Electronic health records and quality of diabetes care. *New England Journal of Medicine*, 365:825–833, 2011.

[21] J. Chen, D. Zhang, and M. Davidian. A monte carlo em algorithm for generalized linear mixed models with flexible random effects distribution. *Biostatistics*, 3(3):347–360, 2002.

[22] W.G. Cochran. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24:295–313, 1968.

[23] J.B. Cologne, R.L. Carter, S. Fujita, and S. Ban. Application of generalized estimating equations to a study of in vitro radiation sensitivity. *Biometrics*, 49:927–934, 1993.

[24] J. Concato, N. Shah, and R.I. Horwitz. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medcine*, 342(25):1887–1892, 2000.

[25] J.L. Czajka, S.M. Hirabayashi, R.J.A. Little, and D.B. Rubin. Projecting from advance data using propensity modeling: an application to income and tax statistics. *Journal of Business and Economic Statistics*, 10:117–131, 1992.

[26] M. Davidian and A.R. Gallant. The nonlinear mixed effects model with a smooth random effects density. *Biometrika*, 80:475–488, 1993.

[27] M. Davidian and D.M. Gltinan. *Nonlinear models for repeated measurement data*. Chapman and Hall, New York, 1995.

[28] M. Davidian and D.M. Gltinan. Nonlinear models for repeated measurement data: an overview and update. *Journal of Agricultural, Biological, and Environmental Statistics*, 8:387–419, 2003.

[29] A.C. Davison and D.V. Hinkley. *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge, 1997.

[30] N.E. Day, D.P. Byar, and S.B. Green. Overadjustment in case-control studies. *American Journal of Epidemiology*, 112(5):696–706, 1980.

[31] R.H. Dehejia and S. Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062, 1999.

[32] R.H. Dehejia and S. Wahba. Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, 84(1):151–161, 2002.

[33] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[34] P.J. Diggle, P. Heagerty, K.Y. Liang, and S.L. Zeger. *Analysis of Longitudinal Data*. Oxford Statistical Science Series, Oxford, 2002.

[35] D.P. Do and B.K. Finch. The link between neighborhood poverty and health: context or composition? *American Journal of Epidemiology*, 168:611–619, 2008.

[36] R.B. DAgostino. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, 17(19):2265–2281, 1998.

[37] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[38] M. Eklind-Cervenka, L. Benson, U. Dahlstrom, M. Edner, M. Rosenqvist, and L.H. Lund. Association of candesartan vs losartan with all-cause mortality in patients with heart failure. *Journal of the American Medical Association*, 305:175–182, 2011.

[39] C.A. Field and A.H. Welsh. Bootstrapping clustered data. *J. R. Statist. Soc. B*, 69:369–390, 2007.

[40] G.M. Fitzmaurice, N.M. Laird, and J.H. Ware. *Applied Longitudinal Analysis*. Wiley, New Jersey, 2004.

[41] A.R. Gallant and D.W. Nychka. Semi-nonparametric maximum likelihood estimation. *Econometrica*, 55(2):363–390, 1987.

[42] A.R. Gallant and G.E. Tauchen. *A nonparametric approach to nonlinear time series analysis: estimation and simulation*. Springer, New York, 1992.

[43] H. Goldstein, W. Browne, and J. Rasbash. Tutorial in biostatistics: Multilevel modeling of medical data. *Statistics in Medicine*, 21:3291–3315, 2002.

[44] S. Greenland, H. Morgenstern, and D.C. Thomas. Considerations in determining matching criteria and stratum sizes for case-control studies. *Int J Epidemiol*, 10:389–392, 1981.

[45] J.A. Hanley, A. Negassa, M.D. Edwardes, and J.E. Forrester. Statistical analysis of correlated data using generalized estimating equations: An orientation. *American Journal of Epidemiology*, 157(4):364–375, 2003.

[46] J. Hardin and J. Hilbe. *Generalized Estimating Equations*. Chapman and Hall/CRC, London, 2003.

[47] H. He and M.P. McDermott. A robust method using propensity score stratification for correcting verification bias for binary tests. *Biostatistics*, 13(1):32–47, 2012.

[48] J. Heckman. Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources*, 32(3):441–462, 1997.

[49] J. Hill. Comments on ”a critical appraisal of propensity-score matching in the medical literature between 1996 and 2003”. *Statistics in Medicine*, 27:2055–2061, 2008.

[50] G. Hong and B. Yu. Effects of kindergarten retention on childrens social-emotional development: An application of propensity score method to multivariate multi-level data. *Developmental Psychology*, 44(2):407–421, 2008.

[51] I.C. Huang, C. Frangakis, F. Dominici, G.B. Diette, and A.W. Wu. Application of a propensity score approach for risk adjustment in profiling multiple physician groups on asthma care. *Health Serv. Res.*, 40(1):253–278, 2005.

[52] The PROTECT Investigators. Dalteparin versus unfractionated heparin in critically ill patients. *New England Journal of Medcine*, 364:1305–1314, 2011.

[53] J.A. Ioannidis and J. Lau. Completeness of safety reporting in randomized trials. *The Journal of the American Medical Association*, 285(4):437–443, 2001.

[54] J.P. Ioannidis, A.B. Haidich, M. Pappa, N. Pantazis, S.I. Kokori, M.G. Tektonidou, D.G. Contopoulos-Ioannidis, and J. Lau. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *The Journal of the American Medical Association*, 286(7):821–830, 2001.

[55] R.S. Jackson, D.C. Chang, and J.A. Freischlag. Comparison of long-term survival after open vs endovascular repair of intact abdominal aortic aneurysm among medicare beneficiaries. *Journal of the American Medical Association*, 307:1621–1628, 2012.

[56] J. Jiang. Conditional inference about generalized linear mixed models. *Annals of Statistics*, 27:1974–2008, 1999.

[57] S.C. Johnston, J.D. Rootenberg, S. Katrak, W.S. Smith, and J.S. Elkins. Effect of a us national institutes of health programme of clinical trials on public health and costs. *Lancet*, 367:1319–1327, 2006.

[58] O.H. Klungel, E.P. Martens, B.M. Psaty, D.E. Grobbee, S.D. Sullivan, B.H. Stricker, H.G. Leufkens, and A. Boer. Methods to assess intended effects of drug treatment in observational studies are reviewed. *Journal of Clinical Epidemiology*, 57:1223–1231, 2004.

[59] C.G. Koch, L. Li, D. Sessler, P. Figueroa, G.A. Hoeltge, T. Mihaljevic, and E.H. Blackstone. Duration of red-cell storage and complication after cardiac surgery. *The New England Journal of Medicine*, 358:1229–1239, 2008.

[60] M.S. Kramer and S.H. Shapiro. Scientific challenges in the application of randomized trials. *The Journal of the American Medical Association*, 252:2739–2745, 1984.

[61] L.L. Kupper, J.M. Karon, D.G. Kleinbaum, H. Morgenstern, and D.K. Lewis. Matching in epidemiologic studies: validity and efficiency considerations. *Biometrics*, 37:271–291, 1981.

[62] N.M. Laird and J.H. Ware. Random effects models for longitudinal data. *Biometrics*, 38:963–974, 1982.

[63] E.C. Lau, F.S. Mowat, M.A. Kelsh, J.C. Legg, N.M. Engel-Nitz, H.N. Watson, H.L. Collins, R.J. Nordyke, and J.L. Whyte. Use of electronic medical records (emr) for oncology outcomes research: assessing the comparability of emr information to patient registry and health claims data. *Clin Epidemiol*, 3:259–272, 2011.

[64] B.K. Lee, J. Lessler, and E.A. Stuart. Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29:33–346, 2010.

[65] Y. Lee and J.A. Nelder. Hierarchical generalized linear models (with discussion). *J. Roy. Statist. Soc. Ser. B*, 58:619–678, 1996.

[66] K.Y. Liang and S. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

[67] D.Y. Lin, B.M. Psaty, and R.A. Kronmal. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, 54:948–963, 1998.

[68] X. Lin and N.E. Breslow. Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, 91:1007–1016, 1996.

[69] M.L. Lindstrom and D.M. Bates. Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of American Statistical Association*, 83(404):1014–1021, 1988.

[70] J. Liu and R. Chen. Sequential monte-carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93(443):1032–1044, 1998.

[71] J.K. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, 23:2937–2960, 2004.

[72] M. Lunt, D. Solomon, K. Rothman, R. Glynn, K. Hyrich, D.P.M. Symmons, and T. Sturmer. Different methods of balancing covariates leading to different effect estimates in the presence of effect modification. *American Journal of Epidemiology*, 169(7):909–917, 2009.

[73] N. Mantel and W. Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst*, 22(4):719–728, 1959.

[74] D.F. McCaffrey, G. Ridgeway, and A.R. Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9:403–425, 2004.

[75] M. McClellan, B.J. McNeil, and J.P. Newhouse. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? analysis using instrumental variables. *Journal of American Medical Association*, 272:859–866, 1994.

[76] P. McCullagh. Re-sampling and exchangeable arrays. *Bernoulli*, 6:285–301, 2000.

[77] C.E. McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of Am. Statist. Ass.*, 92:162–170, 1997.

[78] C.E. McCulloch and S.R. Searle. *Generalized, Linear, and Mixed Models*. Wiley, New York, 2001.

[79] S.M. McKinlay. Pair-matching-a reappraisal of a popular technique. *Biometrics*, 33:725–735, 1977.

[80] O.S. Miettinen. The matched pairs design in the case of all-or-none responses. *Biometrics*, 24:339–352, 1968.

[81] O.S. Miettinen. Stratification by a multivariate confounder score. *American Journal of Epidemiology*, 104:609–620, 1976.

[82] E. Miguel, S. Satyanath, and E. Sergenti. Economic shocks and civil conflict: An instrumental variable approach. *Journal of Political Economy*, 112:725–753, 2004.

[83] M. Mitka. Us government kicks off program for comparative effectiveness research. *The Journal of the American Medical Association*, 304(20):2230–2231, 2010.

[84] J. Monaco, J. Cai, and J. Grizzle. Bootstrap analysis of multivariate failure time data. *Statistics in Medicine*, 24(22):3387–3400, 2005.

[85] Y. Mundlak. On the pooling of time series and cross section data. *Econometrica*, 46(1):69–85, 1978.

[86] K.M. Murphy and R.H. Topel. Estimation and inference in two-step econometric models. *Journal of Business and Economic Statistics*, 3:370–379, 1985.

[87] J. Neily, P.D. Mills, Y.Y. Xu, B.T. Carney, P. West, D.H. Berger, L.M. Mazzia, D.E. Paull, and J.P. Bagian. Association between implementation of a medical team training program and surgical mortality. *Journal of the American Medical Association*, 304:1693–1700, 2010.

[88] J.A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.

[89] E. Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistic*, 33:1065–1076, 1962.

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

[91] T. Permutt and J.R. Hebel. Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight. *Biometrics*, 45:619–622, 1989.

[92] E. Petkova and J. Teresi. Some statistical issues in the analyses of data from longitudinal studies of elderly chronic care populations. *Psychosomatic Medicine*, 64:531–547, 2002.

[93] S.J. Pocock and D.R. Elbourne. Randomized trials or observational tribulations? *New England Journal of Medcine*, 342(25):1907–1909, 2000.

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

[95] S. Rose and M.J. van der Laan. Why match? matched case-control studies. *Int. Journal of Biostat.*, 5(1):1–25, 2009.

[96] P.R. Rosenbaum. Propensity score. *Encyclopedia of Biostatistics*, 5:3551–3555, 1998.

[97] P.R. Rosenbaum and D.B. Rubin. The central role of the propensity scorn observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[98] P.R. Rosenbaum and D.B. Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J R Stat Soc B*, 45:212–218, 1983b.

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

[100] P.R. Rosenbaum and D.B. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39:33–38, 1985.

[101] M. Rosenblatt. Remarks on some nonparametric estimates of density function. *Annals of Mathematical Statistics*, 27:832–837, 1956.

[102] M.B. Rothberg, P.S. Pekow, M. Lahti, O. Brody, D.J. Skiest, and P.K. Lindenauer. Antibiotic therapy and treatment failure in patients hospitalized for acute exacerbations of chronic obstructive pulmonary disease. *Journal of the American Medical Association*, 303:2035–2042, 2010.

[103] K.J. Rothman and S. Greenland. *Modern Epidemiology.* Lippincott-Raven, Philadelphia, 1998.

[104] D.B. Rubin. Matching to remove bias in observational studies. *Biometrics*, 29:159–183, 1973.

[105] D.B. Rubin. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med*, 127:757–763, 1997.

[106] R. Schall. Estimation in generalized linear models with random effects. *Biometrika*, 78:717–727, 1991.

[107] S. Schneeweiss and J. Avorn. A review of uses of health care utilization databases for epidemiological research on therapeutics. *Journal of Clinical Epidemiology*, 58:323–337, 2005.

[108] S. Schneeweiss, A.M. Walker, R.J. Glynn, M. Maclure, C. Dormuth, and S.B. Soumerai. Outcomes of reference pricing for angiotensin-converting-enzyme inhibitors. *The New England Journal of Medicine*, 346:822–829, 2002.

[109] G.E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[110] D. Scott. On optimal and data-based histograms. *Biometrika*, 66:605–610, 1979.

[111] J.D. Seeger, P.L. Williams, and A.M. Walker. An application of propensity score matching using claims data. *Pharmacoepidemiol Drug Saf.*, 14(7):465–476, 2005.

[112] S. Senn, E. Graf, and A. Caputo. Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment or exposure. *Statistics in Medicine*, 26:5529–5544, 2007.

[113] S. Setoguchi, S. Schneeweiss, M.A. Brookhart, R.J. Glynn, and E.F. Cook. Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepi-demiology and Drug Safety*, 17:546–555, 2008.

[114] W.R. Shadish, T.D. Cook, and D.T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Houghton-Mifflin, Boston, 2002.

[115] R.P. Sharpe, R. Gupta, V.H. Gracias, J.P. Pryor, F.M. Pieracci, P.M. Reilly, and C.W. Schwab. Incidence and natural history of below-knee deep venous thrombosis in high-risk trauma patients. *Journal of Trauma-Injury Infection & Critical Care*, 53(6):1048–1052, 2002.

[116] A.D. Shaw, M.S. Smith, W.D. White, B.P. Bute, M. Swaminathan, C. Milano, I.J. Welsby, S. Aronson, J.P. Mathew E.D. Peterson, and M.F. Newman. The effect of aprotinin on outcome after coronary-artery bypass grafting. *The New England Journal of Medicine*, 358:784–793, 2008.

[117] S.J. Sheather and M.C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B*, 53(3):683–690, 1991.

[118] L.B. Shepardson, S.J. Youngner, T. Speroff, and G.E. Rosenthal. Increased risk of death in patients with do-not-resuscitate orders. *Medical Care*, 37:727–737, 1999.

[119] J. Staff, M.E. Patrick, L. Eric, and J.L. Maggs. Teenage alcohol use and educational attainment. *Journal of Studies on Alcohol and Drugs*, 69:848–858, 2008.

[120] W.F. Stewart, N.R. Shah, M.J. Selna, R.A. Paulus, and J.M. Walker. Bridging the inferential gap: The electronic health record and clinical evidence. *Health Affairs*, 26(2):181–191, 2007.

[121] E.W. Steyerberg. *Clinical prediction models: A practical approach to development, validation, and updating*. Springer, New York, 2009.

[122] T. Sturmer, M. Joshi, R.J. Glynn, J. Avorn, K.J. Rothman, and S. Schneeweiss. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J. Clin. Epidemiol.*, 59(5):437–447, 2006.

[123] T. Sturmer, S. Schneeweiss, J. Avorn, and R.J. Glynn. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *American Journal of Epidemiology*, 162(3):279–289, 2005.

[124] S. Suissa and E. Garbe. Primer: administrative health databases in observational studies of drug effects - advantages and disadvantagees. *Nat Clin Pract Rheumatol*, 12:725–732, 2007.

[125] D.B. Suits. Use of dummy variables in regression equations. *Journal of the American Statistical Association*, 52(280):548–551, 1957.

[126] L.M. Sullivan, K.A. Dukes, and E. Losina. Tutorial in biostatistics: An introduction to hierarchical linear modelling. *Statistics In Medicine*, 18:855–888, 1999.

[127] M. Tanasescu, M.F. Leitzmann, E.B. Rimm, W.C. Willett, M.J. Stampfer, and F.B. Hu. Exercise type and intensity in relation to coronary heart disease in men. *Journal of the American Medical Association*, 288:1994–2000, 2002.

[128] R. Tannen, M. Weiner, and D. Xie. Replicated studies of two randomized trials of angiotensincoverting enzyme inhibitors: further empiric validation of the 'prior event rate ratio' to adjust for unmeasured confounding by indication. *Pharmacoepidemiology and Drug Safety*, 17:671–685, 2008.

[129] F.J. Thoemmes and S.G. West. The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research*, 46:514–543, 2011.

[130] M.E. Tinetti and S.A. Studenski. Comparative effectiveness research and patients with multiple chronic conditions. *New England Journal of Medcine*, 364:2478–2481, 2011.

[131] O.C. Ukoumunne, A.C. Davison, M.C. Gulliford, and S. Chinn. Non-parametric bootstrap confidence intervals for the intraclass correlation coefficient. *Statis. Med.*, 22:3805–3821, 2003.

[132] F. Vaida and R. Xu. Proportional hazards model with random effects. *Statistics in Medicine*, 19:3309–3324, 2000.

[133] S. Vansteelandt, J. Bowden, M. Babanezhad, and E. Goetghebeur. On instrumental variables estimation of causal odds ratios. *Statistical Science*, 26(3):403–422, 2011.

[134] G. Verbeke and E. Lesaffre. A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91(433):217–221, 1996.

[135] S. Wacholder, D.T. Silverman, and J.K. McLaughlin. Selection of controls in case-control studies. iii. design options. *Am J Epidemiol*, 135:1042–1050, 1992.

[136] G. Wahba. Optimal convergence properties of variable knot, kernel, and orthogonal series methods for density estimation. *Annals of Statistics*, 3(1):15–29, 1975.

[137] J. Wang and P.T. Donnan. Propensity score methods in drug safety studies: practice, strengths, and limitations. *Pharmacoepidemiology and Drug Safety*, 10:341–344, 2001.

[138] M.G. Weiner, D. Xie, and R.L. Tannen. Replication of the scandinavian simvastatin survival study using a primary care medical record database prompted exploration of a new method to address unmeasured confounding. *Pharmacoepidemiology and Drug Safety*, 17:661–670, 2008.

[139] S. Weitzen, K.L. Lapane, A.Y. Toledano, A.L. Hume, and V. Mor. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiology and Drug Safety*, 13:841–853, 2004.

[140] D. Whitley. A genetic algorithm tutorial. *Stat. Comput.*, 4:65–85, 1994.

[141] E.J. Williamson, A. Forbe, and R. Wolfe. Doubly robust estimators of causal exposure effects with missing data in the outcome, exposure or a confounder. *Statistics in Medicine*, 31:4382–4400, 2012.

[142] H. Wunsch, W.T. Linde-Zwirble, and D.C. Angus. Methods to adjust for bias and confounding in critical care health services research involving observational data. *Journal of Critical Care*, 21:1–7, 2006.

[143] A.E. Wyse, V.A. Keesler, and B.Schneider. Assessing the effects of small school size on mathematics achievement: A propensity score-matching approach. *Teachers College Record*, 110:1879–1900, 2008.

[144] W. Yang, A. Zilov, P. Soewondo, O.M. Bech, F. Sekkal, and P.D. Home. Observational studies: going beyond the boundaries of randomized controlled trials. *Diabetes Research and Clinical Practice*, 88:S3–S9, 2010.

[145] Y. Ye and L.A. Kaskutas. Using propensity scores to adjust for selection bias when assessing the effectiveness of alcoholics anonymous in observational studies. *Drug and Alcohol Dependence*, 104:56–64, 2009.

[146] S.L. Zeger and M.R. Karim. Generalized linear models with random effects: a gibbs sampling approach. *Journal of the American Statistical Association*, 86:79–86, 1991.

[147] D. Zhang and M. Davidian. Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, 57:795–802, 2001.

[148] H. Zhu, R.H. Byrd, and J. Nocedal. Algorithm 778: L-bfgs-b, fortran routines for large scale bound constrained optimization. *ACM Transactions on Mathematical Software*, 23:550–560, 1997.