Health Services and Outcomes Research Methodology (2019) 19:36–60
https://doi.org/10.1007/s10742-019-00196-2

Model-based inference on average causal effect in observational clustered data

Meng Wu[1,2] · **Recai M. Yucel**[1,3]

Received: 2 May 2018 / Revised: 23 October 2018 / Accepted: 4 January 2019 / Published online: 16 January 2019

Abstract
We study causal inference using the framework of potential outcomes in clustered data settings where observational units are clustered in naturally occurring groups (e.g. patients within hospitals). To incorporate the correlated nature of the data, we employ mixed-effects models and a sandwich estimator to make inferences on the average causal effect (ACE). Our methods apply the concept of potential outcomes from the Rubin Causal Model (Holland in J Am Stat Assoc 81(396):945–960, 1986), and extend Schafer and Kang's methods of estimating the variance of the ACE (Schafer and Kang in Psychol Methods 13(4):279–313, 2008). Particularly, we develop two model-based approaches to estimate the ACE and its variance under a dual-modeling strategy which adjusts for the confounding effect through inverse probability weighting. These two approaches use linear mixed-effects models for the estimation of potential outcomes, but differ in how clustering is handled in the treatment assignment model. We present a summary of our comprehensive simulation study assessing the repetitive sampling properties of the two approaches in a pseudo-random simulation environment. Finally, we report our findings from an application to study the ACE of inadequate prenatal care on birth weight among low-income women in New York State.

Keywords ACE · Causal inference · Clustered data · Dual-modeling · Linear mixed-effects model · Potential outcomes · Sandwich estimator

1 Introduction

Drawing causal inference from well-designed randomized trials is an intuitive process. For example, given a dichotomous treatment assignment with perfect compliance, a causal effect can be estimated by direct comparison of the outcomes between the treatment and control groups. Random allocation of subjects to the treatment group implies

✉ Recai M. Yucel
ryucel@albany.edu

[1] Department of Epidemiology and Biostatistics, University at Albany, SUNY, Albany, USA

[2] Office of Quality and Patient Safety, New York State Department of Health, Albany, USA

[3] School of Public Health, State University of New York at Albany, One University Place, Room 139, Corning Tower Room 287, Rensselaer, NY 12144-3456, USA

exchangeability of the treatment and control participants, i.e., the distribution of potential outcomes under each treatment level would be unchanged if all participants' assigned treatments were exchanged (Hernán 2004). Therefore, a difference in the estimated average outcomes can be causally attributed to the treatment, allowing for evaluation of an average effect of treatment (vs control) in randomized trials (Rubin 1974, 1990). When the ideal situation of randomization is disturbed by loss to follow-up or non-compliance, researchers often rely on additional statistical approaches in order to obtain consistent or unbiased estimates. Examples of such approaches include methods for estimating the complier average causal effect (Frangakis and Rubin 1999; Berg et al. 2017; Chan 2014; Cheng 2009; Connell 2009; Gruber et al. 2014), inverse probability censoring weighting (Robins and Finkelstein 2000), and latent ignorability modeling (Frangakis and Rubin 1999; Zhou and Li 2006; Taylor and Zhou 2009).

Although randomized experiments are widely accepted as the "gold standard" for drawing causal inference, it cannot be universally applied to every causal question due to infeasibility, high cost, or ethical issues. For example, under-representation of target populations may pose a challenge for generalization even though the experiment is randomized. When such obstacles exist in studying the causal effect of interest, observational data are typically used in the absence of randomization. Observational data have been historically used in association studies; however, the statistical literature for causal methods has so far mostly focused on methods for randomized experiments. Particular areas of research pertain to methods using propensity scores (Rosenbaum and Rubin 1983; Rubin 2001, 2004; Ho et al. 2007; Pearl 2009; Austin et al. 2005, 2007; Austin 2009, 2011; Austin and Stuart 2015), principal stratification (Frangakis and Rubin 2002; Gallop et al. 2009; Elliott et al. 2010; Pearl 2011), marginal structural models, and g-estimation (Robins 1999; Robins et al. 2000; Hernán et al. 2001, 2002; Robins et al. 2015).

Our work aims to fill the gap between causal methods and their application in surveys and/or administrative data where observational units are correlated. We are particularly interested in the propensity score method due to its wide applicability in social and medical science. This method is based on the counterfactual framework that was originally proposed by Neyman (1923). Rubin (1974, 2006) and other researchers, extended it to a general framework with implication for both experimental and observational studies. In the literature, it is also referred to as the Neyman–Rubin Causal Model, and more frequently called the Rubin Causal Model (RCM) (Holland 1986). Under RCM, each subject is assumed to have two potential outcomes given a dichotomous treatment assignment. One is potentially realized under treatment and the other one is potentially realized under control. An individual causal effect is then defined as the difference between these two potential outcomes. Obviously, however, this individual causal effect can never be identified because only one potential outcome can be observed at a time. This missing data problem is the fundamental issue of causal inference (Holland 1986). With some assumptions, individual causal effect can be predicted by the average causal effect (ACE) which can be estimated at the population level.

The role of propensity scores under the RCM framework is to control confounding factors associated with both the treatment assignment and outcome (Rosenbaum and Rubin 1983). Unbalanced confounding factors in observational data can contribute to bias in the estimation of average treatment effects. Without the correct treatment assignment mechanism to adjust for confounding, valid causal effect cannot be estimated. Classical propensity score strategies include stratified analysis and propensity score matching. The performance of these strategies was recently compared and discussed in a longitudinal study (Leon et al. 2012a, b). A class of semi-parametric methods based on propensity scores

have been increasingly used. The well-known inverse probability weighting (IPW) method originally introduced to handle missing data problems has been extensively applied within a dual-modeling framework to pursue causal inference (Robins et al. 1995, 2000). This strategy has a double robustness property which means that asymptotic bias due to model misspecification of either the regression or propensity-score model is eliminated as long as one of the two models is correctly specified (Robins et al. 1995). However, this asymptotic unbiasedness does not hold if both models are inaccurately specified. A comprehensive review discussion on this matter is given by Schafer and Kang (2008) who also developed a sandwich estimator for estimating the variance of an ACE estimator and discussed its properties in relation to other previous methods .

We extend computational methods by Schafer and Kang (2008) to observational data obtained either from surveys or administrative data with correlated observational units. For example, Medicaid claim data is the reimbursement system for the services provided to the program enrollees. Researchers using this type of data for statistical analysis can often benefit from rich information. As an example, birth certificate data contain mothers' clinical conditions in addition to standard birth outcomes. Therefore various risk factors associated with the birth outcomes can be analyzed. Administrative or survey data can further provide advantages to statistical analysis in terms of cost effectiveness and time efficiency. Existing administrative data do not add costs to data collection, and survey data can be administered in a short time period due to advancements in technology (e.g. web surveys).

Administrative data typically present additional statistical complexity as they are not collected for statistical purposes. This complexity typically increases when records are correlated. Examples include observational units nested within geographical locations such as state or county; or patients nested within service providers, such as hospitals and outpatient clinics. Ignoring such clustering effect can be detrimental to the statistical inference in the sense that variances can be underestimated, leading to erroneous inferences. Our work focuses on inferences for ACE in surveys or administrative data with correlated observational units. In Sect. 2, we provide definitions and assumptions for causal inference in clustered data. In Sects. 4 and 5, we then develop methods for computing ACE and its variance in settings differentiated by inclusion of random effects in treatment assignment. We discuss our comprehensive simulation study on the proposed techniques in Sect. 5. In Sect. 6, we present an application of these methods to study the ACE of inadequate prenatal care on birth weight. Finally, we discuss the limitations of the proposed method and future work in Sect. 7.

2 Causal inference in clustered data

2.1 Potential outcomes framework

Let t_{ij} denote the treatment assignment indicator for the jth observational unit within the ith cluster ($i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n_i$):

$$t_{ij} = \begin{cases} 1 & \text{treatment group} \\ 0 & \text{control group} \end{cases}$$

Further, we let x_{ij} denote the observed covariates and y_{ij} denote the observed outcome measurement. We define potential outcomes Y_1 and Y_0 as the outcome measurement that would have been observed under treatment or control. Our definition of potential outcomes

are non-deterministic because the underlying cluster may have a dynamic confounding effect on the outcomes (Gitelman 2005). Rather than specifying a fixed value to the potential outcome (the deterministic potential outcome) for each unit, we define the potential outcomes Y_1 and Y_0 as random variables. The choice of non-deterministic potential outcomes does not impact the definition of ACE and its point estimation (Hernán 2004). One can estimate the ACE by

$$ACE = E(Y_1) - E(Y_0) \tag{1}$$

We consider a linear mixed-effects model with random intercepts to model the outcomes. This model is written as

$$y_{ij} = \alpha_i + x_{ij}^T \beta + \epsilon_{ij}, \tag{2}$$

where α_i is the random intercept and assumed to be distributed as N(0, σ_α^2). To estimate the ACE, we chose to fit the model separately for each treatment group for a given set of covariates (obviously, excluding the treatment assignment) specified for the units in the relevant groups:

$$y_{ij0} = \alpha_{i0} + x_{ij0}^T \beta_0 + \epsilon_{ij0}, \quad y_{ij1} = \alpha_{i1} + x_{ij1}^T \beta_1 + \epsilon_{ij1} \tag{3}$$

The parameters estimated in these two separate models are then used to predict potential outcomes y_{ij1} and y_{ij0} for all observational units.

To make the causal inference from this framework, some additional assumptions are needed. The first assumption is exchangeability, which states that the groups are comparable and that the potential outcomes are independent of treatment assignment (Hernán 2004). This assumption is sometimes called ignorability of treatment assignment. Exchangeability holds unconditionally if treatment is randomized, but generally does not hold in observational data due to confounding factors that can not be balanced. To overcome this, an alternative assumption of conditional exchangeability is proposed (Hernán and Robins 2006). Conditional exchangeability allows outcome to be independent of treatment assignment conditional on confounding variables, and treatment assignment mechanism would then need to be modeled. The second assumption is the positivity assumption, which states that there is a positive probability of receiving every level of treatment for every combination of values of the treatment and covariates (Hernán and Robins 2006; Westreich and Stephen 2010). Under the potential outcomes framework, we also need to make the "stable unit treatment value assumption (SUTVA)" (Rubin 1974), which states that (a) the treatment effect is the same for all the units, and (b) the potential outcomes of a unit are not affected by the treatment assignment of other units such that there is no interference.

2.2 Inverse probability weighting

Propensity scores, the probability of treatment assignment, have been commonly used to adjust for confounding in non-randomized studies. The effect of confounders on outcomes and treatments can be minimized through propensity score matching. Propensity scores can also be utilized to post-adjust outcome estimates through inverse probability weighting (IPW), which weighs subjects by inverse probability of selection. IPW was originally proposed for missing data problem. Analysis is conducted on complete data, and each unit is weighted by its inverse probability of being complete. For example, one can weigh subjects in a randomized study due to non-compliance in treatment (Robins et al. 2000) or survival data with censoring (Cain

and Cole 2009). A recent review of this method for dealing with missing data was given by Seaman and White (2013).

For causal inference, dual-modeling IPW estimators are frequently used to obtain unbiased estimate of treatment effects when there are unbalanced covariates between the treatment groups. One particular choice is residual weighting (Robins et al. 1995; Schafer and Kang 2008). It adds a bias-corrected term to the estimate of average potential outcomes. In a clustered data setting, one can estimate the average potential outcome under treatment by

$$\widehat{E(Y_1)} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}\left(x_{ij}^t\hat{\beta}_1 + \hat{\alpha}_{i1}\right) + \frac{\sum_i^m\sum_j^{n_i} t_{ij}\hat{\pi}_{ij}^{-1}\left(y_{ij} - x_{ij}^t\hat{\beta}_1 - \hat{\alpha}_{i1}\right)}{\sum_i^m\sum_j^{n_i} t_{ij}\hat{\pi}_{ij}^{-1}},$$

where $\hat{\beta}_1$ and $\hat{\alpha}_{i1}$ are the estimates from the potential outcome model fitted for units observed in the treatment group, and $\hat{\pi}_{ij}$ is the estimated propensity score for jth observational unit in cluster i. Weighted residuals in this equation are used to approximate the residuals for all the subjects in the sample because residuals cannot be computed for unobserved potential outcomes. The average of these weighted residuals provides a bias correction to the estimation of $E(Y_1)$. If the potential outcome model is correctly specified, an unbiased estimate can be obtained because the average weighted residual is zero regardless of whether the treatment model is correct. If the potential outcome model is incorrectly specified, the average weighted residual will not be zero, and thus adds a bias corrected term to the estimate. With this correction, we can still obtain an accurate estimate if the propensity score model is correctly specified. This is the well-known double robust property, i.e., the estimate for $E(Y_1)$ is consistent if either the model for the treatment assignment or the model for the potential outcomes is correctly specified. Similarly, the estimate of average potential outcome under control is

$$\widehat{E(Y_0)} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}\left(x_{ij}^t\hat{\beta}_0 + \hat{\alpha}_{i0}\right) + \frac{\sum_i^m\sum_j^{n_i}(1-t_{ij})(1-\hat{\pi}_{ij})^{-1}\left(y_{ij} - x_{ij}^t\hat{\beta}_0 - \hat{\alpha}_{i0}\right)}{\sum_i^m\sum_j^{n_i}(1-t_{ij})(1-\hat{\pi}_{ij})^{-1}}.$$

We consider two propensity score models for clustered data. If the probability of treatment assignment is independent of cluster-level characteristics, whether observable or not (i.e., random-effects), a standard logistic regression can be fit to estimate the propensity scores. Let π_{ij} denote the probability of treatment assignment; z_{ij} denote the covariates in the logistic regression to distinguish from the notation in the linear mixed-effects model; and γ denote the regression coefficients. Then the propensity scores or treatment assignment probabilities are estimated by:

$$\pi_{ij} = \left(1 + exp\left(-z_{ij}^T\gamma\right)\right)^{-1}, \tag{4}$$

where z_{ij} are the covariates that are associated with treatment assignment. If the treatment assignment for the subjects within the same cluster are correlated with each other due to cluster-level covariates, allocation of the clusters will have a confounding effect on the treatment assignment. In this paper, we focus on the situation where these cluster-level characteristics are not measured, and we consequently fit a random intercept logistic regression model to incorporate the clustering effect:

$$\pi_{ij} = \left(1 + exp\left(-z_{ij}^T\gamma - \zeta_i\right)\right)^{-1}, \tag{5}$$

where ζ_i is the random intercept assumed to be distributed as $N(0, \sigma^2)$ independently across the clusters $i = 1, 2, \ldots, m$.

Propensity scores predicted from these two models may have extreme values due to outliers in confounders. Unusually large or small weights generated from these extreme values can impact the performance of the estimation. When both models are misspecified, even moderately, point estimate and its corresponding standard error can be significantly biased (Kang and Schafer 2007). In such situations, other methods that do not depend on propensity scores such as multiple imputations could be a preferable choice (Rubin 2004; Westreich et al. 2015). Therefore valid causal inference requires careful selection of confounders, which is a crucial process and often requires expert knowledge. Procedures and strategies on model selection under misspecified dual models have been discussed by several authors in recent literature (Vansteelandt et al. 2012; Waernbaum 2012; Gruber and Van Der Laan 2015; Schnitzer et al. 2016).

2.3 Estimation of ACE variance

Schafer and Kang (2008) developed a method to estimate the ACE variance for non-clustered settings by expressing ACE as a linear equation of a set of parameters. Consider a simple data setting without clusters, let Y_1 denote the potential outcome under treatment and Y_0 denote the potential outcome under control. ACE can be estimated by $E(Y_1) - E(Y_0)$. To illustrate the method, a linear model for the potential outcomes is assumed. $E(Y_1)$ and $E(Y_0)$ are obtained through ordinary least squares (OLS) estimation and are not corrected by residual-weighting. Let $\mu_1 = E(Y_1)$ and $\mu_0 = E(Y_0)$, ACE can then be expressed as $ACE = \mu_1 - \mu_0 = a^T\theta$, where $a = (0, 0, -1, 1)^T$ and $\theta = (\beta_0, \beta_1, \mu_0, \mu_1)^T$. The OLS estimates $\hat{\theta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\mu}_0, \hat{\mu}_1)$ from the linear regression can be treated as a solution of a set of joint estimation equations $\sum_{i=1}^{N} \varphi_i(\theta) = 0$, where $\varphi_i(\theta) = (\varphi_{i\beta_0}, \varphi_{i\beta_1}, \varphi_{i\mu_0}, \varphi_{i\mu_1})$, $\varphi_{i\beta_0} = (1 - t_i)(y_i - x_i^T\beta_0)$, $\varphi_{i\beta_1} = t_i(y_i - x_i^T\beta_1)$, $\varphi_{i\mu_0} = (1 - t_i)(y_i - \mu_0) + t_i(x_i^T\beta_0 - \mu_0)$, and $\varphi_{i\mu_1} = t_i(y_i - \mu_1) + (1 - t_i)(x_i^T\beta_1 - \mu_1)$. By the central limiting theory and Taylor approximation,

$$\hat{\theta} \approx N\left(\theta, \hat{J}(\phi(\hat{\theta}))^{-1} V(\phi(\hat{\theta}))\left(\hat{J}(\phi(\hat{\theta}))^{-1}\right)^T\right), \tag{6}$$

where $\hat{J}(\varphi(\hat{\theta})) = E\left(\frac{\partial\varphi(\theta)}{\partial\theta}\right)$, $V(\varphi(\hat{\theta})) = E(\varphi(\theta)\varphi(\theta)^T)$. Variance of $\hat{ACE}$ can then be approximated by:

$$\widehat{V(\hat{ACE})} \approx \frac{1}{N} a^T A^{-1} \hat{B} (A^{-1})^T a \tag{7}$$

where $B = E(\varphi\varphi^T), A = \hat{J}(\varphi(\theta))$.

Building upon the methods described by Schafer and Kang (2008), we propose two new approaches for estimation of ACE and its variance in clustered data with particular emphasis on administrative and/or survey data. Both approaches estimate ACE by random intercept only regression under the framework of potential outcomes. A dual-modeling strategy is employed to adjust the estimates with the inverse probability of treatment assignment. The first approach, referred as Method 1, ignores the clustering effect on treatment

assignment and uses a standard logistic regression to model the probability of treatment assignment. The second approach, referred as Method 2, incorporates the clustering effect on the treatment assignment using random-effects in the underlying logistic regression model.

3 Method 1: ACE estimation under linear mixed-effects model and IPW by standard logistic regression

As described in the definition of potential outcomes for clustered data, we assume a linear mixed-effects model as the underlying potential outcome model. Coefficients in Eq. (3) are estimated using data from the control group and treatment group, respectively. Under the dual-modeling strategy, residuals are weighted by the inverse probabilities of treatment assignment estimated from the standard logistic model given in Eq. (4). ACE is then estimated as follows:

$$\hat{ACE} = \widehat{E(Y_1)} - \widehat{E(Y_0)} = \hat{\mu}_1 - \hat{\mu}_0 + \hat{\alpha}_1 - \hat{\alpha}_0 + \hat{\epsilon}_1 - \hat{\epsilon}_0 \tag{8}$$

where

$$\begin{aligned}
\hat{\mu}_1 &= \frac{1}{N}\sum_i^m \sum_j^{n_i} \left\{ x_{ij}^T \hat{\beta}_1 \right\}, \\
\hat{\mu}_0 &= \frac{1}{N}\sum_i^m \sum_j^{n_i} \left\{ x_{ij}^T \hat{\beta}_0 \right\}, \\
\hat{\alpha}_1 &= \frac{1}{N}\sum_i^m \sum_j^{n_i} \left\{ \hat{\alpha}_{i1} \right\}, \\
\hat{\alpha}_0 &= \frac{1}{N}\sum_i^m \sum_j^{n_i} \left\{ \hat{\alpha}_{i0} \right\}, \\
\hat{\epsilon}_1 &= \frac{\sum_i^m \sum_j^{n_i} t_{ij} \hat{\pi}_{ij}^{-1} \left(y_{ij} - x_{ij}^T \hat{\beta}_1 - \hat{\alpha}_{i1} \right)}{\sum_i^m \sum_j^{n_i} t_{ij} \hat{\pi}_{ij}^{-1}}, \\
\hat{\epsilon}_0 &= \frac{\sum_i^m \sum_j^{n_i} (1 - t_{ij})(1 - \hat{\pi}_{ij})^{-1} \left(y_{ij} - x_{ij}^T \hat{\beta}_0 - \hat{\alpha}_{i0} \right)}{\sum_i^m \sum_j^{n_i} (1 - t_{ij})(1 - \hat{\pi}_{ij})^{-1}}.
\end{aligned}$$

Note that

$$\hat{ACE} = \hat{\mu}_1 - \hat{\mu}_0 + \hat{\alpha}_1 - \hat{\alpha}_0 + \hat{\epsilon}_1 - \hat{\epsilon}_0 = a^T \hat{\theta} \tag{9}$$

where $a^T = (0, 0, 0, -1, 1, -1, 1, -1, 1)$ and $\hat{\theta} = (\hat{\gamma}, \hat{\beta}_0, \hat{\beta}_1, \hat{\mu}_0, \hat{\mu}_1, \hat{\alpha}_0, \hat{\alpha}_1, \hat{\epsilon}_0, \hat{\epsilon}_1)^T$. $\hat{\theta}$ can be thought as the solution of a set of joint estimation equations $\sum_i^m \sum_j^{n_i} \varphi_{ij}(\theta) = 0$, where

$$
\begin{aligned}
\varphi_{ij\gamma} &= (t_{ij} - \pi_{ij}) z_{ij},\\
\varphi_{ij\beta_0} &= (1 - t_{ij}) x_{ij}^T \left(\bar{y}_{i0} - x_{ij0}^T \beta_0\right),\\
\varphi_{ij\beta_1} &= t_{ij} x_{ij}^T \left(\bar{y}_{i1} - x_{ij1}^T \beta_1\right),\\
\varphi_{ij\mu_0} &= x_{ij0}^T \beta_0 - \mu_0,\\
\varphi_{ij\mu_1} &= x_{ij1}^T \beta_1 - \mu_1,\\
\varphi_{ij\alpha_0} &= \bar{y}_{ij0} - \bar{x}_{i0}^T \beta_0 - \alpha_0,\\
\varphi_{ij\alpha_1} &= \bar{y}_{ij1} - \bar{x}_{i1}^T \beta_1 - \alpha_1,\\
\varphi_{ij\epsilon_0} &= (1 - t_{ij})(1 - \pi_{ij})^{-1} \left(y_{ij} - x_{ij}^T \beta_0 - \alpha_{i0} - \epsilon_0\right),\\
\varphi_{ij\epsilon_1} &= t_{ij} \pi_{ij}^{-1} \left(y_{ij} - x_{ij}^T \beta_1 - \alpha_{i1} - \epsilon_1\right).
\end{aligned}
$$

Then the variance of $\hat{ACE}$ can be estimated using Eq. (7).

4 Method 2: ACE estimation under linear mixed-effects model and IPW by mixed-effects logistic regression

Here we explicitly allow random perturbations in the intercept around the population average intercept term in the probability of treatment assignment as described in Eq. (5). The formula for $\hat{ACE}$ is the same as Eq. (8), but with different parameters in the re-written equation $\hat{ACE} = a^T \theta \ : \ a^T = (0, 0, 0, 0, -1, 1, -1, 1, -1, 1)$ and $\hat{\theta} = (\hat{\gamma}, \hat{\sigma}^2, \hat{\beta}_0, \hat{\beta}_1, \hat{\mu}_0, \hat{\mu}_1, \hat{\alpha}_0, \hat{\alpha}_1, \hat{\epsilon}_0, \hat{\epsilon}_1)^T$, where the added parameters σ^2 is from treatment assignment model accounting for cluster-specific effects. Accordingly, the φ is expressed as:

$$
\begin{aligned}
\varphi_{ij\gamma} &= t_{ij} z_{ij} - \frac{1}{n_i} \frac{\int_{-\infty}^{\infty} \sum_{j}^{n_i} z_{ij} \frac{e^{z_{ij}^T \gamma + \zeta_i}}{1 + e^{z_{ij}^T \gamma + \zeta_i}} e^{h_i(\gamma, \zeta)} d\zeta}{\int_{-\infty}^{\infty} e^{h_i(\gamma, \zeta)} d\zeta},\\
\varphi_{ij\sigma^2} &= -\frac{1}{2\sigma^2 n_i} + \frac{1}{2\sigma^4 n_i} \frac{\int_{-\infty}^{\infty} \alpha^2 e^{h_i(\gamma, \zeta)} d\zeta}{\int_{-\infty}^{\infty} e^{h_i(\gamma, \zeta)} d\zeta},\\
\varphi_{ij\beta_0} &= (1 - t_{ij}) x_{ij}^T \left(\bar{y}_{i0} - x_{ij0}^T \beta_0\right)\\
\varphi_{ij\beta_1} &= t_{ij} x_{ij}^T \left(\bar{y}_{i1} - x_{ij1}^T \beta_1\right), \varphi_{ij\mu_0} = x_{ij0}^T \beta_0 - \mu_0, \varphi_{ij\mu_1} = x_{ij1}^T \beta_1 - \mu_1,\\
\varphi_{ij\alpha_0} &= \bar{y}_{ij0} - \bar{x}_{i0}^T \hat{\beta}_0 - \alpha_0, \varphi_{ij\alpha_1} = \bar{y}_{ij1} - \bar{x}_{i1}^T \beta_1 - \alpha_1,\\
\varphi_{ij\epsilon_0} &= (1 - t_{ij})(1 - \pi_{ij})^{-1} \left(y_{ij} - x_{ij}^T \beta_0 - \alpha_{i0} - \epsilon_0\right),\\
\varphi_{ij\epsilon_1} &= t_{ij} \pi_{ij}^{-1} \left(y_{ij} - x_{ij}^T \beta_1 - \alpha_{i1} - \epsilon_1\right).
\end{aligned}
$$

These estimation functions are derived from the OLS estimates for the outcome models (i.e., random intercept linear model), and the maximum likelihood estimates for the treatment assignment model (i.e., random intercept logistic regression model). To obtain the estimate of variance, we approximate the integrals using numeric approaches (Demidenko 2004). Computational details are described in the "Appendix". ACE variance is estimated by Eq. (7).

5 Simulation study

To assess the operational performance of our two methods in terms of repetitive sampling characteristics, we conduct a simulation study under three scenarios. We first assess the impact of sample size on the estimation performance. The second scenario aims to assess the impact of the magnitude of the clustering effect on the estimates, and the third scenario investigates sensitivity to skewed covariates.

5.1 Data generation

We first simulate a set of three covariates from independent normal distributions with varying means and variances: $x_1 \sim N(3,4)$, $x_2 \sim N(4,4)$ and $x_3 \sim N(12,4)$. Next, treatment t_{ij} is simulated from a binomial distribution with probability

$$\pi_{ij} = (1 + exp(-(-3 + x_{1ij} + 3x_{2ij} - x_{3ij} + \zeta_i)))^{-1},$$

where ζ_i is a cluster-specific random intercept assumed to follow a normal distribution $N(0, \sigma^2)$, independently across the clusters $i = 1, 2, \ldots, m$. Based on the simulated covariates x_1, x_2 and x_3 as well as the treatment assignment t_{ij}, the outcome variable y_{ij} is then computed under the following linear mixed-effects model:

$$y_{ij} = \alpha_i + 18 + 2x_{1ij} + 3x_{2ij} + 0.8x_{3ij} + 4t_{ij} + \epsilon_{ij},$$

where ϵ_{ij} and α_i refer to the residual error term and cluster-specific random intercepts, respectively. They are further assumed to be independent and normally distributed:

$$\epsilon_{ij} \sim N(0,1),$$
$$\alpha_i \sim N(0, \sigma_1^2).$$

This framework is used to simulate a population of 3,000,000 observational units that are grouped under 3000 clusters. The true ACE is 4 as set by the coefficient of the treatment assignment in the model above. In the first scenario, three populations are simulated based on a large ICC values of 0.6 for the outcome y_{ij} (ICC1) and three ICC values (0.08, 0.15 and 0.3) for the treatment assignment t_{ij}(ICC2). ICC values are used to determine the specific values for σ^2 and σ_1^2 in the simulation. Sampling from each population is repeated 100 times for varying sample size. We chose a range of 30–150 clusters and 40–100 subjects within each cluster.

In the second simulation experiment, we simulate the population data using the same models as the first experiment, but with different values of ICC. Specifically, we select ICC1 varying between 0.12 and 0.85 for the outcome and ICC2 varying from 0.08 to 0.3 for the treatment assignment. To avoid negative impact due to small sample size, we set a relatively large sample size with 100 clusters and 60 units within each cluster. We then

Table 1 Performance of methods for estimating ACE and standard error from data with various sample size: average bias (Bias), root-mean-square error (RMSE), standard deviation of ACE estimates (SD), average of SE estimates (SE), percent coverage rate of nominal 95% confidence intervals (CR). ICC1 = 0.59, ICC2 = 0.3

n_i	Method 1. IPW by logistic model ($ACE = 4.0$)					Method 2. IPW by mixed logistic model ($ACE = 4.0$)				
	Bias	*RMSE*	*SD*	*SE*	*CR*	*Bias*	*RMSE*	*SD*	*SE*	*CR*
$m = 150$										
100	2.67	2.99	1.35	1.28	33	1.57	2.88	2.41	2.98	92
90	2.57	3.05	1.65	1.37	38	1.77	2.24	1.37	2.46	87
80	2.39	3.42	2.45	1.44	36	1.53	2.81	2.74	3.14	85
70	2.85	3.39	1.84	1.34	32	1.57	3.05	2.62	2.89	88
60	2.68	3.55	2.32	1.48	28	1.79	2.37	1.56	2.33	82
50	2.88	3.6	2.16	1.54	35	1.87	2.82	2.12	2.99	90
40	2.92	3.63	2.15	1.69	36	2.29	2.74	1.51	7.39	86
$m = 100$										
100	2.48	3.3	2.17	1.53	38	1.96	2.44	1.46	3.82	95
90	2.45	3.11	1.92	1.57	44	2.1	2.84	1.92	3.16	88
80	2.84	3.42	1.9	1.54	33	1.93	2.59	1.72	3.89	95
70	2.8	3.27	1.7	1.62	41	1.92	2.68	1.86	4.15	95
60	2.52	3.49	2.42	1.75	43	2.14	2.71	1.66	2.91	89
50	2.84	3.54	2.12	1.82	46	2.25	2.63	1.36	4.23	93
40	3.1	3.84	2.27	1.87	39	2.48	2.8	1.29	6.87	90
$m = 50$										
100	2.61	3.3	2.02	1.87	50	2.06	2.89	2.02	6.26	97
90	2.86	3.78	2.47	1.8	42	2.46	3.15	1.96	5.78	93
80	3.04	3.41	1.55	1.88	44	2.21	2.68	1.53	6.88	98
70	3.14	3.69	1.94	1.89	42	2.28	2.88	1.75	8.24	98
60	2.95	4.13	2.89	2.06	46	2.52	3.11	1.83	4.94	94
50	3.33	3.86	1.94	2.22	48	2.16	2.75	1.7	7.92	98
40	3.18	3.95	2.34	2.44	56	2.33	3.15	2.11	9.25	95
$m = 30$										
100	3.5	3.84	1.58	1.81	37	2.27	2.82	1.67	10.11	98
90	3.12	3.64	1.88	2.04	44	1.82	2.98	2.36	10.23	98
80	3.54	3.95	1.76	1.9	40	2.26	2.97	1.93	11.04	99
70	3.22	3.83	2.06	2.2	46	2.9	3.22	1.39	12.05	99
60	3.93	4.1	1.17	2.02	41	2.18	3.3	2.48	8.05	97
50	3.5	3.96	1.86	2.44	52	2.52	3.17	1.92	11.67	98
40	3.37	4.03	2.21	2.68	62	2.8	3.3	1.75	16.4	100

study the performance criteria stated below across 100 repetition to assess the sampling properties of our methods.

In the third simulation experiment, the impact of asymmetrical distributions on the covariates is studied. We have the same distributional specification for x_1, and replace x_2 and x_3 with χ^2_{10} and log normal distribution $lnN(2, 0.36)$, respectively. Similar to the first two scenarios, we set a similar sample size of 100 clusters with 60 subjects within each

Table 2 Performance of methods for estimating ACE and standard error from data with various sample size: average bias (Bias), root-mean-square error (RMSE), standard deviation of ACE estimates (SD), average of SE estimates (SE), percent coverage rate of nominal 95% confidence intervals (CR). ICC1 = 0.59, ICC2 = 0.15

n_i	Method 1. IPW by logistic model ($ACE = 4.0$)					Method 2. IPW by mixed logistic model ($ACE = 4.0$)				
	Bias	*RMSE*	*SD*	*SE*	*CR*	*Bias*	*RMSE*	*SD*	*SE*	*CR*
$m = 150$										
100	2.58	3.58	2.48	1.21	27	1.69	2.31	1.59	4.7	99
90	2.74	3.38	1.99	1.25	20	1.83	2.33	1.44	5.57	98
80	2.97	3.28	1.38	1.19	31	1.94	2.28	1.2	10.86	99
70	2.44	3.04	1.79	1.52	40	1.89	2.64	1.84	2.64	99
60	2.57	3.87	2.89	1.44	31	1.93	2.56	1.67	10.53	92
50	2.94	3.26	1.41	1.47	34	1.62	2.93	2.44	4.58	90
40	2.87	3.47	1.94	1.63	38	1.58	2.99	2.54	1.78	51
$m = 100$										
100	2.65	3.58	2.41	1.38	33	1.99	2.83	2.01	11.44	98
90	2.85	3.43	1.92	1.46	33	2.01	2.49	1.48	7.49	98
80	3.07	3.4	1.47	1.44	32	2.05	2.52	1.47	20.4	100
70	3.07	3.42	1.5	1.48	36	2.34	2.62	1.17	19	100
60	2.56	3.53	2.43	1.74	41	1.91	2.63	1.81	16.46	98
50	3.43	3.65	1.24	1.53	31	1.84	2.79	2.1	9.45	97
40	3.34	3.93	2.06	1.69	29	1.89	2.87	2.16	2.26	74
$m = 50$										
100	2.84	3.7	2.38	1.7	36	2.19	2.72	1.62	2.72	100
90	3.34	3.83	1.86	1.63	27	1.81	2.87	2.23	16.65	99
80	3.25	3.8	1.99	1.72	36	2.49	2.81	1.29	46.67	100
70	3.48	3.83	1.6	1.74	34	2.02	3.03	2.26	32.56	99
60	3.36	4.06	2.27	1.88	35	2.4	2.83	1.5	61.6	97
50	3.22	3.74	1.9	2.16	51	2.3	3.14	2.14	30.6	97
40	3.69	4.19	1.97	2.15	40	2.26	2.78	1.62	75.3	93
$m = 30$										
100	3.1	3.95	2.45	2	43	2.12	3.01	2.14	28	100
90	2.97	3.67	2.15	2.13	54	1.99	2.98	2.21	25.79	100
80	3.28	3.76	1.83	2.19	56	1.67	2.89	2.35	18.13	100
70	3.24	3.84	2.06	2.23	50	2.51	3.36	2.23	52.51	100
60	3.41	3.96	2.01	2.42	54	2.43	3.2	2.1	31.89	97
50	3.77	4.29	2.05	2.4	43	2.65	3.43	2.18	40.82	99
40	3.85	4.24	1.78	2.62	56	2.49	3.09	1.84	130.49	97

cluster and repeatedly sample 100 times to study the sampling properties with respect to the following criteria: average bias, root-mean-square error (RMSE), standard deviation of ACE estimates, average of standard error estimates and the percent coverage rate of nominal 95% confidence intervals. Average bias is the pure accuracy measure, SD and SE are efficiency measures, and CR and RMSE are the hybrid measures (Demirtas 2007).

Table 3 Performance of methods for estimating ACE and standard error from data with various sample size: average bias (Bias), root-mean-square error (RMSE), standard deviation of ACE estimates (SD), average of SE estimates (SE), percent coverage rate of nominal 95% confidence intervals (CR). ICC1 = 0.59, ICC2 = 0.08

n_i	Method 1. IPW by logistic model ($ACE = 4.0$)					Method 2. IPW by mixed logistic model ($ACE = 4.0$)				
	Bias	*RMSE*	*SD*	*SE*	*CR*	*Bias*	*RMSE*	*SD*	*SE*	*CR*
$m = 150$										
100	2.06	3.2	2.45	1.41	41	1.56	2.38	1.79	81.5	100
90	2.16	3.44	2.66	1.44	40	1.13	2.52	2.26	6.49	96
80	2.68	3.09	1.54	1.4	39	1.99	2.42	1.38	3.89	99
70	2.79	3.18	1.53	1.39	37	1.46	2.79	2.38	3.52	95
60	2.7	3.36	1.99	1.48	39	1.76	2.74	2.1	2.02	73
50	2.73	3.37	1.98	1.61	37	1.57	2.45	1.88	2.06	78
40	2.9	3.58	2.11	1.67	36	1.74	2.78	2.17	2.45	78
$m = 100$										
100	2.81	3.34	1.8	1.4	38	1.47	2.68	2.24	19.5	100
90	2.57	3.34	2.13	1.56	39	1.91	2.82	2.08	96	100
80	2.52	3.2	1.98	1.62	46	1.49	2.83	2.4	9.4	99
70	2.8	3.32	1.77	1.63	42	1.57	2.57	2.03	10.87	97
60	2.59	3.72	2.67	1.69	37	1.98	2.43	1.41	3.23	92
50	3.01	3.73	2.2	1.68	40	1.88	2.66	1.89	2.94	83
40	2.88	3.56	2.09	1.9	46	2.19	2.67	1.53	2.87	82
$m = 50$										
100	2.78	3.57	2.24	1.72	41	2.72	2.72	1.49	9.8	99
90	2.57	3.42	2.25	1.95	51	1.67	2.55	1.92	13.25	99
80	3	3.74	2.23	1.87	40	2.32	3.07	2	154.72	100
70	2.95	3.61	2.08	2.04	40	2.08	2.79	1.86	34.04	100
60	3.59	3.93	1.59	1.9	39	2.02	3.04	2.28	43.84	98
50	3.35	3.87	1.95	2.13	42	2.29	2.8	1.6	7.24	96
40	3.39	3.97	2.07	2.32	47	1.95	2.97	2.24	5.02	92
$m = 30$										
100	3.09	3.71	2.06	1.99	48	2.53	2.98	1.56	31.66	100
90	3.38	3.77	1.66	2.04	48	2.26	3.26	2.35	44.93	100
80	3.43	3.98	2.02	2.2	50	2.31	2.94	1.82	83.33	100
70	3.31	3.9	2.07	2.3	58	NA	NA	NA	NA	NA
60	3.65	4.22	2.13	2.33	42	NA	NA	NA	NA	NA
50	3.74	4.35	2.22	2.46	45	NA	NA	NA	NA	NA
40	3.28	4.01	2.31	2.95	67	NA	NA	NA	NA	NA

In addition, we calculate the standard deviation (SD) from the 100 ACE estimates in each of the simulation scenario in Tables 1, 2, 3, 4 and 5. This is a measure to gauge the variation of the ACE estimates. Our specific goal is to see if our estimate of the ACE variance is unbiased by comparing this SD with the average of the SE estimates (the square root of the variance estimate) from the 100 samples. This indicates whether our variance estimate for ACE is unbiased or not.

Table 4 Performance of methods for estimating ACE and standard error based on various intra class correlations: average bias (AB), root-mean-square error (RMSE), standard deviation of ACE estimates (SD), average of SE estimates (SE), percent coverage rate of nominal 95% confidence intervals (CR)

ICC1	Method 1. IPW by logistic model ($ACE = 4.0$)					Method 2. IPW by mixed logistic model ($ACE = 4.0$)				
	Bias	*RMSE*	*SD*	*SE*	*CR*	*Bias*	*RMSE*	*SD*	*SE*	*CR*
$ICC_2 = 0.08$										
0.12	2.68	3.19	1.72	1.74	53	2.42	2.78	1.36	5.35	86
0.40	2.95	3.44	1.77	1.65	41	2.00	2.76	1.90	8.93	92
0.59	3.01	3.49	1.77	1.66	39	1.84	2.70	1.99	7.62	92
0.85	3.16	3.60	1.73	1.55	34	2.05	2.62	1.63	4.45	90
$ICC_2 = 0.30$										
0.12	3.13	3.40	1.32	1.45	32	1.60	2.73	2.21	2.79	85
0.40	3.52	3.68	1.06	1.34	21	2.08	2.60	1.57	2.96	87
0.59	3.52	3.68	1.06	1.35	21	2.21	2.81	1.74	3.23	87
0.85	3.26	3.59	1.52	1.42	30	1.67	2.74	2.17	3.57	86

5.2 Summary of results

Tables 1, 2 and 3 summarize the results for the first simulation scenario. Graphical visualization of the performance for the two methods is further displayed in Fig. 1. With correct specification for the outcome model, both methods can generate consistent estimates of ACE under the dual-modeling strategy. While increasing sample size can help reduce bias in both methods, a better estimate is obtained in method 2 with correct specification for the treatment assignment model. Results show that the impact of sample size on the efficiency should not be of concern because of the consistent estimates of ACE standard error in the settings with a large number of clusters (m = 150). On average, we observe low coverage rates under method 1, which indicates that ignoring clustering effect in the treatment assignment model can underestimate the variance of the ACE estimate. When the treatment assignment model accounts for the clustering effect (i.e., method 2), we observe a significant improvement on the performance (e.g., with respect to the coverage rate). However, we also note that this improvement is diminished and can produce an inaccurate high coverage when the number of clusters is small, particularly with a small ICC for the treatment assignment. Method 2 leads to subpar performance when the number of observations within a cluster drops below 70 in the settings with 30 clusters and 0.08 of ICC for the treatment assignment. Overall, the performance of method 2 becomes unstable with small ICC values for the treatment, and the variance of ACE tends to be overestimated.

For the second simulation scenario, the results for the settings with varying ICC1 and ICC2 are presented in Table 4. It appears that the change in the ICC for the outcome model has not much impact on the estimate of ACE in both methods when the outcome model is correctly specified. With the clustering effect on the treatment assignment being ignored in method 1, the smaller ICC2 (0.08) tends to result in a higher coverage rate, and the higher ICC2 (0.3) causes a notable 5–20% decline on the coverage rate. On average, we observe a lower coverage rate of 34% in method 1, which indicates ignoring the clustering effect on the treatment assignment model can underestimate the variance of ACE. Such effect should be evaluated when method 1 is applied. The average coverage rate of 88% in method 2 is

Table 5 Performance of methods for estimating ACE and standard error from samples when covariates are highly assymetric: average bias (Bias), root-mean-square error (RMSE), standard deviation of ACE estimates (SD), average of SE estimates (SE), percent coverage rate of nominal 95% confidence intervals (CR)

ICC_1	Method 1. IPW by logistic model ($ACE = 4.0$)					Method 2. IPW by mixed logistic model ($ACE = 4.0$)				
	Bias	*RMSE*	*SD*	*SE*	*CR*	*Bias*	*RMSE*	*SD*	*SE*	*CR*
$ICC_2 = 0.08$										
0.12	1.41	2.38	1.91	1.61	83	1.23	2.00	1.58	1.64	79
0.40	0.97	2.34	2.12	1.70	85	0.98	1.86	1.58	1.70	80
0.59	1.41	2.03	1.46	1.66	87	0.26	2.18	2.19	1.68	95
0.85	1.50	2.27	1.70	1.66	82	− 0.25	2.42	2.43	1.63	93
$ICC_2 = 0.30$										
0.12	1.37	1.73	1.05	1.38	82	1.22	1.79	1.31	1.40	87
0.40	1.24	1.82	1.32	1.46	85	− 0.04	2.05	2.06	1.68	92
0.59	1.48	1.83	1.09	1.37	86	− 0.06	1.49	1.49	1.47	95
0.85	1.37	2.21	1.74	1.52	82	− 0.21	2.51	2.52	1.77	94

desirable, as both models are correctly specified. However, the estimate of ACE variance is more accurate with larger ICC2 (0.3), i.e., the difference between SD and SE is relatively small.

We note that the SE is consistently underestimated in method 1 when the clustering effect on treatment is ignored. While method 2 works well in the settings with large sample size, particularly large number of clusters, and with significant clustering effect on the treatment assignment, it tends to overestimate the ACE variance when the sample size drops or the clustering effect on the treatment assignment decreases.

The third simulation experiment tests the efficiency of our methods in settings with skewed covariates while the other simulation conditions remain the same. We find that the performance is consistently satisfactory in both methods. As seen in Table 5, the average coverage rate is about 84% in method 1 and about 89% in method 2. The small difference between the two methods in this simulation experiment indicates that ignoring clustering effect in such settings leads to moderate impact on the estimate of variance. Given that the true ACE is set to 4, we obtain a less bias (average around 0.4) in method 2 than method 1 (average around 1.3). This finding is similar in the other two simulation experiments, indicating that accounting for correlated data in the propensity score model can help reduce the bias in the estimate of ACE.

In some of the simulation scenarios, we see that there is considerable bias given the relative large magnitude of the average bias to the true bias value. In observational studies, the primary source of bias is the confounding effect on the outcome and treatment assignment. Given our simulation models, we found that the difference of the mean outcomes between the two population groups (treatment and control) is around 9. Ideally, we hope this baseline difference can be eliminated by the dual-modeling strategy if all the confounding variables are captured and outcome model is correctly specified. However, we found it can only be reduced to certain degree in the simulation, perhaps due to sampling error and clustering effect. In the causal framework, treatment assignment is conditionally independent of potential outcomes. Therefore we didn't see a significant increase in bias when the true ACE changes, e.g., 40 and 400. In our additional simulation studies (not

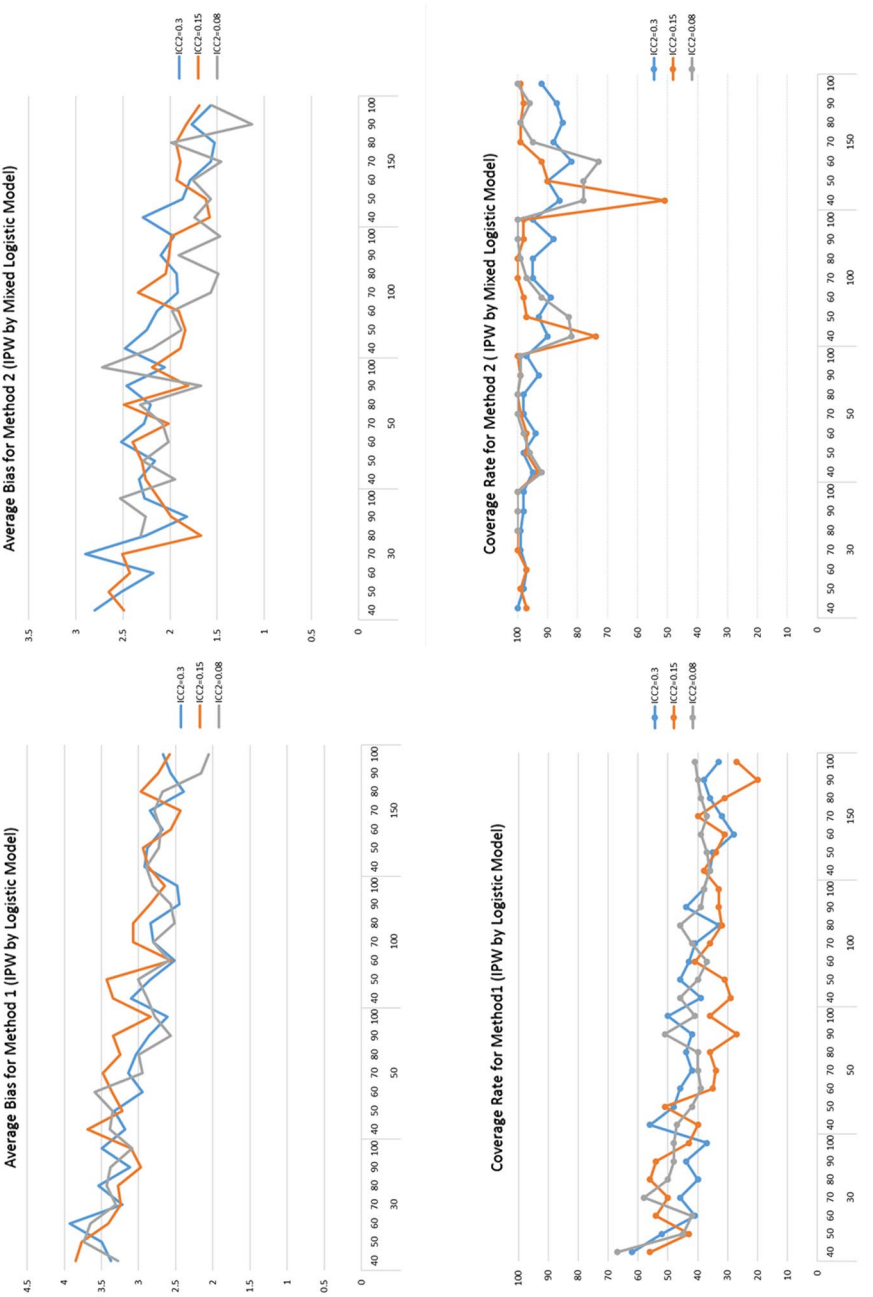

Fig. 1 Average bias and coverage rate by various sample size (ICC1 = 0.6). ICC2: intraclass correlation coefficient for the treatment assignment

reported here), in settings with larger true ACE, bias becomes negligible (the percentage bias is less than 5% when true ACE is 40 or 400). We refer to Schafer and Kang (2008) on similar performances with in-depth discussion.

Our simulation results lead to some important messages for the practitioners of these methods. As discussed above, we see lower percent biases when true ACE is large, and such biases are primarily contributed by the systematic differences in mean outcomes between the two groups in the simulated populations. Given the same set of confounding variables explaining these differences, the amount of bias is relatively consistent across the settings in each method. Overall, method 2 has a better performance than method 1 with respect to lower average bias, higher coverage rate and lower RMSE. However, in some cases with small sample size (as summarized in Table 1) and low clustering effect on treatment assignment model (Tables 2, 3), the ACE variance can be overestimated, which leads to a high coverage rate. This is due to extreme outliers from the inverse propensity scores, and is a limitation in our approach. In such situations, we prefer to use RMSE for the evaluation of the performance. Method 2 has consistently lower RMSE across all simulation scenarios. Given this limitation, we recommend the methods be applied to large survey or administrative data.

6 Application

Our simulation experiments show that method 2 is the preferable analytical method, as it accounts for the distinct sources of variations between and within the clusters. Here, we study the ACE of inadequate prenatal care on birth weight among low income women in New York State.

It has been widely recognized that prenatal care plays an important role in keeping pregnant women and their babies healthy. Early and adequate prenatal care can help identify mothers at risks of adverse pregnancy conditions and reduce the chance of delivering low birth weight babies. Studies in the past decades have shown that inadequate prenatal care is significantly associated with infants born with low birth weight (Donaldson and Billy 1984; Scholl et al. 1987; Hueston 1995; Pedraza et al. 2013; Loftus et al. 2015). Other associated risk factors can be mother's health insurance coverage, demographic, socioeconomic status, drug use and medical conditions such as placental abruption and pre-eclampsia. For example, a recent study shows that Medicaid was protective against low birth weight (Xaverius et al. 2016), but the Medicaid beneficiaries who receive care through public prenatal program are less likely to have a low birth weight baby than those Medicaid beneficiaries who receive care primarily from private-practice physicians (Buescher et al. 1987; Jamieson and Buescher 1992). Although the benefit of adequate care on the reduction of incidents of low birth weight has been widely discussed in literature, the explicit quantitative effect of inadequate prenatal care on birth weight for full term babies has rarely been reported. Our goal in applying method 2 is to investigate to what extent the inadequate prenatal care would affect birth weight. We are particularly interested in the low income population who have Medicaid coverage insurance.

Our primary data source is New York State 2009 vital records which consist of birth certificates which were collected separately from the five boroughs of New York City and the rest of the state. To meet the positivity assumption in causal inference, we restricted the sample to live births from certified hospitals in which unmeasured confounding variables are reduced to a minimum. We kept 54,880 birth records from 120 hospitals with

Table 6 Descriptive statistics for confounding variables

Binary variable	Inadequate care		Adequate care	
	N	%	*N*	%
Previous low birth weight ***	177	0.6	201	0.8
Preexisting hypertension	252	0.8	205	0.8
Pregnancy induced diabetes	1094	3.6	846	3.4
Adverse event	718	2.4	612	2.5
Less than HS education ***	9804	32.4	9563	38.9
SSI eligible	366	1.2	342	1.4
Maternal race: Black ***	6233	20.6	6644	27.0
Maternal race: Hispanic	10,376	34.3	8559	34.8
Maternal race: White ***	10,270	33.9	6877	28.0
No previous live births ***	14,853	49.0	11,190	45.5
Maternal smoking	3942	13.0	3250	13.2
Continuous variable	Mean	SD	Mean	SD
Infants' gestational age	39.4	0.93	39.4	0.97
Mothers' BMI ***	35.6	26.2	33.6	21
Mothers' age	26.1	5.7	25.6	5.9

Adverse event includes abruptio placenta or eclampsia or infection or pregnancy induced hypertension

*** Significantly different between the two groups

a gestational age greater than 37 weeks and that were all covered by the New York State Medicaid program. The number of births in each hospital ranges from 47 to 1677. Covariates in this study include clinical factors reported in birth certificate and demographic collected in Medicaid enrollment records. The two type of information are linked through the mothers' Medicaid member identification number. We define a dichotomous variable for the adequacy of prenatal care based on the Kessner index (Kessner et al. 1973). Women who received more than nine prenatal care visits during their pregnancy are grouped as having received adequate care. Women who received fewer than nine visits are grouped as having received inadequate care. Using this definition, we identify that 55.2% of the mothers have received adequate care. When comparing the covariate distribution between the inadequate care and adequate care group, we observe significant differences in previous low birth weight birth, education, SSI status, race, previous live births, smoking and mother's BMI (Table 6).

In this data, we allow the observational records within each of the hospitals to be correlated. Mothers who gave birth in the same hospital may receive prenatal care due to their provider's affiliation with the hospital. Our descriptive analysis shows that the average birth weight for mothers in the adequate care group was 3397 g with a standard deviation of 70 while the average birth weight for mothers in the inadequate care group is 3351 g with a standard deviation of 75. Therefore, the observed average difference in birth weight between the two groups was 46 g. Based on our knowledge of confounding effects, we include education, race, mothers' age, whether the mom had previous live birth and mothers' BMI in the propensity score model. All of the 14 covariates shown in Table (6) are included in the potential outcome models. Birth weight for infants born to each women is computed based on a random intercept model, as described in method 2. Our results show

that the estimated ACE of inadequate care on birth weight is −24.1 g, which means that receiving inadequate care will reduce infants' birth weight 24.1 g on average. We compute the standard error as 4.7 with the 95% confidence interval (−14.9, −33.3) for the ACE.

7 Discussion

Bias reduction of ACE in non-randomized studies is generally thought as the primary concern in causal inference (Rubin 2006). One of the primary sources of bias is the confounding effect that differs between the treatment and control group due to non-randomization in observational studies. The dual-modeling strategy in our methods has shown an efficient elimination of selection bias through weighting the residuals with inverse propensity scores. Although the logistic regression in method 1 is a misspecified model due to unaddressed clustering effect on the treatment assignment, the estimate of ACE is consistent. Ignoring the clustering effect under the dual-model strategy mainly impacts the estimate of the variance of ACE. Without incorporating the uncertainty of this effect in the treatment assignment model, the standard error of ACE can be potentially underestimated and a low coverage rate would be expected. When using method 1, one should evaluate the clustering effect and make more realistic model specification to avoid this type of underestimation.

This shortcoming has been overcome in method 2 which correctly reflects the clustered data structure into both models. However, when the clustering effect on the treatment assignment is small, there may be issues with the estimation. These issues could be attributed to potential outliers in the predicted propensity scores, which can lead to extreme values in the estimation equations as well as numerical approximation used in integral computations. This indicates that the goodness of fit of the treatment assignment model plays a critical role in the estimation of the variance of the ACE estimate. As a result, we recommend this method for the settings with both large clustering effect on the treatment assignment and number of clusters.

The framework of potential outcomes can be treated as a missing data problem since the potential outcomes cannot be observed (Rubin 2005). Multiply imputing potential outcomes could theoretically improve the performance of the approaches. Additionally, missing data are common in observational studies, and ignoring or deleting the missing data could potentially bias the ACE estimate. We assume complete data in our methods and will address the problem of missing data in the future work.

Acknowledgements The authors are grateful to the referees and Associate Editor whose reviews improved the manuscript significantly. The code used in the computations is available upon request from the authors.

Compliance with ethical standards

Human and animals rights This article does not contain any studies with human participants or animals performed by any of the authors.

Appendix: Formulas for ACE variance estimation in clustered data

Some of the equations for ACE estimate and its variance estimate have been presented in the paper. Here we show the formulas for the computation behind the equations. In method 1, residuals from linear mixed-effects model are adjusted by inverse propensity scores that are estimated by logistic regression. The OLS estimates from linear mixed-effects model (Demidenko 2004) are:

$$\hat{\beta}_0 = \left(x_{ij0}^T x_{ij0}\right)^{-1} x_{ij0}^T \bar{y}_{i0},$$

$$\hat{\beta}_1 = \left(x_{ij1}^T x_{ij1}\right)^{-1} x_{ij1}^T \bar{y}_{i1},$$

$$\hat{\alpha}_0 = \frac{1}{N}\sum_i^m \sum_j^{n_i} (\bar{y}_{ij0} - \bar{x}_{i0}\hat{\beta}_0),$$

$$\hat{\alpha}_1 = \frac{1}{N}\sum_i^m \sum_j^{n_i} (\bar{y}_{ij1} - \bar{x}_{i1}\hat{\beta}_1),$$

$$\hat{\mu}_0 = \frac{1}{N}\sum_i^m \sum_j^{n_i} (x_{ij}\beta_0),$$

$$\hat{\mu}_1 = \frac{1}{N}\sum_i^m \sum_j^{n_i} (x_{ij}\beta_1),$$

$$\hat{\epsilon}_1 = \frac{\sum_i^m \sum_j^{n_i} \hat{\pi}_{ij}^{-1}(y_{ij} - x_{ij}\hat{\beta}_1 - \hat{\alpha}_{i1})}{\sum_i^m \sum_j^{n_i} t_{ij}\hat{\pi}_{ij}^{-1}},$$

$$\hat{\epsilon}_0 = \frac{\sum_i^m \sum_j^{n_i} (1 - t_{ij})(1 - \hat{\pi}_{ij})^{-1}(y_{ij} - x_{ij}\hat{\beta}_0 - \hat{\alpha}_{i0})}{\sum_i^m \sum_j^{n_i} (1 - t_{ij})(1 - \hat{\pi}_{ij})^{-1}}$$

The matrix A is a 9 by 9 lower triangle matrix:

$$\begin{bmatrix} A_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & A_{22} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & A_{33} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & A_{42} & 0 & A_{44} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & A_{53} & 0 & A_{55} & 0 & 0 & 0 & 0 \\ 0 & A_{62} & 0 & 0 & 0 & A_{66} & 0 & 0 & 0 \\ 0 & 0 & A_{73} & 0 & 0 & 0 & A_{77} & 0 & 0 \\ A_{81} & A_{82} & 0 & 0 & 0 & A_{86} & 0 & A_{88} & 0 \\ A_{91} & 0 & A_{93} & 0 & 0 & 0 & A_{97} & 0 & A_{99} \end{bmatrix}$$

where

$$\hat{A}_{11} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}\hat{\pi}_{ij}(1-\hat{\pi}_{ij})z_{ij}z_{ij}^T, \quad \hat{A}_{22} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}(1-t_{ij})x_{ij}^T x_{ij},$$

$$\hat{A}_{33} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}t_{ij}x_{ij}^T x_{ij}, \quad \hat{A}_{42} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}x_{ij}, \quad \hat{A}_{44} = -1, \quad \hat{A}_{53} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}x_{ij},$$

$$\hat{A}_{55} = -1, \quad \hat{A}_{62} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}\bar{x}_{i0}, \quad \hat{A}_{66} = -1, \quad \hat{A}_{73} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}\bar{x}_{i1}, \quad \hat{A}_{77} = -1,$$

$$\hat{A}_{81} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}(1-t_{ij})\hat{\pi}_{ij}(1-\hat{\pi}_{ij})^{-1}(y_{ij} - x_{ij}\hat{\beta}_0 - \alpha_0 - \epsilon_0)z_{ij}^T,$$

$$\hat{A}_{91} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}t_{ij}\hat{\pi}_{ij}^{-1}(1-\hat{\pi}_{ij})(y_{ij} - x_{ij}\hat{\beta}_1 - \alpha_1 - \epsilon_1)z_{ij}^T,$$

$$\hat{A}_{82} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}(1-t_{ij})(1-\hat{\pi}_{ij})^{-1}x_{ij}, \quad \hat{A}_{93} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}t_{ij}\hat{\pi}_{ij}^{-1}x_{ij},$$

$$\hat{A}_{86} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}(1-t_{ij})(1-\hat{\pi}_{ij})^{-1}, \quad \hat{A}_{97} = t_{ij}\hat{\pi}_{ij}^{-1},$$

$$\hat{A}_{88} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}(1-t_{ij})(1-\hat{\pi}_{ij})^{-1}, \quad \hat{A}_{99} = t_{ij}\hat{\pi}_{ij}^{-1}$$

ACE can be estimated by Eq. (8) and its variance can be estimated by Eq. (7).

In method 2, residuals are adjusted by inverse propensity scores that are estimated by logistic mixed-effects model. The re-written equation $\hat{ACE} = a^T\theta$ has a=(0,0,0,0,-1,1,-1,1,-1,1)^T and $\hat{\theta} = (\hat{\gamma}^T, \hat{\sigma}^2, \hat{\beta}_0^T, \hat{\beta}_1^T, \hat{\mu}_0^T, \hat{\mu}_1^T, \hat{\alpha}_0^T, \hat{\alpha}_1^T, \hat{\epsilon}_0^T, \hat{\epsilon}_1^T)$. The maximum likelihood estimates for the linear mixed-effects model are the same as in method 2. $\hat{\gamma}^T$ and $\hat{\sigma}^2$are estimated from logistic mixed-effects model: $P_{t_{ij}=1}(z_{ij};\gamma,\zeta_i) = \pi_{ij}$ and $P_{t_{ij}=0}(z_{ij};\gamma,\zeta_i) = 1-\pi_{ij}$, where $\pi_{ij} = \frac{e^{z_{ij}^T\gamma+\zeta_i}}{1+e^{z_{ij}^T\gamma+\zeta_i}}$ and $\zeta \sim N(0,\sigma)^2)$. The log-likelihood for treatment group is: $l(\gamma,\sigma^2) = -\frac{N}{2}ln(2\pi\sigma^2) + \gamma M + \sum_{i-1}^{m} ln\int_{-\infty}^{\infty} e^{h_i(\gamma,\zeta)}d\zeta$, where $M = \sum_{i}^{m}\sum_{j}^{n_i}t_{ij}z_{ij}, h_i(\gamma,\zeta_i) = k_i\zeta_i - \frac{\zeta_i^2}{2\sigma^2} - \sum_{j=1}^{n_i} ln(1+e^{z_{ij}^T\gamma+\zeta_i}), K_i = \sum_{j=1}^{n_i} t_{ij}$. The maximum likelihood estimates can be obtained from:

$$\frac{\partial l(\gamma,\sigma^2)}{\partial\gamma} = M - \sum_{i=1}^{m}\frac{I_{i3}}{I_{i1}}, \frac{\partial l(\gamma,\sigma^2)}{\partial\sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{m}\frac{I_{i2}}{I_{i1}} \tag{10}$$

where $I_{i1} = \int_{-\infty}^{\infty} e^{h_i(\gamma,\zeta)}d\zeta, I_{i2} = \int_{-\infty}^{\infty}\zeta^2 e^{h_i(\gamma,\zeta)}d\zeta, I_{i3} = \int_{-\infty}^{\infty}\sum_{j=1}^{n_i} Z_{ij}\frac{e^{h_i(\gamma,\zeta)}}{1+e^{h_i(\gamma,\zeta)}}e^{h_i(\gamma,\zeta)}]d\zeta$

Matrix A is a 10 by 10 lower triangle matrix as:

$$\begin{bmatrix} A_{11} & A_{12} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A_{21} & A_{22} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & A_{33} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & A_{44} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & A_{53} & 0 & A_{55} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & A_{64} & 0 & A_{66} & 0 & 0 & 0 & 0 \\ 0 & 0 & A_{73} & 0 & 0 & 0 & A_{77} & 0 & 0 & 0 \\ 0 & 0 & 0 & A_{84} & 0 & 0 & 0 & A_{88} & 0 & 0 \\ A_{91} & A_{92} & A_{93} & 0 & 0 & 0 & A_{97} & 0 & A_{99} & 0 \\ A_{101} & A_{102} & 0 & A_{104} & 0 & 0 & 0 & A_{108} & 0 & A_{1010} \end{bmatrix}$$

where

$$\hat{A}_{11} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}\left\{-\frac{1}{\int_{-\infty}^{\infty} e^{h_i(\gamma,\zeta)}d\zeta}\int_{-\infty}^{\infty}\sum_{j=1}^{n_i} z_{ij}z_{ij}^T\frac{e^{z_{ij}^T\gamma+\zeta_i}}{(1+e^{z_{ij}^T\gamma+\zeta_i})^2}e^{h_i(\gamma,\zeta)}d\zeta + \left(\frac{\int_{-\infty}^{\infty}\sum_{j=1}^{n_i} z_{ij}\frac{e^{z_{ij}^T\gamma+\zeta_i}}{1+e^{z_{ij}^T\gamma+\zeta_i}}e^{h_i(\gamma,\zeta)}d\zeta}{\int_{-\infty}^{\infty} e^{h_i(\gamma,\zeta)}d\zeta}\right)^2\right\},$$

$$\hat{A}_{21} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}\frac{1}{2\sigma^4}\left\{-\frac{\int_{-\infty}^{\infty}\sum_{j=1}^{n_i} z_{ij}\frac{e^{z_{ij}^T\gamma+\zeta_i}}{1+e^{z_{ij}^T\gamma+\zeta_i}}e^{h_i(\gamma,\zeta)}\zeta^2 d\zeta}{\int_{-\infty}^{\infty} e^{h_i(\gamma,\zeta)}d\zeta} + \frac{\int_{-\infty}^{\infty}\zeta^2 e^{h_i(\gamma,\zeta)}d\zeta\int_{-\infty}^{\infty}\sum_{j=1}^{n_i} z_{ij}\frac{e^{z_{ij}^T\gamma+\zeta_i}}{1+e^{z_{ij}^T\gamma+\zeta_i}}e^{h_i(\gamma,\zeta)}d\zeta}{\int_{-\infty}^{\infty} e^{h_i(\gamma,\zeta)}d\zeta\int_{-\infty}^{\infty} e^{h_i(\gamma,\zeta)}d\zeta}\right\},$$

$$\hat{A}_{91} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i} t_{ij}\hat{\pi}_{ij}(1-\hat{\pi}_{ij})^{-1}(y_{ij}-x_{ij}\hat{\beta}_1-\alpha_1-\epsilon_1)z_{ij}^T, \hat{A}_{101} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i} t_{ij}\hat{\pi}_{ij}^{-1}(1-\hat{\pi}_{ij})(y_{ij}-x_{ij}\hat{\beta}_1-\alpha_1-\epsilon_1)z_{ij}^T,$$

$$\hat{A}_{12} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}\left\{-\frac{\int_{-\infty}^{\infty}\sum_{j=1}^{n_i} z_{ij}\frac{e^{z_{ij}^T\gamma+\zeta_i}}{1+e^{z_{ij}^T\gamma+\zeta_i}}e^{h_i(\gamma,\zeta)}\frac{\zeta^2}{2\sigma^4}d\zeta}{\int_{-\infty}^{\infty} e^{h_i(\gamma,\zeta)}d\zeta}, + \frac{\int_{-\infty}^{\infty}\frac{\zeta^2}{2\sigma^4}e^{h_i(\gamma,\zeta)}d\zeta\int_{-\infty}^{\infty}\sum_{j=1}^{n_i} z_{ij}\frac{e^{z_{ij}^T\gamma+\zeta_i}}{1+e^{z_{ij}^T\gamma+\zeta_i}}e^{h_i(\gamma,\zeta)}d\zeta}{\int_{-\infty}^{\infty} e^{h_i(\gamma,\zeta)}d\zeta\int_{-\infty}^{\infty} e^{h_i(\gamma,\zeta)}d\zeta}\right\},$$

$$\hat{A}_{22} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}\left\{\frac{1}{2\sigma^4}-\frac{1}{\sigma^6}\frac{\int_{-\infty}^{\infty}\zeta^2 e^{h_i(\gamma,\zeta)}d\zeta}{\int_{-\infty}^{\infty} e^{h_i(\gamma,\zeta)}d\zeta} + \frac{1}{2\sigma^4}\left(\frac{\int_{-\infty}^{\infty}\frac{\zeta^4}{2\sigma^4}e^{h_i(\gamma,\zeta)}d\zeta}{\int_{-\infty}^{\infty} e^{h_i(\gamma,\zeta)}d\zeta} - \frac{\int_{-\infty}^{\infty}\zeta^2 e^{h_i(\gamma,\zeta)}d\zeta\int_{-\infty}^{\infty}\frac{\zeta^2}{2\sigma^4}e^{h_i(\gamma,\zeta)}d\zeta}{\int_{-\infty}^{\infty} e^{h_i(\gamma,\zeta)}d\zeta\int_{-\infty}^{\infty} e^{h_i(\gamma,\zeta)}d\zeta}\right)\right\},$$

$$\hat{A}_{92} = 0, \quad \hat{A}_{102} = 0, \quad \hat{A}_{33} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}(1-t_{ij})x_{ij}^T x_{ij}, \quad \hat{A}_{53} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i} x_{ij},$$

$$\hat{A}_{73} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}\bar{x}_{i0}, \quad \hat{A}_{93} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}(1-t_{ij})(1-\hat{\pi}_{ij})^{-1}x_{ij},$$

$$\hat{A}_{44} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i} t_{ij}x_{ij}^T x_{ij}, \quad \hat{A}_{64} = \frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i} x_{ij},$$

$$\hat{A}_{84} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}\bar{x}_{i1}, \quad \hat{A}_{104} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i} t_{ij}\hat{\pi}_{ij}^{-1}x_{ij}, \quad \hat{A}_{55} = -1, \quad \hat{A}_{66} = -1, \quad \hat{A}_{77} = -1,$$

$$\hat{A}_{97} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}(1-t_{ij})(1-\hat{\pi}_{ij})^{-1}, \quad \hat{A}_{88} = -1, \quad \hat{A}_{108} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i} t_{ij}\hat{\pi}_{ij}^{-1},$$

$$\hat{A}_{99} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i}(1-t_{ij})(1-\hat{\pi}_{ij})^{-1}, \quad \hat{A}_{1010} = -\frac{1}{N}\sum_{i}^{m}\sum_{j}^{n_i} t_{ij}\hat{\pi}_{ij}^{-1}$$

The 8 integrals in the elements of matrix A can be approximated by the method of Gauss-Hermite quadrature for integrals as described by Demidenko (2004).

$$f_i(1) = \int_{-\infty}^{\infty} e^{h_i(\gamma,\zeta)}d\zeta, \quad f_i(2) = \int_{-\infty}^{\infty} \zeta^2 e^{h_i(\gamma,\zeta)}d\zeta, \quad f_i(3) = \int_{-\infty}^{\infty} \sum_{j=1}^{n_i} z_{ij} \frac{e^{z_{ij}^T\gamma+\zeta_i}}{1+e^{z_{ij}^T\gamma+\zeta_i}} e^{h_i(\gamma,\zeta)}d\zeta,$$

$$f_i(4) = \int_{-\infty}^{\infty} \frac{\zeta^2}{2\sigma^4} e^{h_i(\gamma,\zeta)}d\zeta, \quad f_i(5) = \int_{-\infty}^{\infty} \sum_{j=1}^{n_i} z_{ij}z_{ij}^T \frac{e^{z_{ij}^T\gamma+\zeta_i}}{(1+e^{z_{ij}^T\gamma+\zeta_i})^2} e^{h_i(\gamma,\zeta)}d\zeta,$$

$$f_i(6) = \int_{-\infty}^{\infty} \frac{\zeta^4}{2\sigma^4} e^{h_i(\gamma,\zeta)}d\zeta, \quad f_i(7) = \int_{-\infty}^{\infty} \sum_{j=1}^{n_i} z_{ij} \frac{e^{z_{ij}^T\gamma+\zeta_i}}{1+e^{z_{ij}^T\gamma+\zeta_i}} e^{h_i(\gamma,\zeta)} \frac{\zeta^2}{2\sigma^4}d\zeta,$$

$$f_i(8) = \int_{-\infty}^{\infty} \sum_{j=1}^{n_i} z_{ij} \frac{e^{z_{ij}^T\gamma+\zeta_i}}{1+e^{z_{ij}^T\gamma+\zeta_i}} e^{h_i(\gamma,\zeta)}\zeta^2 d\zeta$$

where $k_i = \sum_{j=1}^{n_i} T_{ij}$ and $h_i(\gamma,\zeta) = k_i\zeta - \frac{\zeta^2}{2\sigma^2} - \sum_{j=1}^{n_i} ln(1+e^{\gamma z_{ij}+\zeta})$.

Better precision can be achieved when the approximation is around the point of maximum of the integrand. For example, to approximate $f_i(1) = \int_{-\infty}^{\infty} e^{h_i(\gamma,\zeta)}d\zeta$, we need to find ζ at the maximum value of $e^{h_i(\gamma,\zeta)}d\zeta$, which is the minimum of $h_i(\gamma,\zeta) = -k_i\zeta + \frac{\zeta^2}{2\sigma^2} + \sum_{j=1}^{n_i} ln(1+e^{\gamma z_{ij}+\zeta})$. Given $\frac{dh_i}{d\zeta} = -k_i + \frac{\zeta}{\sigma^2} + e^{\zeta}\sum_{j=1}^{n_i} \frac{B_j}{1+B_j e^{\zeta}}$ and $\frac{d^2h_i}{d\zeta^2} = \frac{1}{\sigma^2} + e^{\zeta}\sum_{j=1}^{n_i} \frac{B_j}{(1+B_j e^{\zeta})^2}$, where $B_j = e^{\gamma z_{ij}}$, the second derivative is positive. Therefore, the function has a unique minimum. Newton-Raphson algorithm can be used to solve $\frac{dh_i}{d\zeta} = 0$.

$$\zeta_{s+1} = \zeta_s - \left(\frac{dh_i}{d\zeta}\right)\left(\frac{d^2h_i}{d\zeta^2}\right)^{-1} \tag{11}$$

Starting from zero, if ζ_{max} is the limiting point of the iterations, then $f_i(1)$ can be approximated by

$$\int_{-\infty}^{\infty} e^{h(\zeta)}d\zeta \approx \sqrt{2}\hat{v}_h \sum_{k=1}^{K} w_k exp[\zeta_k^2 + h(\zeta_{max} + \sqrt{2}\hat{v}_h\zeta_k)] \tag{12}$$

where $\hat{v}_h = (-\frac{d^2h_i}{d\zeta^2}|_{\zeta=\zeta_{max}})^{-\frac{1}{2}}$. All the integrals can be approximated by this algorithm. However, if the integral is not a unimodal function, the integral needs to be split into two intervals $(-\infty, 0)$ and $(0, \infty)$ for the approximation. Similarly, ACE can be estimated by Eq. (8) and its variance can be estimated by Eq. (7).

References

Austin, P.C.: Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat. Med. **28**(25), 3083–3107 (2009)

Austin, P.C.: An introduction to propensity-score methods for reducing the effects of confounding in observational studies. Multivar. Behav. Res. **46**, 399–424 (2011)

Austin, P.C., Stuart, E.A.: Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat. Med. **34**(28), 3661–3679 (2015)

Austin, P.C., Mamdani, M.M., Stukel, T.A., Anderson, G.M., Tu, J.V.: The use of the propensity score for estimating treatment effects: administrative versus clinical data. Stat. Med. **24**(10), 1563–1578 (2005)

Austin, P.C., Grootendorst, P., Anderson, G.M.: A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a monte carlo study. Stat. Med. **26**(4), 734–753 (2007)

Berg, J.K., Bradshaw, C.P., Jo, B., Lalongo, N.S.: Using complier average causal effect estimation to determine the impacts of the good behavior game preventive intervention on teacher implementers. Adm. Policy Mental Health **44**(4), 558–571 (2017)

Buescher, P.A., Smith, C., Holliday, J.L., Levine, R.H.: Source of prenatal care and infant birth weight: the case of a North Carolina county. Am. J. Obstet. Gynecol. **156**(1), 204–210 (1987)

Cain, L.E., Cole, S.R.: Inverse probability-of-censoring weights for the correction of time-varying noncompliance in the effect of randomized highly active antiretroviral therapy on incident aids or death. Stat. Med. **28**(12), 1725–1738 (2009)

Chan, K.C.: A note about the identifiability of causal effect estimates in randomized trials with non-compliance. Stat. Methodol. **16**, 68–71 (2014)

Cheng, J.: Estimation and inference for the causal effect of receiving treatment on a multinomial outcome. Biometrics **65**(1), 96–103 (2009a)

Connell, A.M.: Employing complier average causal effect analytic methods to examine effects of randomized encouragement trials. Am. J. Drug Alcohol Abuse **35**(4), 253–259 (2009)

Demidenko, E.: Mixed Models Theory Applications. Wiley, New York (2004). ISBN 978-0-471-60161-6

Demirtas, H.: Practical advice on how to impute continuous data when the ultimate interest centers on dichotomized outcomes through pre-specified thresholds. Commun. Stat. Simul. Comput. **36**(4), 871–889 (2007)

Donaldson, P.J., Billy, J.O.: The impact of prenatal care on birth weight: evidence from an international data set. Med. Care **22**(2), 177–188 (1984)

Elliott, M.R., Raghunathan, T.E., Li, Y.: Bayesian inference for causal mediation effects using principal stratification with dichotomous mediators and outcomes. Biostatistics **11**(2), 353–372 (2010)

Frangakis, C.E., Rubin, D.B.: Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment non-compliance and subsequent missing outcomes. Biometrika **86**(2), 365–379 (1999)

Frangakis, C.E., Rubin, D.B.: Principal stratification in causal inference. Biometrika **58**(1), 21–29 (2002)

Gallop, R., Small, D.S., Lin, J.Y., Elliott, M.R., Joffe, M., Ten Have, T.R.: Mediation analysis with principal stratification. Stat. Med. **28**(7), 1108–1130 (2009)

Gitelman, A.I.: Estimating causal effects from multilevel group-allocation data. J. Educ. Behav. Stat. **30**(4), 397–412 (2005)

Gruber, S., Van Der Laan, M.J.: Consistent causal effect estimation under dual misspecification and implications for confounder selection procedures. Stat. Methods Med. Res. **24**(6), 1003–1008 (2015)

Gruber, J.S., Arnold, B.F., Reyqadas, F., Hubbard, A.E., Colford Jr., J.M.: Estimation of treatment efficacy with complier average causal effects (CACE) in a randomized stepped wedge trial. Am. J. Epidemiol. **179**(9), 1134–1142 (2014)

Hernán, M.A.: A definition of causal effect for epidemiological research. J. Epidemiol. Commun. Health **58**(4), 265–271 (2004)

Hernán, M.A., Robins, J.M.: Estimating causal effects from epidemiological data. J. Epidemiol. Commun. Health **60**(7), 578–586 (2006)

Hernán, M.A., Brumback, B., Robins, J.M.: Marginal structural models to estimate the joint causal effect of non-randomized treatments. J. Am. Stat. Assoc. **96**(454), 440–448 (2001)

Hernán, M.A., Brumback, B., Robins, J.M.: Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. Stat. Med. **21**(12), 1689–1709 (2002)

Ho, D.E., Lmai, K., King, G., Stuart, E.A.: Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Polit. Anal. **15**(3), 199–236 (2007)

Holland, P.W.: Statistics and causal inference. J. Am. Stat. Assoc. **81**(396), 945–960 (1986)

Hueston, W.J.: Prenatal care and low-birth-weight rates in urban and rural Wisconsin. J. Am. Board Fam. Med. **8**(1), 17–21 (1995)
Jamieson, D.J., Buescher, P.A.: The effect of family planning participation on prenatal care use and low birth weight. Fam. Plan. Perspect. **24**(5), 214–218 (1992)
Kang, J.D., Schafer, J.L.: Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Stat. Sci. **22**(4), 523–539 (2007)
Kessner, D.M., Singer, J., Kalk, C.E., Schlesinger, E.R.: Infant Death: An Analysis by Matemal Risk and Health Care. Institute of Medicine and National Academy of Sciences, Washington, DC (1973)
Leon, A.C., Demirtas, H., Li, C., Hedeker, D.: Two propensity score-based strategies for a three decade observational study: investigating psychotropic medications and suicide risk. Stat. Med. **31**(27), 3255–3260 (2012a)
Leon, A.C., Hedeker, D., Li, C., Demirtas, H.: Performance of a propensity score adjustment in longitudinal studies with covariate-dependent representation. Stat. Med. **31**(20), 2262–2274 (2012b)
Loftus, C.T., Stewart, O.T., Hensley, M.D., Enquobahrie, D.A., Hawes, S.E.: A longitudinal study of changes in prenatal care utilization between first and second births and low birth weight. Matern. Child Health J. **19**(12), 2627–2635 (2015)
Neyman, J.: On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Stat. Sci. **5**(4), 465–472 (1923)
Pearl, J.: Causality: Models, Reasoning, and Inference, 2nd edn. Cambridge University Press, New York (2009). ISBN 978-0-521-89560-6
Pearl, J.: Principal stratification: a goal or a tool? Int. J. Biostat. **7**(1), 1–13 (2011)
Pedraza, D.F., Rocha, A.C., Cardoso, M.V.: Prenatal care and birth weight: an analysis in the context of family health basic units. Rev. Bras. Ginecol. Obstet. **35**(8), 349–356 (2013)
Robins, J.M.: Marginal structural models versus structural nested models as tools for causal inference. In: Halloran, M.E., Berry, D. (eds.) Statistical Models in Epidemiology: The Environment and Clinical Trials, vol. 116. Springer, New York (1999)
Robins, J.M., Finkelstein, D.M.: Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. Biometrics **56**(3), 779–788 (2000)
Robins, J.M., Rotnitzky, A., Zhao, L.: Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J. Am. Stat. Assoc. **90**(429), 106–121 (1995)
Robins, J.M., Hernán, M.A., Brumback, B.: Marginal structural models and causal inference in epidemiology. Epidemiology **11**(5), 550–560 (2000)
Robins, J.M., Hernán, M.A., Wasserman, L.: Discussion of ön Bayesian estimation of marginal structural models. Biometrics **71**(2), 296–299 (2015)
Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**(1), 41–55 (1983)
Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. **66**(5), 688–701 (1974)
Rubin, D.B.: On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Stat. Sci. **5**(4), 472–480 (1990). [Comment: Neyman (1923) and causal inference in experiments and observational studies]
Rubin, D.B.: Using propensity scores to help design observational studies: application to the tobacco litigation. Health Serv. Outcomes Res. Methodol. **2**(3–4), 169–188 (2001)
Rubin, D.B.: On principles for modeling propensity scores in medical research. Pharmacoepidemiol. Drug Saf. **13**(12), 855–857 (2004)
Rubin, D.B.: Causal inference using potential outcomes: design, modeling, decisions. J. Am. Stat. Assoc. **100**(469), 322–331 (2005)
Rubin, D.B.: Matched Sampling for Causal Effects. Cambridge University Press, New York (2006). ISBN 9780521674362
Schafer, J.L., Kang, J.: Average causal effects from nonrandomized studies: a practical guide and simulated example. Psychol. Methods **13**(4), 279–313 (2008)
Schnitzer, M.E., Lok, J.J., Gruber, S.: Variable selection for confounder control, flexible modeling and collaborative targeted minimum loss-based estimation in causal inference. Int. J. Biostat. **12**(1), 97–115 (2016)
Scholl, T.O., Miller, L.K., Salmon, R.W., Cofsky, M.C., Sheare, J.: Prenatal care adequacy and the outcome of adolescent pregnancy: effects on weight gain, preterm delivery, and birth weight. Am. J. Obstet. Gynecol. **69**(3 Pt 1), 312–316 (1987)
Seaman, S.R., White, I.R.: Review of inverse probability weighting for dealing with missing data. Stat. Methods Med. Res. **22**(3), 278–295 (2013)

Taylor, L., Zhou, X.H.: Relaxing latent ignorability in the ITT analysis of randomized studies with missing data and noncompliance. Stat. Sin. **19**(2), 749–764 (2009)
Vansteelandt, S., Bekaert, M., Claeskens, G.: On model selection and model misspecification in causal inference. Stat. Methods Med. Res. **21**(1), 7–30 (2012)
Waernbaum, L.: Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation. Stat. Med. **31**(15), 1572–1581 (2012)
Westreich, D., Stephen, R.C.: Invited commentary: positivity in practice. Am. J. Epidemiol. **184**(9), 678–681 (2010)
Westreich, D., Edwards, J.K., Cole, S.R.: Imputation approaches for potential outcomes in causal inference. Int. J. Epidemiol. **44**(5), 1731–1737 (2015)
Xaverius, P., Alman, C., Holtz, L., Yarber, L.: Risk factors associated with very low birth weight in a large urban area, stratified by adequacy of prenatal care. Matern. Child Health J. **20**(3), 623–629 (2016)
Zhou, X.H., Li, S.: ITT analysis of randomized encouragement design studies with missing data. Stat. Med. **25**(16), 2737–2761 (2006)