Research Article

Statistics in Medicine

Received 12 March 2013, Accepted 21 August 2013 Published online 9 September 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5969

Inverse probability weighting for covariate adjustment in randomized studies

Changyu Shen,[a*†] Xiaochun Li[a] and Lingling Li[b]

Covariate adjustment in randomized clinical trials has the potential benefit of precision gain. It also has the potential pitfall of reduced objectivity as it opens the possibility of selecting a 'favorable' model that yields strong treatment benefit estimate. Although there is a large volume of statistical literature targeting on the first aspect, realistic solutions to enforce objective inference and improve precision are rare. As a typical randomized trial needs to accommodate many implementation issues beyond statistical considerations, maintaining the objectivity is at least as important as precision gain if not more, particularly from the perspective of the regulatory agencies. In this article, we propose a two-stage estimation procedure based on inverse probability weighting to achieve better precision without compromising objectivity. The procedure is designed in a way such that the covariate adjustment is performed before seeing the outcome, effectively reducing the possibility of selecting a 'favorable' model that yields a strong intervention effect. Both theoretical and numerical properties of the estimation procedure are presented. Application of the proposed method to a real data example is presented. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: clinical trials; covariate adjustment; efficiency; inverse probability weighting; objectivity

1. Introduction

As many clinical trials suffer from high cost, difficulty in recruitment, and subject attrition, efficient estimation of the treatment effect is important. Typical clinical trials collect a fairly rich list of covariates at baseline, including demographics, medical history, medications, lab results, pre-treatment outcome and so on. The rationale behind baseline covariate adjustment is that characteristics predictive of the outcome can be used to account for some of the variation in the outcome, which results in more precise estimate or more powerful test.

The most popular and well-studied method to adjust for baseline covariates for a continuous outcome is the Analysis of Covariance (ANCOVA)[1–3]. The ANCOVA essentially is a regression model by treating the outcome as the dependent variable and baseline covariates and treatment assignment indicator (and possibly their interactions) as the independent variables. In this article, we propose an alternative approach based on inverse probability weighting (IPW) for both continuous and binary outcomes, which also uses baseline covariates to improve efficiency. The IPW estimator is intuitively appealing as weighting is a well-understood approach for estimation problems, and there is no concern of model misspecification because the fitted model is always correct (Section 2.3). Our main theoretical conclusion is that the IPW estimator with the weight estimated from a logistic regression is asymptotically equivalent to the ANCOVA estimator that includes both the main and interaction terms of the covariates and treatment indicator [3] and behaves essentially the same with finite samples. The IPW estimator initially originated from sample survey studies [4] and later was extended to address issues of incomplete

[a]*Department of Biostatistics, School of Medicine, Fairbanks School of Public Health, Indiana University, Indianapolis, IN 46202, U.S.A.*

[b]*Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA 02215, U.S.A.*

**Correspondence to: Changyu Shen, Department of Biostatistics, School of Medicine, Fairbanks School of Public Health, Indiana University, Indianapolis, IN 46202, U.S.A.*

[†]*E-mail: chashen@iupui.edu*

data and confounding in observational studies [5–7]. Following these work in the literature, we estimate the average treatment effect by the difference of the weighted arm-specific means with the weights being the inverse of the 'estimated' probability of receiving the assigned treatment. By 'estimated' probability, we mean fitting a model for the treatment arm indicator given baseline covariates and then estimating the probability based on the model and estimated parameters. This might sound a bit strange as in a randomized trial the treatment assignment is known and does not depend on observed and unobserved covariates. Nevertheless, as explained in Section 2 and verified in Section 3, the 'chance' variation of treatment assignment proportions among different covariate strata allows the model to improve precision.

We want to emphasize that the intention of this article is not to seek the estimator with optimal precision. Instead, our aim is to seek a balance of precision gain and objective inference in clinical trials. There is a rich literature on covariate adjustment for the improvement of precision [1, 2, 8–12], including the recent IPW approach with augmentation by regression function [13, 14]. Nevertheless, there is still debate on whether or not such an adjustment should be made, or if needed, how it should be made [2, 8, 15, 16]. Center to the debate is the issue of objective inference in randomized trials, as covariate adjustment opens the possibility of selecting 'favorable' models. For instance, when ANCOVA is used to adjust for covariates, variable and/or model selection can be used to choose an 'optimal' model that 'best accentuates the estimate and/or statistical significance of the treatment difference' [16].

To objectively incorporate covariate adjustment, Tsiatis *et al.* [3] proposed a principled strategy that relies on analyzing data from the control and intervention arms separately using the regression approach followed by a comparison based on the two fitted regression models. In addition, the analyst is blinded to which treatment arm is being analyzed. The approach allows variable and model selection to optimize efficiency whereas in the meanwhile reduces the probability of fitting a 'favorable' model. However, it is possible that the data analysts might still obtain information from other sources (e.g., study protocol, certain aspects of the data set) that allow them to guess the treatment arm of the data given to them, which may influence the analysis. A more rigorous control is to pre-specify what and how covariates are to be adjusted in a study protocol [17], which also has its limitations. For instance, once the study starts, extra information from other studies or literature may suggest a predictive covariate that is being collected but is not included in the list of covariates to be adjusted. Protocol revision is needed to incorporate this new covariate, which involves many layers of processes and can be time-consuming. As can be seen in Section 2.4, our approach offers a balanced strategy for efficiency improvement and objectivity. A key feature of our approach is that the covariate adjustment is conducted in a way such that there is no need for the covariates and outcome to be in the same data set throughout the analysis process, effectively reducing the possibility of 'cherry picking' by examining the relationship between covariates and the outcome. In addition, our approach offers some level of flexibility as the covariates to be adjusted for do not need to be pre-specified in the study protocol.

In what follows, we provide the background on ANCOVA estimators in clinical trials and IPW estimators in observational studies in Sections 2.1 and 2.2, respectively. Our main result is reported in Sections 2.3 and 2.4. A simulation study is described in Section 3 and a real data application is presented in Section 4. We conclude the article with a discussion section (Section 5).

2. Method

2.1. Background

We consider a typical randomized trial comparing an experimental intervention and a suitable control for a continuous or binary end point. Let Y be the outcome, A be the treatment assignment indicator such that $A = 1$ indicates experimental intervention and $A = 0$ indicates control. The corresponding randomization probabilities are r and $1 - r$, respectively. Denote by X a vector of baseline covariates under consideration. Without loss of generality, we will assume that the mean of X is 0. Finally, let $(Y_i, A_i, X_i), i = 1, 2, \ldots, n$ be the data collected on n subjects, and let Σ denote the summation over the n subjects.

Often the primary interest of the trial is the inference on the difference of the population means under intervention and control. If $Y(0)$ and $Y(1)$ are the potential outcomes under control and intervention, respectively, then $Y = AY(1) + (1 - A)Y(0)$. The goal is to infer

$$\theta = E(Y(1)) - E(Y(0)) = E(Y|A = 1) - E(Y|A = 0).$$

Here, the second equality is due to the randomization. A simple and widely employed estimator of θ is the difference of the sample means

$$\hat{\theta}_1 = \frac{1}{\sum A_i}\sum Y_i A_i - \frac{1}{\sum (1-A_i)}\sum Y_i(1-A_i) = \frac{1}{n}\sum\left(\frac{Y_i A_i}{\hat{r}} - \frac{Y_i(1-A_i)}{1-\hat{r}}\right), \tag{1}$$

where $\hat{r} = \sum A_i/n$. As baseline covariates that correlate with the outcome can be used to improve precision, alternative estimators based on ANCOVA have been proposed in the literature [3]. Specifically, least square estimators $\hat{\theta}_2$ of the model

$$E(Y|X,A) = \beta_0 + \beta_1^T X + \theta_2 A, \tag{2}$$

and $\hat{\theta}_3$ of the model

$$E(Y|X,A) = \gamma_0 + \gamma_1^T X + \gamma_2^T XA + \theta_3 A \tag{3}$$

have been shown to be consistent estimators of θ and asymptotically normal. An important feature of these two estimators is that the aforementioned properties hold even if models (2) and (3) are incorrect. In the succeeding text, we summarize some results from Tsiatis *et al.* [3] regarding the estimators $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$. First, all three estimators belong to a general class of augmented semi-parametric estimators Θ,

$$\Theta: \frac{1}{n}\sum\left\{\left(\frac{A_i}{r} - \frac{1-A_i}{1-r}\right)Y_i - \frac{A_i - r}{r(1-r)}\left(rg_0(X_i) + (1-r)g_1(X_i)\right)\right\} + o_p(1/\sqrt{n}), \tag{4}$$

where g_0 and g_1 are arbitrary scalar functions of X that distinguish different estimators. For example, the two functions are constants for $\hat{\theta}_1$ with $g_0(X) = E(Y|A=0)$ and $g_1(X) = E(Y|A=1)$, whereas the two functions are of the form $a + b^T X$ with different values of a (scalar) and b (vector) for $\hat{\theta}_2$ and $\hat{\theta}_3$. Second, $\hat{\theta}_3$ is efficient among all estimators with both g_0 and g_1 linear in X. Third, $\hat{\theta}_2$ and $\hat{\theta}_3$ are asymptotically equivalent when $r = 0.5$. Lastly, the estimator with the smallest variance in this class is the one with $g_0(X) = E(Y|X, A=0)$ and $g_1(X) = E(Y|X, A=1)$, that is, the true regression model within each arm. The last point is not surprising because in that case the second term in Equation (4) is the projection of the first term to the space of functions of X and, therefore, minimizes the variance of the random variable in Equation (4). In a similar sense, $g_0(X)$ and $g_1(X)$ for $\hat{\theta}_3$ essentially are predictions of the outcome based on a linear regression model within each arm.

The class of estimators in Equation (4) is what is called the augmented inverse probability weighting estimator [18] with the probability being the marginal randomization probability. It turns out that a simple IPW estimator with the probability being the estimated propensity score accounting for X also belongs to this class. In addition, if the propensity score is estimated by a logistic regression, then the simple IPW estimator is asymptotically equivalent to $\hat{\theta}_3$. In the rest of Section 2, we provide more details along these lines.

2.2. *Inverse probability weighting in observational studies*

In a typical observational study, the setting and research goal is similar to the randomized trial in Section 2.1, except that the treatment assignment indicator A may stochastically depend on X and even unobserved variables. In this setting, $\hat{\theta}_1$ usually is biased because components of X associated with Y may not have the same distribution between the control and intervention arms. The concept of propensity score [19] was introduced to correct the bias. A key assumption for the propensity score-based approach is no uncontrolled confounding

$$(Y(0), Y(1)) \perp A|X, \tag{5}$$

which means conditional on X, the potential outcomes are independent of treatment assignment indicator. The propensity score is then defined as the conditional probability of $A = 1$ given X, $p_0(X) = \Pr[A = 1|X]$, which can be used in different ways to eliminate or reduce bias [20]. Among them, the IPW approach essentially estimates θ by the difference of weighted sample means with the weight being the reciprocal of the probability of receiving the treatment actually assigned to the subject (the propensity score is assumed to be bounded away from 0 and 1):

$$\tilde{\theta}_I = \frac{1}{n}\sum\left(\frac{Y_i A_i}{p_0(X_i)} - \frac{Y_i(1-A_i)}{1-p_0(X_i)}\right). \tag{6}$$

Straight forward algebra shows that $\tilde{\theta}_I$ is an unbiased estimator of θ.

In practice, the propensity score usually is unknown and has to be estimated. A common practice is to fit a parametric model (e.g., logistic regression) to the data to obtain an estimated propensity score $p(X,\hat{\alpha})$, where $\hat{\alpha}$ is the parameter estimate usually obtained through maximum likelihood estimation. As long as the propensity model is correctly specified, that is, there exists some α_0 such that $p(X,\alpha_0) = p_0(X)$ for all X, the estimator

$$\hat{\theta}_I = \frac{1}{n}\sum\left(\frac{Y_i A_i}{p(X_i,\hat{\alpha})} - \frac{Y_i(1-A_i)}{1-p(X_i,\hat{\alpha})}\right) \tag{7}$$

is consistent for θ and asymptotically normal [5].

An interesting and seemingly counter-intuitive phenomenon is that $\hat{\theta}_I$ is asymptotically at least as efficient as $\tilde{\theta}_I$ [6, 7]. The rationale behind this is that the estimator $\hat{\alpha}$ results in an empirical propensity $(p(X_i,\hat{\alpha}))$ that is more effective in reducing the variation of the numerators in Equation (7) than the true propensity. Moreover, we say $p(X,\alpha)$ and $q(X,W,\alpha,\eta)$ are two nested identifiable parametric models with covariates X and (X, W) in the sense that there exists an $\eta_0(\alpha)$ such that $p(X,\alpha) = q(X,W,\alpha,\eta_0(\alpha))$ for all α, X, and W. Suppose the true propensity score is such that $p_0(X,W) = q(X,W,\alpha_0,\eta_0(\alpha_0)) = p(X,\alpha_0)$, then $\hat{\theta}_I^L$ using the estimated propensity score based on $q(X,W,\alpha,\eta)$ is at least as efficient as $\hat{\theta}_I^S$ based on $p(X,\alpha)$ [6, 7]. In other words, inclusion of extra covariates not related to the treatment assignment will at least not compromise the efficiency asymptotically. We provide the proof of these results in A.1 and A.2 in the Appendix. Brookhart *et al.* [21] has shown the same finding through simulation studies.

2.3. *Inverse probability weighting in randomized trials*

In many randomized trials, there exist baseline covariates that are correlated with the endpoint (e.g., prognostic factors), and therefore, 'chance' imbalance of these covariates between the two arms will increase the variation of the difference in sample means. The results in Section 2.2 suggests that fitting 'larger' propensity score models can improve precision in observational studies, which motivated the idea of IPW estimation in randomized trials for the improvement of efficiency. For a randomized trial, the propensity score is known with a simple form: $p_0(X) = r$ for all$X$. Therefore, any family of distributions covering this simple distribution is a correct model. In fact, $\hat{\theta}_1$ in Section 2.1 is an IPW estimator based on a model with only a constant. The following theorem shows that IPW estimator $\hat{\theta}_I$ based on a parametric model $p(X,\alpha)$ belongs to the same class of estimators as $\hat{\theta}_1$,$\hat{\theta}_2$ and $\hat{\theta}_3$. The proof is included in A.3 of the Appendix.

Theorem 1
Let $p(X,\alpha)$ be a smooth parametric model such that there exists an α_0 with $p(X,\alpha_0) = r$ for all X. Let $\hat{\alpha}$ be the maximum likelihood estimator of α ,then

$$\hat{\theta}_I = \frac{1}{n}\sum\left(\frac{Y_i A_i}{p(X_i,\hat{\alpha})} - \frac{Y_i(1-A_i)}{1-p(X_i,\hat{\alpha})}\right)$$

belongs to the class Θ in Equation (4).

Theorem 1 establishes the potential efficiency gain by using a parametric family to 'model' the treatment assignment mechanism in a randomized trial. Specifically, on the basis of the results in 2.1, if the chosen model $p(X,\alpha)$ is such that asymptotically $g_0(X) = E(Y|X,A=0)$ and $g_1(X) = E(Y|X,A=1)$ for $\hat{\theta}_I$, then $\hat{\theta}_I$ will be efficient. Certainly, such a parametric model might be difficult to identify as that would require the knowledge of the true regression model within each arm. Nonetheless, as summarized in Section 2.1, a simplification with g_0 and g_1 linear in X can still improve efficiency over the naive estimator $\hat{\theta}_1$. Logistic regression is the most widely used parametric model for the propensity score. The following corollary states that $\hat{\theta}_I$based on logistic regression is the most efficient estimator with both g_0 and g_1 linear in X. The proof is included in A.4 of the Appendix.

Corollary 1

If $p(X,\alpha) = 1/\left[1+\exp\left(-\alpha^T X^*\right)\right], X^* = (1, X^T)^T$, then $\hat{\theta}_I$ is asymptotically equivalent to $\hat{\theta}_3$ in Section 2.1. In other words, $\hat{\theta}_I$ is efficient among all estimators with both g_0 and g_1 linear in X. Moreover,

$$\begin{aligned}
&\sqrt{n}\left(\hat{\theta}_I - \theta\right) \to N(0,\Gamma), \Gamma = B - C^T DC \\
&B = Var\left[\left(\frac{A}{r} - \frac{(1-A)}{1-r}\right) Y\right], C = rE(X^*Y|A=0) + (1-r)E(X^*Y|A=1), \\
&D = \frac{1}{r(1-r)}\left[E\left(X^*X^{*T}\right)\right]^{-1}.
\end{aligned} \tag{8}$$

In many trials, $r = 0.5$, which leads to

$$\begin{aligned}
\Gamma &= 4Var((2A-1)Y) - 4E\left(X^{*T}Y\right)\left[E\left(X^*X^{*T}\right)\right]^{-1} E(X^*Y) \\
&= 4\left[Var\left((2A-1)Y\right) - [E(Y)]^2 - E\left(X^TY\right)\left[E\left(XX^T\right)\right]^{-1} E(XY)\right] \\
&= 4\left[Var\left((2A-1)Y\right) - [E(Y)]^2 - R^2 Var(Y)\right].
\end{aligned} \tag{9}$$

The R^2 in Equation (9) represents the R-square statistic by fitting a linear regression model to infinite number of data points of (X, Y) with half from the control and the other half from the intervention. Clearly, R^2 in this case determines the amount of efficiency gain. More importantly, because R^2 will not decrease with added covariates, Equation (9) indicates that adding more covariates to the logistic regression model for the propensity score will at least not compromise the asymptotic efficiency. A generalization of this result is that models with more covariates are at least as efficient as models with fewer covariates.

Theorem 2

Let X and W be non-overlapping baseline covariate vectors. Let $p(X,\alpha)$ and $q(X,W,\alpha,\eta)$ be smooth nested identifiable parametric models in the sense that there exists an $\eta_0(\alpha)$ such that $p(X,\alpha) = q(X,W,\alpha,\eta_0(\alpha))$ for every α, X and W, and there is an α_0 such that $p(X,\alpha_0) = r$ for all X. Let $\hat{\alpha}$ and $(\breve{\alpha},\breve{\eta})$ be the maximum likelihood estimators of the model parameters for the two models, respectively. Let $\hat{\theta}_I$ and $\breve{\theta}_I$ be the corresponding IPW estimators

$$\hat{\theta}_I = \frac{1}{n}\sum\left(\frac{Y_i A_i}{p(X_i,\hat{\alpha})} - \frac{Y_i(1-A_i)}{1-p(X_i,\hat{\alpha})}\right), \breve{\theta}_I = \frac{1}{n}\sum\left(\frac{Y_i A_i}{q(X_i,W_i,\breve{\alpha},\breve{\eta})} - \frac{Y_i(1-A_i)}{1-q(X_i,W_i,\breve{\alpha},\breve{\eta})}\right).$$

Then $\breve{\theta}_I$ is at least as efficient as $\hat{\theta}_I$ asymptotically.

The proof of Theorem 2 essentially is the same as the proof in A.2 and is omitted here.

2.4. *Two-stage analysis strategy based on inverse probability weighting estimator*

The variance of the limiting distribution of the IPW estimator based on a logistic regression, Γ, can be consistently estimated by

$$\begin{aligned}
&\hat{\Gamma} = V - M^2 - H, \\
&V = \tfrac{1}{n}\sum\left[\left(\tfrac{A_i}{r} - \tfrac{(1-A_i)}{1-r}\right) Y_i\right]^2, M = \tfrac{1}{n}\sum\left[\left(\tfrac{A_i}{r} - \tfrac{(1-A_i)}{1-r}\right) Y_i\right], H = \tfrac{1}{nr^4(1-r)^4}\sum\left[Y_i^2(A_i-r)^4 q_i^2\right],
\end{aligned} \tag{10}$$

where q_i is the ordinary least squares (OLS) fitted value of $A_i - r$ by a linear regression of $A - r$ on $X^* = (1, X^T)^T$. The proof of the consistency of Equation (10) is included in A.5 of the Appendix.

A unique feature of the variance estimator in Equation (10) is that the calculation only requires (Y, A, q) and does not need X. The property allows a two-stage analysis strategy for objective inference in randomized clinical trials, which is easy to implement. The main feature of this approach is that the analysis is performed in two stages by either the same data analyst or two analysts. In either case, the data analyst(s) never see(s) the outcome data and the baseline covariates together in the same data set for drawing the primary conclusion of a study, effectively reducing the possibility of selecting a favorable

model through examination of the relationship between the covariates and the outcome. In addition, the covariates to be adjusted do not need to be specified until the time of fitting the propensity score. In the succeeding text are the steps on how this approach is implemented with two data analysts.

(1) Stage 1 is performed by analyst 1 with data on (A,X):
   (a) Fit a logistic regression model for A (treatment assignment indicator) with whatever covariate vector X deemed appropriate. Obtain the estimated propensity score $p(X,\hat{\alpha})$ for each subject.
   (b) Fit a linear regression model by OLS for $A-r$ using the same covariate vector X. Obtain the fitted value q for each subject.
   (c) Pass $p(X,\hat{\alpha})$ and q to analyst 2.
(2) Stage 2 is performed by analyst 2 with data on $(Y, A, p(X,\hat{\alpha}), q)$:
   Calculate the IPW estimate $\hat{\theta}_I$ as in Theorem 1 and the standard error using formulae (10) for statistical inference.

Note that X needs not to be centered in the procedure earlier. Clearly, X and Y are never in the same data during the two-stage analysis process. Compared with the approach by Tsiatis *et al.* [3], our strategy has more stringent control on what data the analysts have access to. On the other hand, the estimator might not be the most efficient because our approach does not allow the search for an 'optimal' model for $E(Y|X, A)$. However, this is necessary to avoid potential bias introduced from the variable/model selection process so as to maintain the objectivity of the study. In summary, the two-stage IPW estimator offers an improvement in precision without compromising objectivity. The same procedure can be implemented with one analyst, where the data set (A,X) can be made unavailable to the analyst after stage 1 through data management arrangement.

3. A simulation study

Theorem 2 suggests that larger models with more covariates will not compromise the asymptotic efficiency, which might not be the case in finite samples. In particular, when logistic regression is used, an added covariate with zero partial correlation with the outcome will not increase R^2 and may induce efficiency loss [22].

We first conducted a simulation study to understand the finite sample performance of the IPW estimator for a continuous outcome. The main inference target is the difference of the population means under intervention and control, θ. We generated data using the following model:

$$[Y|X, A] = N\left(\sum_{j=1}^{k}\beta_j X_j + \gamma A, \delta^2\right), X_j \sim_{i.i.d} N(0,1), \Pr[A=1|X] = 0.5.$$

On the basis of this model, $\theta = \gamma$. Let $\sigma^2 = \sum_{j=1}^{k}\beta_j^2$. Then, simple algebra shows that $R^2 = \frac{\sigma^2}{\sigma^2+\delta^2}$. In our simulation, we fix $\sigma^2 = \delta^2 = 2$ so that $R^2 = 0.5$. An R-square value of 0.5 is realistic as often times, there exists a baseline outcome measure that is correlated with the post-treatment outcome measure with a Pearson correlation coefficient as high as 0.7 [23], leading to an R-square of 0.49 based on a single covariate. We generated a total of 20 independent standard normal variables and selected $k = 5$, 10, and 20 of them in the regression model. For a given k, we divide the covariate into five groups of equal size and assume the same coefficient values for covariates within the same group. The five unique coefficients are two-fold apart at square scale. Therefore, as k goes from 5 to 20, the coefficients for the involved X gets smaller with the summation at the square scale fixed at $\sigma^2 = 2$. This reflects realistic situations of either several strong prognostic factors or a number of weak prognostic factors. We generated 2000 Monte Carlo data sets each composed of 200, 500, or 1000 subjects.

Three estimation procedures are considered, including the unadjusted estimator $\hat{\theta}_1$ (UNADJ), $\hat{\theta}_I$ based on logistic regression (IPW), and the optimal estimator that is in the form of formulae (4) with $g_0(X_i)$ and $g_1(X_i)$ replaced by the fitted values from separately estimated linear models within each of the two arms (OPT). Actually in this case, IPW, OPT, and the two ANCOVA estimators ($\hat{\theta}_2$ and $\hat{\theta}_3$ in Section 2.1) are all asymptotically equivalent. For both IPW and OPT estimators, all 20 covariates are included in the corresponding models. Note that this means 15 and 10 covariates are not associated with the outcome when $k = 5$ and 10. Although asymptotically equivalent, the OPT with all 20 covariates may have diffe-

rent finite sample performance from the OPT with the relevant covariates only. Nonetheless, we will still keep all 20 covariates and call it OPT to reflect the realistic situation where some of the covariates are not correlated to the outcome conditional on the covariates already included in the model. The results are summarized in Tables I and II. As all three estimators are essentially unbiased (bias<0.01), we omitted the summary of bias here. From Table I, it is clear that IPW and OPT are more efficient than UNADJ, with the standard error about 70% that of the UNADJ for all scenarios. It is quite interesting to see this is the case even when k=5, where 15 out of the 20 included covariates are not predictive of the outcome. The IPW and OPT estimators have essentially the same efficiency. When $n = 200$, both IPW and OPT show clear trends in underestimating the standard error, leading coverage probability below the nominal level. This is particularly apparent for IPW. As sample size gets large, the standard error estimate gets more accurate, except a slight downward bias leading to a bit under-coverage when $n = 500$ and $k = 20$. Table II summarizes the rejection probability for the hypothesis $H_0 : \theta = 0$. Because IPW and

Table I. Simulation results for a continuous endpoint based on 2000 Monte Carlo simulations.

		$SE_M/SE_A(\times 1000)$			Cov. Prob.		
		UNADJ	IPW	OPT	UNADJ	IPW	OPT
$n = 200$	$k = 5$	290/282	220/189	204/197	0.945	0.914	0.941
	10	287/282	221/189	207/194	0.943	0.900	0.932
	20	287/283	223/189	214//189	0.949	0.904	0.913
$n = 500$	$k = 5$	176/178	129/124	127/126	0.952	0.943	0.951
	10	178/179	127/124	126/125	0.955	0.941	0.946
	20	174/179	129/124	128/124	0.956	0.938	0.939
$n = 1000$	$k = 5$	125/126	89/89	89/89	0.958	0.949	0.953
	10	130/126	90/88	90/89	0.947	0.942	0.944
	20	126/126	90/88	90/88	0.953	0.945	0.946

The average treatment effect is zero. $\delta^2 = \sigma^2 = 2$. For IPW, a logistic regression including all 20 X variables is fitted to the data. For OPT, all 20 X variables are included in separate linear models for the two arms. SE_M is the standard error based on Monte Carlo samples; SE_A is the average of large sample standard error estimates over the Monte Carlo samples. Cov.Prob. is the coverage probability of the 95% confidence interval based on large-sample normal approximation.

Table II. Rejection probability for a continuous endpoint.

			UNADJ	IPW	OPT
$\theta = 0$	n=200	k=5	0.055	0.086	0.060
		10	0.057	0.100	0.069
		20	0.051	0.096	0.088
	$n = 500$	$k = 5$	0.048	0.057	0.049
		10	0.045	0.059	0.054
		20	0.044	0.063	0.061
	$n = 1000$	$k = 5$	0.043	0.051	0.048
		10	0.053	0.059	0.057
		20	0.047	0.056	0.054
$\theta = 0.25$	$n = 1000$	$k = 5$	0.508	0.808	0.803
		10	0.505	0.790	0.794
		20	0.519	0.803	0.805
$\theta = 0.5$	$n = 200$	$k = 5$	0.434	0.714	0.706
		10	0.430	0.729	0.724
		20	0.413	0.725	0.735
	$n = 500$	$k = 5$	0.807	0.974	0.975
		10	0.797	0.971	0.974
		20	0.804	0.978	0.979

All simulation and estimation parameters are the same as Table I, except more values on the true treatment effect.

OPT underestimate the standard error when $n = 200$, there is an apparent inflation of type I error. When sample size is large, both IPW and OPT perform equally well, both are more powerful than the UNADJ. When $n = 500$ and $k = 20$, the slight downward bias in estimating the standard error leads to a slight inflation in the type I error rate.

We also conducted a simulation study for a binary endpoint. The primary interest is the difference of the proportions under the intervention and control. We generated data from the following model:

$$\Pr[Y = 1|X, A] = \Phi\left(\alpha + \sum_{j=1}^{k} \beta_j X_j + \gamma A\right), X_j \sim_{i.i.d} N(0, 1), \Pr[A = 1|X] = 0.5,$$

where Φ is the cumulative distribution function of a standard normal variable. We fix $\sigma^2 = \sum_{j=1}^{k} \beta_j^2 = 3$ and consider $k = 5$, 10, and 20 out of 20 X variables as in the case of a continuous endpoint. The β values are also assigned the same way as in the continuous case. The true proportions for the control and intervention arms are $\Phi(\alpha/2)$ and $\Phi((\alpha + \gamma)/2)$, respectively. Because the true regression model is not linear anymore, the OPT is in the form of formulae (4) with $g_0(X_i)$ and $g_1(X_i)$ replaced by the fitted value from separately estimated probit models within each of the two arms. We also included $\hat{\theta}_2$ (ANCOVA) for comparison. Tables III and IV summarize the results based on 2000 Monte Carlo data sets of size 1000. Overall, IPW and ANCOVA are very similar; both are more efficient and powerful than

Table III. Simulation results for a binary endpoint based on 2000 Monte Carlo data sets of size 1000.

		SE_M/SE_A(×10000)				Cov. Prob.			
(Δ_0, Δ_1)		UNADJ	IPW	ANCOVA	OPT	UNADJ	IPW	ANCOVA	OPT
(0.1, 0.15)	$k = 5$	209/209	180/174	176/174	157/152	0.950	0.941	0.948	0.943
	10	208/209	174/174	170/174	152/150	0.948	0.952	0.955	0.946
	20	208/204	177/174	173/174	155/149	0.950	0.943	0.946	0.941
(0.3, 0.39)	$k = 5$	288/300	214/220	214/220	200/205	0.965	0.959	0.966	0.962
	10	301/300	227/220	224/220	210/205	0.942	0.944	0.946	0.947
	20	296/300	223/220	220/220	209/203	0.953	0.945	0.949	0.939

For IPW, a logistic regression for the propensity score including all 20 X variables is fitted to the data. For ANCOVA, all 20 X variables are included in the model. For OPT, all 20 X variables are included in separate logistic outcome regression models for the two arms. SE_M is the standard error based on Monte Carlo samples; SE_A is the average of large sample standard error estimates over the Monte Carlo samples. Cov.Prob. is the coverage probability of the 95% confidence interval based on large-sample normal approximation. (Δ_0, Δ_1) are the true proportions in the control and intervention arms, respectively.

Table IV. Rejection probability for a binary endpoint.

(Δ_0, Δ_1)		UNADJ	IPW	ANCOVA	OPT
(0.1, 0.1)	$k = 5$	0.047	0.050	0.043	0.048
	10	0.050	0.053	0.050	0.064
	20	0.054	0.056	0.052	0.075
(0.3, 0.3)	$k = 5$	0.047	0.051	0.049	0.050
	10	0.052	0.055	0.051	0.058
	20	0.048	0.058	0.054	0.050
(0.1, 0.15)	$k = 5$	0.696	0.835	0.826	0.924
	10	0.691	0.838	0.830	0.926
	20	0.675	0.822	0.814	0.921
(0.3, 0.39)	$k = 5$	0.876	0.989	0.989	0.997
	10	0.866	0.985	0.981	0.995
	20	0.861	0.984	0.980	0.991

All simulation and estimation parameters are the same as Table III. (Δ_0, Δ_1) are the true proportions in the control and intervention arms, respectively.

UNADJ but less so than the OPT. This certainly is expected. When the true event rates in the control and intervention arms are both 0.1, the OPT shows an inflation of type I error rate (Table IV) for $k = 10$ and 20. Therefore, the gain in power by OPT over IPW/ANCOVA when the true event rates are 0.1 and 0.15 might actually be less than what is shown in Table IV.

4. Application to the primary biliary cirrhosis data

The Mayo Clinic conducted a double-blinded randomized trial in primary biliary cirrhosis of the liver, in which the drug D-penicillamine (DPCA) was compared with a placebo [24]. There were 424 patients who met the eligibility criteria, and 312 agreed to participate in the randomized trial. We focus on the 312 subjects in the randomized trial. Our primary aim is to compare the 2-year mortality between DPCA and the placebo. For this purpose, the mortality status by year two can be ascertained for 311 subjects with one subject censored before year two due to liver transplantation. Therefore, our analysis includes 311 subjects, of whom 157 received DPCA and 154 received the placebo. The 2-year mortality rates for the DPCA and placebo are $14/157 = 8.9\%$ and $19/154 = 12.3\%$.

For each of the 311 subjects, a large number of clinical, biochemical, serologic, and histologic parameters were collected at baseline, which may be used to improve the precision of the comparison. To apply the proposed method, we included all covariates in the data set PBC [25] except those with missing values. There are in total 12 covariates included in our propensity model, including sex, age in years at study registration, presence of ascites (yes/no), presence of hepatomegaly (yes/no), presence of spiders (yes/no), presence of edema (0: no edema and no diuretic therapy for edema; 1: edema present for which no diuretic therapy was given, or edema resolved with diuretic therapy; 2: edema despite diuretic therapy), serum bilirubin (mg/dl), albumin (gm/dl), alkaline phosphatase (U/l), SGOT (U/ml), prothrombin time (s), and histologic stage (1, 2, 3, or 4). Here, presence of edema and histologic stage are treated as categorical variables in the model. We summarize the distribution of the 12 covariates in Table V. It can be seen that most of the covariates are balanced well. Proportion of hepatomegaly is slightly higher in the placebo arm, and serum bilirubin is slightly higher in the DPCA arm.

The analysis result is shown in Table VI, where UNADJ, IPW, and OPT are the same estimators as in the simulation studies. The efficiency gain of IPW and OPT over UNADJ is apparent. IPW reduces the standard error by about 20% as compared with UNADJ or a relative efficiency of about 1.6. Clearly, some of the covariates are predictive of mortality. In fact, it was shown through model selection that a

Table V. Baseline covariates for the 311 subjects in the primary biliary cirrhosis data.

	Placebo (n=157)	D-penicillamine (n=154)
Male	20 (12.7%)	15 (9.7%)
Age (year)	51.4 (11.0)	48.6 (10.0)
Presence of ascites	143 (91.1%)	144 (93.5%)
Presence of hepatomegaly	85 (54.1%)	67 (43.5%)
Presence of spiders	112 (71.3%)	109 (70.1%)
Presence of edema		
0	131 (83.4%)	131 (85.1%)
1	16 (10.2%)	13 (8.4%)
2	10 (6.4%)	10 (6.5%)
Serum bilirubin (mg/dl)	2.88 (3.64)	3.65 (5.28)
Albumin (gm/dl)	3.52 (0.44)	3.52 (0.40)
Alkaline phosphatase (U/l)	156 (87)	144 (83)
SGOT (U/ml)	91.3 (53.7)	90.7 (54.6)
Prothrombin time (s)	18.2 (16.1)	21.4 (16.5)
Histologic stage		
1	12 (7.6%)	4 (2.6%)
2	35 (22.3%)	32 (20.8%)
3	56 (35.7%)	64 (41.6%)
4	54 (34.4%)	54 (35.1%)

Numbers in parenthesis for continuous variables are the standard deviations.
SGOT, serum glutamic oxaloacetic transaminase.

Table VI. Treatment effect on 2-year mortality rate using 311 subjects in the PBC data.

Method	Estimate of the difference in mortality rates (%) (DPCA minus placebo)	Standard error (%)	p-value
UNADJ	−3.42	3.49	0.33
IPW	−4.62	2.75	0.09
OPT	−5.23	2.65	0.05

For IPW, logistic regression with the main effects of the twelve covariates was used to fit the propensity model. For OPT, separate logistic regression models for the outcome were fitted to the two arms with the main effects of the twelve covariates.
IPW, inverse probability weighting; DPCA, D-penicillamine.

Cox model including age, serum bilirubin, albumin, prothrombin time, and presence of edema predicted mortality well [24]. The interesting aspect of our analysis is that even with extra variables that are not predictive of the endpoint, the efficiency gain is still quite pronounced. On a cautionary note, our simulation studies showed that the asymptotic standard error estimate has downward bias when sample size is relatively small (e.g., $n = 200$). Thus, the actual efficiency gain in this case may be smaller than 20%.

Serum bilirubin is positively correlated with 2-yr mortality, and it has higher value in the DPCA arm, which will make the estimate more negative (i.e., more treatment effect) after adjustment. Yet, presence of hepatomegaly is also positively correlated with the outcome, and it has higher proportion in the placebo arm, which will make the estimate more positive (i.e., less treatment effect). Estimates from IPW and OPT are lower than the UNADJ, which is a net consequence of adjusting for multiple variables.

5. Discussion

The IPW estimator is shown to be asymptotically equivalent to the ANCOVA estimator that includes both the main effects of baseline covariates and their interactions with the treatment assignment indicator. The performance under finite samples is also similar. Our simulation study shows that substantial efficiency gain can be achieved with IPW as long as the propensity model includes covariates with moderate to small correlations with the outcome. This is the case even when a significant proportion of the covariates included are not correlated with the outcome, and the sample size is relatively small. Our simulation studies also suggest that there may be downward bias in the asymptotic standard error estimate when the sample size is relatively small, which could lead to inflation of type-I error. Thus, we suggest simulation studies be performed in practice to properly calibrate standard error estimate when sample size is small.

It is well established that stratification variables in a randomized trial should be included in the analysis of variance model to for efficiency gain, which is also true for the proposed IPW method. Other covariates can include those deemed to be related to the outcomes. One advantage of our method is that the list of covariates to be included in the propensity model does not need to be specified until the actual fitting of the propensity score. The analysis enjoys some level of flexibility on which covariates to use based on updated literature and other sources of information. In addition, if feasible, one can perform a 'stage 0' analysis to examine the relationship between covariates and the outcome without the presence of treatment assignment indicator to select a list of covariates with possibly better prediction power for efficiency improvement.

The two-stage analysis strategy to draw inference on the treatment effect based on IPW offers an improvement in precision without compromising objectivity. As covariate adjustment using IPW is more efficient than the unadjusted estimator asymptotically, and is so under realistic finite samples, it provides a simple and easy-to-implement solution for clinical trials striving for both cost reduction and objective interpretation of the data. The IPW approach prevents the analyst from predicting the outcome through variable and model selection. This may sound limited because model and variable selection through the examination of the relation of a large covariate list and outcome may offer better precision. However, such analyses can be prone to subjectivity and subsequently leads to misleading results [22]. In addition, in certain cases, the gain in efficiency through variable selection over pre-specified variable list might be limited as the extra variables outside the pre-specified list often have weak partial correlation with the outcome. It should be noted that IPW approach may allow the data analyst to select covariates after data on (X,A) are seen (e.g., select 'unbalanced' covariates in addition to the pre-specified list). Nevertheless,

without extra knowledge on the correlations between X and Y beyond what was known at the design stage, such practice might not help much [22].

One difficulty of IPW estimators encountered in observational studies is that the estimated probability can be close to 0 or 1, leading to numerical instability. This is unlikely to occur in randomized trials as the treatment assignment ratio is fixed at some value far from 0 and 1. Therefore, the typical concern of extreme weights in applying IPW is not relevant here.

Appendix A

A.1. Asymptotic efficiency of $\hat{\theta}_I$

By Taylor expansion at the true value α_0, we have

$$\begin{aligned}
\sqrt{n}(\hat{\theta}_I-\theta) &= \frac{1}{\sqrt{n}}\sum\left[\left(\frac{Y_iA_i}{p_0(X_i)}-\frac{Y_i(1-A_i)}{1-p_0(X_i)}\right)-\theta\right]-\frac{1}{\sqrt{n}}\sum\left(\frac{Y_iA_i}{p_0(X_i)}-\frac{Y_i(1-A_i)}{1-p_0(X_i)}\right)S_i^T(\hat{\alpha}-\alpha_0)+o_p(1)\\
&= \frac{1}{\sqrt{n}}\sum\left[\left(\frac{Y_iA_i}{p_0(X_i)}-\frac{Y_i(1-A_i)}{1-p_0(X_i)}\right)-\theta\right]-\left[\frac{1}{n}\sum\left(\frac{Y_iA_i}{p_0(X_i)}-\frac{Y_i(1-A_i)}{1-p_0(X_i)}\right)S_i^T\right][\sqrt{n}(\hat{\alpha}-\alpha_0)]+o_p(1)\\
&= \frac{1}{\sqrt{n}}\sum\left[\left(\frac{Y_iA_i}{p_0(X_i)}-\frac{Y_i(1-A_i)}{1-p_0(X_i)}\right)-\theta\right]-\left[E\left\{\left(\frac{Y_iA_i}{p_0(X_i)}-\frac{Y_i(1-A_i)}{1-p_0(X_i)}\right)S_i^T\right\}+o_p(1)\right]\\
&\quad\times\left[\frac{1}{\sqrt{n}}\sum I^{-1}S_i+o_p(1)\right]+o_p(1)\\
&= \frac{1}{\sqrt{n}}\sum\left[\underbrace{\left(\frac{Y_iA_i}{p_0(X_i)}-\frac{Y_i(1-A_i)}{1-p_0(X_i)}\right)-\theta}_{U_i}-\underbrace{E\left\{\left(\frac{Y_iA_i}{p_0(X_i)}-\frac{Y_i(1-A_i)}{1-p_0(X_i)}\right)S_i^T\right\}I^{-1}S_i}_{V_i}\right]+o_p(1)\\
&= \frac{1}{\sqrt{n}}\sum(U_i-V_i)+o_p(1)
\end{aligned}$$

Here, S_i is the score function for subject i evaluted at the true value α_0, and I is the Fisher information matrix evaluted at α_0. Therefore, by central limit theorem, $\sqrt{n}\left(\hat{\theta}_I-\theta\right)$ is asymptotically normal with variance equal to $Var(U-V)$. Because $E(U)=E(V)=0$ and $Var(V)=E(UV)$), it follows

$$Var(U-V)=Var(U)+Var(V)-2E(UV)=Var(U)-Var(V)$$

Because $Var(U)$ is the variance of the limiting distribution of $\sqrt{n}\left(\tilde{\theta}_I-\theta\right)$, it follows that $\hat{\theta}_I$ is at least as efficient as $\tilde{\theta}_I$.

A.2. Asymptotic efficiency of inverse probability weighting estimators based on two nested parametric models

Let $\hat{\theta}_I^L$ be the IPW estimator based on the fitted propensity score using the 'large' model $q(\cdot,\alpha,\eta)$, and $\hat{\theta}_I^S$ be the IPW estimator based on the fitted propensity score using the 'small' model $p(\cdot,\alpha)$. Based on the result in A.1, we have

$$\sqrt{n}\left(\hat{\theta}_I^L-\theta\right)\to N\left(0,Var(U)-Var\left(V^L\right)\right)\text{ and }\sqrt{n}\left(\hat{\theta}_I^S-\theta\right)\to N\left(0,Var(U)-Var\left(V^S\right)\right).$$

Here, $V^L=E\left\{\left(\frac{YA}{p_0(X)}-\frac{Y(1-A)}{1-p_0(X)}\right)S^T\right\}I^{-1}S$, $V^S=E\left\{\left(\frac{YA}{p_0(X)}-\frac{Y(1-A)}{1-p_0(X)}\right)S_1^T\right\}I_{11}^{-1}S_1$;
S and I are the score and Fisher information matrix evaluated at (α_0,η_0) for the large model, and S_1 and I_{11} are the score and Fisher information matrix evaluated at α_0 for the small model.

Write

$$\begin{aligned}
&S^T=\left(S_1^T,S_2^T\right),I=\begin{pmatrix}I_{11}&I_{12}\\I_{21}&I_{22}\end{pmatrix},Z_1^T=E\left\{\left(\frac{YA}{p_0(X)}-\frac{Y(1-A)}{1-p_0(X)}\right)S_1^T\right\},\\
&Z_2^T=E\left\{\left(\frac{YA}{p_0(X)}-\frac{Y(1-A)}{1-p_0(X)}\right)S_2^T\right\},\\
&\Omega=\left(I_{22}-I_{21}I_{11}^{-1}I_{12}\right)^{-1}.
\end{aligned}$$

By the formula of blockwise inversion,

$$I^{-1} = \begin{pmatrix} I_{11}^{-1} + I_{11}^{-1} I_{12} \Omega I_{21} I_{11}^{-1} & -I_{11}^{-1} I_{12} \Omega \\ -\Omega I_{21} I_{11}^{-1} & \Omega \end{pmatrix}.$$

$$\begin{aligned} Var(V^L) &= Z_1^T \left(I_{11}^{-1} + I_{11}^{-1} I_{12} \Omega I_{21} I_{11}^{-1} \right) Z_1 - Z_1^T I_{11}^{-1} I_{12} \Omega Z_2 - Z_2^T \Omega I_{21} I_{11}^{-1} Z_1 + Z_2^T \Omega Z_2 \\ &= Z_1^T I_{11}^{-1} Z_1 + \left(I_{21} I_{11}^{-1} Z_1 - Z_2 \right)^T \Omega \left(I_{21} I_{11}^{-1} Z_1 - Z_2 \right) \\ &= Var(V^S) + \left(I_{21} I_{11}^{-1} Z_1 - Z_2 \right)^T \Omega \left(I_{21} I_{11}^{-1} Z_1 - Z_2 \right) \end{aligned}$$

Because Ω is positive definite, $\left(I_{21} I_{11}^{-1} Z_1 - Z_2 \right)^T \Omega \left(I_{21} I_{11}^{-1} Z_1 - Z_2 \right) \geqslant 0$. Thus, $Var(V^L) \geqslant Var\left(V^S\right)$.
$\hat{\theta}_I^L$ is at least as efficient as $\hat{\theta}_I^S$.

A.3. Proof of theorem 1

Under randomization, the true propensity score $p_0(X) = r$ for all X. On the basis of the result of A.1, the influence function for $\hat{\theta}_I$ can be written as

$$IF = \frac{YA}{r} - \frac{Y(1-A)}{1-r} - \theta - E\left\{ \left(\frac{YA}{r} + \frac{Y(A-1)}{1-r} \right) S^T \right\} I^{-1} S$$

The score function S and Fisher information matrix can be written as

$$S = \left[\frac{A-r}{r(1-r)} \right] \frac{\partial p(X, \alpha_0)}{\partial \alpha}.$$

Because $X \perp A$, the Fisher information matrix can be written as

$$I = E\left[\frac{(A-r)^2}{r^2(1-r)^2} \frac{\partial p(X, \alpha_0)}{\partial \alpha} \frac{\partial p(X, \alpha_0)}{\partial \alpha^T} \right] = \frac{1}{r(1-r)} E\left(\frac{\partial p(X, \alpha_0)}{\partial \alpha} \frac{\partial p(X, \alpha_0)}{\partial \alpha^T} \right).$$

By plugging the expression of S and I in the expression of IF,we have

$$IF = \left(\frac{A}{r} - \frac{1-A}{1-r} \right) Y - \theta - \frac{A-r}{r(1-r)} [r g_0(X) + (1-r) g_1(X)],$$

where

$$\begin{aligned} g_0(X) &= E\left(\frac{Y(A-1)(A-r)}{r(1-r)} \frac{\partial p(X, \alpha_0)}{\partial \alpha^T} \right) \left[E\left(\frac{\partial p(X, \alpha_0)}{\partial \alpha} \frac{\partial p(X, \alpha_0)}{\partial \alpha^T} \right) \right]^{-1} \frac{\partial p(X, \alpha_0)}{\partial \alpha}, \\ g_1(X) &= E\left(\frac{YA(A-r)}{r(1-r)} \frac{\partial p(X, \alpha_0)}{\partial \alpha^T} \right) \left[E\left(\frac{\partial p(X, \alpha_0)}{\partial \alpha} \frac{\partial p(X, \alpha_0)}{\partial \alpha^T} \right) \right]^{-1} \frac{\partial p(X, \alpha_0)}{\partial \alpha}. \end{aligned}$$

A.4. Proof of corollary 1

Under logistic regression model, $\partial p(X, \alpha_0)/\partial \alpha = r(1-r)X^*$. Plugging in the result in Theorem 1 leads to

$$\begin{aligned} g_0(X) &= \frac{1}{r(1-r)} E\left[Y(A-1)(A-r) X^{*T} \right] \left[E\left(X^* X^{*T} \right) \right]^{-1} X^* \\ &= \frac{1}{r(1-r)} \left[r(1-r) E(X^{*T} Y) \right] \left[E\left(X^* X^{*T} \right) \right]^{-1} X^* \\ &= E(X^{*T} Y | A = 0) \left[E\left(X^* X^{*T} \right) \right]^{-1} X^* \\ &= E(Y | A = 0) + E\left(X^T Y | A = 0 \right) \left[E\left(X X^T \right) \right]^{-1} X \end{aligned}$$

Similarly,

$$\begin{aligned} g_1(X) &= \frac{1}{r(1-r)} E\left[YA(A-r)X^{*T}\right]\left[E\left(X^*X^{*T}\right)\right]^{-1} X^* \\ &= E\left(X^{*T}Y|A=1\right)\left[E\left(X^*X^{*T}\right)\right]^{-1} X^* \\ &= E(Y|A=1) + E\left(X^TY|A=1\right)\left[E\left(XX^T\right)\right]^{-1} X \end{aligned}$$

Therefore, $\hat{\theta}_I$ has the same g_0 and g_1 as $\hat{\theta}_3$ (Tsiatis et al.,2008), and they are asymptotically equivalent. Because the influence function is

$$IF = \frac{YA}{r} - \frac{Y(1-A)}{1-r} - \theta - E\left\{\left(\frac{YA}{r} + \frac{Y(A-1)}{1-r}\right)S^T\right\} I^{-1}S,$$

The asymptotic variance of $\hat{\theta}_I$ is

$$\text{Avar}(\hat{\theta}_I) = Var\left[\left(\frac{A}{r} - \frac{(1-A)}{1-r}\right)Y\right] - E\left\{\left(\frac{A}{r} + \frac{(A-1)}{1-r}\right)YS^T\right\} I^{-1} E\left\{\left(\frac{A}{r} + \frac{(A-1)}{1-r}\right)YS\right\}$$

Because $S = (A-r)X^*$ and $I = r(1-r)E\left(X^*X^{*T}\right)$,

$$\begin{aligned} \text{Avar}\left(\hat{\theta}_I\right) &= B - C^T DC \\ B &= Var\left[\left(\frac{A}{r} - \frac{(1-A)}{1-r}\right)Y\right], C = rE(X^*Y|A=0) + (1-r)E(X^*Y|A=1), \\ D &= \frac{1}{r(1-r)}\left[E\left(X^*X^{*T}\right)\right]^{-1} \end{aligned}$$

A.5. Proof of consistency of the estimator $\hat{\Gamma}$ in Equation (10)

First, it is easy to see that $V - M^2 \to_P B = Var\left[\left(\frac{A}{r} - \frac{(1-A)}{1-r}\right)Y\right]$.
Let

$$\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T, \mathbf{A} = (A_1, A_2, \ldots, A_n)^T, \mathbf{X}^* = \left(X_1^*, X_2^*, \ldots, X_n^*\right)^T, \mathbf{Q} = (q_1, q_2, \ldots, q_n)^T,$$

$$\mathbf{r} = \left(\underbrace{r, r, \ldots, r}_{n}\right)^T,$$

and let $\mathbf{J}$ be a diagonal matrix with the i-th diagnonal entry being $(A_i - r)^2$.
By the property of OLS, $\mathbf{Q} = \mathbf{X}^*(\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}(\mathbf{A}-\mathbf{r})$. Then

$$\begin{aligned} H &= \frac{1}{nr^4(1-r)^4}\mathbf{Y}^T\mathbf{J}\mathbf{X}^*(\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}(\mathbf{A}-\mathbf{r})(\mathbf{A}-\mathbf{r})^T\mathbf{X}^*(\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}\mathbf{J}\mathbf{Y} \\ &= \frac{1}{r^4(1-r)^4}\left[\frac{1}{n}\mathbf{Y}^T\mathbf{J}\mathbf{X}^*\right]\left[n(\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\right]\left[\frac{1}{n}\mathbf{X}^{*T}(\mathbf{A}-\mathbf{r})(\mathbf{A}-\mathbf{r})^T\right]\left[n(\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\right]\left[\frac{1}{n}\mathbf{X}^{*T}\mathbf{J}\mathbf{Y}\right] \end{aligned}$$

By the law of the large numbers and the continuous mapping theorem

$$n(\mathbf{X}^{*T}\mathbf{X}^*)^{-1} = n\left(\sum X_i^*X_i^{*T}\right)^{-1} \to_P \left[E\left(X^*X^{*T}\right)\right]^{-1},$$

$$\frac{1}{n}\mathbf{X}^{*T}(\mathbf{A}-\mathbf{r})(\mathbf{A}-\mathbf{r})^T\mathbf{X}^* = \frac{1}{n}\sum(A_i-r)^2X_i^*X_i^{*T} \to_P E\left((A-r)^2X^*X^{*T}\right) = r(1-r)E\left(X^*X^{*T}\right),$$

$$\frac{1}{n}\mathbf{Y}^T\mathbf{J}\mathbf{X}^* = \frac{1}{n}\sum(A_i-r)^2X_i^{*T}Y_i \to_P E\left((A-r)^2X^{*T}Y\right).$$

It follows that

$$H \to_P \frac{1}{r^3(1-r)^3}E\left((A-r)^2X^{*T}Y\right)\left[E\left(X^*X^{*T}\right)\right]^{-1}E\left((A-r)^2X^*Y\right)$$

Because $E\left((A-r)^2 X^{*T} Y\right) = r(1-r)^2 E(X^{*T} Y|A=1) + r^2(1-r)E(X^{*T} Y|A=0)$,
It follows that

$$H \to_P C^T DC, C = rE(X^* Y|A=0) + (1-r)E(X^* Y|A=1), D = \frac{1}{r(1-r)}\left[E\left(X^* X^{*T}\right)\right]^{-1}$$

Hence, $\hat{\Gamma} \to_P \Gamma$ by the definition of Γ in Equation (8).

Acknowledgements

This work is supported in part by National Institutes of Health grant R21 CA152463 and the Indiana University Health-Indiana University School of Medicine Strategic Research Initiative in Cardiology.

References

1. Leon S, Tsiatis AA, Davidian M. Semiparametric estimation of treatment effect in a pretest-posttest study. *Biometrics* 2003; **59**(4):1046–1055.
2. Lesaffre E, Senn S. A note on non-parametric ANCOVA for covariate adjustment in randomized clinical trials. *Statistics in Medicine* 2003; **22**(23):3583–3596.
3. Tsiatis AA, Davidian M, Zhang M, Lu X. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statistics in Medicine* 2008; **27**(23):4658–4677.
4. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 1952; **47**(260):663–685.
5. Li L, Shen C, Li X, Robins JM. On weighting approaches for missing data. *Statistical Methods in Medical Research* 2013; **22**(1):14–30.
6. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 1994; **89**(427):846–866.
7. Rotnitzky A, Li L, Li X. A note on overadjustment in inverse probability weighted estimation. *Biometrika* 2010; **97**(4):997–1001.
8. Koch GG, Tangen CM, Jung JW, Amara IA. Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. *Statistics in Medicine* 1998; **17**(15-16):1863–1892.
9. Senn SJ. Covariate imbalance and random allocation in clinical trials. *Statistics in Medicine* 1989; **8**(4):467–475.
10. Yang L, Tsiatis AA. Efficiency study for a treatment effect in a pretest-posttest trial. *The American Statistician* 2001; **55**(4):314–321.
11. Zhang M, Davidian M. "Smooth" semiparametric regression analysis for arbitrarily censored time-to-event data. *Biometrics* 2008; **64**(2):567–576.
12. Tian L, Cai T, Zhao L, Wei LJ. On the covariate-adjusted estimation for an overall treatment difference with data from a randomized comparative clinical trial. *Biostatistics* 2012; **13**(2):256–273.
13. Rotnitzky A, Lei Q, Sued M, Robins JM. Improved double-robust estimation in missing data and causal inference models. *Biometrika* 2012; **99**(2):439–456.
14. Rubin DB, van der Laan MJ. Empirical efficiency maximization: improved locally efficient covariate adjustment in randomized experiments and survival analysis. *International Journal of Biostatistics* 2008; **4**(1). article 5.
15. Grouin JM, Day S, Lewis J. Adjustment for baseline covariates: an introductory note. *Statistics in Medicine* 2004; **23**(5):697–699.
16. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine* 2002; **21**(19):2917–2930.
17. Chan AW, Tetzlaff JM, Gotzsche PC, Altman DG, Mann H, Berlin JA, Dickersin K, Hrobjartsson A, Schulz KF, Parulekar WR, Krleza-Jeric K, Laupacis A, Moher D. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. *British Medical Journal* 2013; **346**:e7586.
18. Tsiatis AA. *Semiparametric Theory and Missing Data*. Springer: New York, 2006.
19. Rosenbaum P, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**(1):41–55.
20. Rosenbaum P. *Observational Studies*. Springer: New York, 1995.
21. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. *American Journal of Epidemiology* 2006; **163**(12):1149–1156.
22. Raab GM, Day S, Sales J. How to select covariates to include in the analysis of a clinical trial. *Controlled Clinical Trials* 2000; **21**(4):330–342.
23. Frison L, Pocock SJ. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. *Statistics in Medicine* 1992; **11**(13):1685–1704.
24. Dickson ER, Grambsch PM, Fleming TR, Fisher LD, Langworthy A. Prognosis in primary biliary cirrhosis: model for decision making. *Hepatology* 1989; **10**(1):1–7.
25. Fleming TR, Harrington DP. *Counting Process* and Survival Analysis, (2nd edn). Wiley: New York, 2005.