Biometrics 65, 640–649
June 2009

DOI: 10.1111/j.1541-0420.2008.01066.x

A Comparison of Methods for Estimating the Causal Effect of a Treatment in Randomized Clinical Trials Subject to Noncompliance

Roderick J. Little,[1,*] Qi Long,[2,] and Xihong Lin[3,***]**

[1]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
[2]Department of Biostatistics, Emory University, Atlanta, Georgia 30322, U.S.A.
[3]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, U.S.A.
[*]*email:* rlittle@umich.edu
[**]*email:* qlong@sph.emory.edu
[***]*email:* xlin@hsph.harvard.edu

Summary. We consider the analysis of clinical trials that involve randomization to an active treatment ($T = 1$) or a control treatment ($T = 0$), when the active treatment is subject to all-or-nothing compliance. We compare three approaches to estimating treatment efficacy in this situation: as-treated analysis, per-protocol analysis, and instrumental variable (IV) estimation, where the treatment effect is estimated using the randomization indicator as an IV. Both model- and method-of-moment based IV estimators are considered. The assumptions underlying these estimators are assessed, standard errors and mean squared errors of the estimates are compared, and design implications of the three methods are examined. Extensions of the methods to include observed covariates are then discussed, emphasizing the role of compliance propensity methods and the contrasting role of covariates in these extensions. Methods are illustrated on data from the Women Take Pride study, an assessment of behavioral treatments for women with heart disease.

Key words: As-treated analysis; Causal inference; Efficacy; Instrumental variables; Per-protocol analysis; Principal stratification; Propensity scores.

1. Introduction

Randomized clinical trials that compare treatments are generally straightforward to analyze, but the analysis and interpretation is complicated when individuals do not comply with their assigned treatments. The gold-standard analysis of such trials in drug approval processes is intention to treat (ITT), where individuals are classified in treatment comparisons according to their assigned treatments, regardless of whether they complied with the treatment. ITT analysis preserves the benefits of randomization, and it provides valid measures of the effect of treatment assignment, sometimes called treatment *effectiveness*. The analysis is less compelling for estimating treatment *efficacy*, the effectiveness of a treatment when it is in fact taken.

Simple approaches to estimating treatment efficacy are as-treated (AT) analysis, where participants are classified according the treatment actually received, and per-protocol (PP) analysis, which restricts analysis to participants who comply with the assigned treatment. These analyses estimate treatment efficacy, because they classify participants according to received treatment, but they are subject to selection bias, in that participants who comply with a treatment may be a biased sample of participants randomized to that treatment. The bias may be reduced by adjustment for covariates, but it remains a concern. Thus much current clinical trial practice for estimating efficacy involves an unappealing choice between ITT analysis, which is protected from bias by randomization but is really estimating effectiveness, and PP and AT analyses, which measure efficacy but are potentially biased by treatment noncompliance.

Another more recent approach to estimating efficacy treats the randomization as an instrumental variable (IV), in economic parlance. In simple terms the IV estimator corrects the ITT estimator for noncompliance, based on certain assumptions about the outcomes for noncompliers under both treatments. This approach yields a direct estimate of treatment efficacy, and is protected from selection bias by the randomization. On the other hand, it does require certain assumptions to be valid, and it also yields estimators with potentially high variance, particularly if the treatment compliance rate is low. Model-based versions of the IV estimator have been proposed that are potentially more efficient, although they make stronger distributional assumptions.

This article has two objectives. First, we provide a side-by-side comparison of the PP, AT, and IV estimators in the situation with no covariates. Two key ideas here are the definition of a *causal effect* of an active treatment as the difference in hypothetical outcomes under that treatment and a control treatment (Rubin, 1974, 1977, 1978); and *principal stratification*, where individuals are stratified according to the values of the posttreatment variable under *both* treatments, rather than simply under the treatment actually observed (Frangakis and

Rubin, 2002). The latter paper generalized the application of the potential outcomes framework in the compliance setting (Baker and Lindeman, 1994; Angrist, Imbens, and Rubin, 1996, henceforth denoted as AIR) to any postrandomization variable. White (2005) adopts a similar framework in arguing for IV over PP analysis in the context of randomized trials. We include AT in our analysis, provide a more overt discussion of the contrasting assumptions of the methods, and compare their precisions and mean squared errors (MSEs). Second, we discuss and compare the role of covariates, and in particular compliance propensities proposed in this article, in improving the performance of the PP, AT, and IV estimators. We show that covariates can reduce bias and weaken the assumptions for PP and AT estimation, but not for ITT or IV estimation; however, they can improve precision for all the methods. We propose covariate-adjusted IV estimators. Finally we outline extensions to more general settings.

A subset of the data from the "Women Take Pride" (WTP) study, a behavior cardiovascular intervention study (Janevic et al., 2003), is used to compare the various estimators of treatment efficacy. Participants are older women with heart disease, and the intervention consists of 6 weekly classes to groups of six to eight women, where strategies for managing the disease are developed. Compliance is defined here as attendance at least one class, and hence the analysis does not distinguish between women who fully comply and women who partially comply with treatment—see Baker (2000) for more discussion of this issue. We compare subjects randomized to this treatment group ($R = 1$, $n = 190$) with subjects randomized to a control "usual care" treatment ($R = 0$, $n = 184$) on one of the main study outcomes. We apply the IV, AT, and PP methods to the WTP study without and with covariates. We discuss the underlying assumptions of these methods in this study, and note that the assumptions for estimating efficacy that are suitable for clinical trials, such as the exclusion restriction (ER) that is required by the IV, might not hold for behavior intervention studies, which are often unblinded.

2. Estimation of Treatment Efficacy: The Case of No Covariates

2.1 *Estimands of Treatment Efficacy*

We consider randomized studies involving random assignment to an active treatment ($R = 1$) and a control treatment ($R = 0$). We assume the treatments are subject to all-or-nothing compliance (Baker, 1997), so that the actual treatment received (say $T(R)$) can differ from the treatment assigned (R). Specifically, we assume that the population can then be divided into three groups: *never-takers* ($C = n$), who take the control treatment whether they are assigned to the control or active treatment ($T(1) = T(0) = 0$), *compliers* who take the treatment they are assigned ($C = c$) ($T(R) = R$), and *always-takers* ($C = a$), who take the active treatment whether assigned the active or control treatment ($T(1) = T(0) = 1$). We assume that there are no *defiers* who take the opposite treatment to that assigned. This is also called the *monotonicity* assumption. We make the stable unit-treatment value assumption (SUTVA; Rubin, 1978), which implies that compliance and outcomes for individuals are not affected by the assignments and outcomes of other individuals in the sample (AIR). It is justified in the WTP study, because individuals given the control treatment did not have access to the classes in the intervention arm.

We call C *principal compliance*, because it is a special case of principal stratification where the posttreatment variable is compliance; subjects are classified by their compliance under both treatments (Frangakis and Rubin, 2002). Principal compliance differs from *observed compliance*, which concerns only whether a participant complied with the assigned treatment. Observed noncompliers in the treatment group are never-takers ($C = n$), observed compliers in the treatment group are compliers or always-takers ($C = c$ or a), observed noncompliers in the control group are always-takers ($C = a$), and observed compliers in the control group are compliers or never-takers ($C = c$ or n). Thus C is only partly observed. Because it is unaffected by the treatment assigned, it can be used as a stratification variable in treatment comparisons, if the missing data problem can be solved.

Table 1A shows a classification of the population by R and C, assuming a proportion α of the population is assigned to the treatment, and population proportions π_n, π_c, π_a of never takers, compliers, and always takers, respectively. The entries reflect independence of R and C, which is a consequence of random treatment assignment.

Consider now an outcome variable Y, and let μ_{rj} denote the mean of Y when assigned $R = r$ ($r = 0, 1$) for the subpopulation with $C = j$, ($j = n$, c, or a); let $\bar{y}_{rj}$ denote the corresponding sample mean, and m_{rj} the corresponding sample size. Table 1B displays population means of Y, with square parentheses when corresponding sample quantities are not observed. The observed sample counts and means are shown in Table 1C. Because there are six cell means in Table 1B, and only four observed means, two restrictions on the means are needed to just identify the model. See White (2005) for a similar presentation of the data.

An ITT estimate in this setting is:

$$\hat{\delta}_{\text{ITT}} = \bar{y}_{1+} - \bar{y}_{0+}. \tag{1}$$

It is protected from selection bias by the randomization, and it measures treatment *effectiveness*, the causal effect of assigning the treatment without regard to compliance. It arguably does not provide a satisfactory estimate of *efficacy*, that is, the effect of the treatment itself, because a treatment that is not taken cannot be expected to be effective. We present ITT estimates here for completeness, but focus on other approaches to estimating efficacy that use compliance information.

Two estimands of treatment efficacy have been considered (Robins, 1989; Robins and Greenland, 1996; Imbens and Rubin, 1997b). The complier-average causal effect (CACE) is the average treatment effect (ATE) in the subpopulation of principal compliers:

$$\delta_{\text{CACE}} = \mu_{1c} - \mu_{0c}, \tag{2}$$

(AIR). The ATE is defined as the difference in mean outcome if all individuals had been assigned and complied with the treatment ($T = 1$) and the mean if all individuals had been assigned and complied with the control treatment ($T = 0$). The ATE requires assumptions about the treatment outcome for noncompliers, in the counterfactual event that they had complied with the treatment. Whether this counterfactual

Table 1
Classifications by treatment and principal compliance: (A) population proportions; (B) population mean outcomes; (C) observed means (sample counts)

(A) Population proportions		Principal compliance C A	C	N	ALL
Randomized treatment R	0	$(1-\alpha)\pi_a$	$(1-\alpha)\pi_c$	$(1-\alpha)\pi_n$	$1-\alpha$
	1	$\alpha\pi_a$	$\alpha\pi_c$	$\alpha\pi_n$	α
	ALL	π_a	π_c	π_n	

(B) Population mean outcomes		Principal compliance C A	C	N	ALL
Randomized treatment R	0	μ_{0a}	$[\mu_{0c}]$	$[\mu_{0n}]$	μ_{0+}
	1	$[\mu_{1a}]$	$[\mu_{1c}]$	μ_{1n}	μ_{1+}
	ALL	$[\mu_{+a}]$	$[\mu_{+c}]$	$[\mu_{+n}]$	

[.] = quantity in parentheses not directly estimable without assumptions

C. Observed means (sample counts)		Principal compliance C A	C	N	ALL
Randomized treatment R	0	$\bar{y}_{0a}\ (m_{0a})$	$\overbrace{\bar{y}_{0(c+n)}\ (m_{0(c+n)})}$		$\bar{y}_{0+}\ (m_{0+})$
	1	$\overbrace{\bar{y}_{1(c+a)}\ (m_{1(c+a)})}$		$\bar{y}_{1n}\ (m_{1n})$	$\bar{y}_{1+}\ (m_{1+})$
	ALL	?	?	?	

event is meaningful arguably depends on context. For example, noncompliance to a behavioral treatment such as an exercise regime might plausibly be changed by increased motivation, as might occur if evidence of success of the treatment becomes widely known. On the other hand, if noncompliance to a drug is the result of intolerable side effects, then compliance may require a reformulation of the drug to remove the side effects. This may change the properties of the drug, and estimation of the ATE is consequently more speculative.

In the absence of covariates, the ATE and CACE are the same if the ATE is the same for compliers as for noncompliers if they had in fact complied. When this assumption does not hold, the ATE and CACE differ, but additional information is needed to estimate the difference. In the case where covariate information is available, the usual additional assumption to identify the ATE is that the ATE is the same for compliers and noncompliers *within strata defined by the covariates*. The CACE and ATE are then the same within strata, but the overall CACE weights the stratum effects by the covariate distribution of compliers, and the overall ATE weights the stratum effects by the covariate distribution of compliers and noncompliers. This difference is likely to be minor in many applications. We focus here on the CACE, in order to avoid the need for assumptions about counterfactual conditions. Bang and Davis (2007) compare estimators of the ATE by simulation.

2.2 *The Assumptions of AT, PP, and IV Estimators*

The quantity $\bar{y}_{1c} - \bar{y}_{0c}$ directly estimates the CACE in equation (2), but $\bar{y}_{1c}$ and $\bar{y}_{0c}$ are not observed, and additional assumptions are needed to identify the estimate. One possibility is to assume

$$\text{NCEC}_\mu : \mu_{0c} = \mu_{0n}, \ \text{NCET}_\mu : \mu_{1c} = \mu_{1a}, \tag{3}$$

which asserts that the mean outcome under the control treatment is the same for compliers and never-takers ("no compliance effect for controls," or NCEC), and the mean outcome under the active treatment is the same for compliers and always-takers ("no compliance effect for treatment," or NCET). Conditions (3) for the means are implied by the conditional independence assumptions

$$\begin{aligned} &\text{NCEC} : [Y \wedge C \,|\, C = n \text{ or } c, R = 0], \\ &\text{NCET} : \ [Y \wedge C \,|\, C = a \text{ or } c, R = 1], \end{aligned} \tag{4}$$

where the symbol $\wedge$ denotes independence. White (2005) calls deviations from these assumptions "selection effects." Under NCEC_μ and NCET_μ, it is natural to estimate both μ_{0c} and μ_{0n} by $\bar{y}_{0(c+n)}$ and both μ_{1c} and μ_{1a} by $\bar{y}_{1(c+a)}$, yielding the PP estimate

$$\hat{\delta}_{\text{PP}} = \bar{y}_{1(c+a)} - \bar{y}_{0(c+n)}, \tag{5}$$

of the CACE. The problem is that the underlying NCEC_μ and NCET_μ assumptions are strong and widely viewed as unacceptable, because compliers and never-takers may differ on various unobserved characteristics related to the outcome under the control treatment, and similarly compliers and always-takers may differ on characteristics related to the outcome under the active treatment. White (2005) argues that NCEC_μ and NCET_μ may be plausible in double-blind prevention trials where the active agent has low rates of adverse events, and noncompliance relates to treatment discontinuation, because treatment discontinuation is relatively unlikely to be related to prognosis. NCEC_μ and NCET_μ can be weakened

by adjusting for known covariates, as discussed Section 4, but they remain strong and questionable assumptions that need to be critically evaluated.

A different, potentially more palatable way of identifying the CACE is to note that participants in the subpopulation of never-takers ($C = n$) are randomly assigned to treatment or control, and in both cases they receive $T = 0$. Similarly always-takers ($C = a$) receive $T = 1$ whether assigned to treatment or control. If outcomes do not depend on the treatment assigned in these two subpopulations (e.g., Baker and Kramer, 2005), then the ER assumption follows:

$$\text{ER}: [Y \wedge R \,|\, C = n]; \;\; [Y \wedge R \,|\, C = a]. \tag{6}$$

The term ER originates in the econometric literature (e.g., AIR), although Greenland (2000) cautions that equation (6) differs from other ERs in that independence is defined within principal compliance strata. The ER implies that the means in Table 1B are such that:

$$\text{ER}_\mu: \mu_{0n} = \mu_{1n}; \;\; \mu_{0a} = \mu_{1a}. \tag{7}$$

The label ER_μ denotes "exclusion restriction for means"; and is weaker than ER because it equates the means rather than the full distribution. The ER assumptions equation (6) or (7) are generally considered more plausible than NCEC and NCET, but they remain assumptions, because the outcome may be affected by whether treatment or control is assigned even though the resulting treatment remains the same, particularly in trials of behavioral interventions. Under ER or ER_μ, $\bar{y}_{1n}$ is an unbiased estimate of both μ_{0n} and μ_{1n}, and $\bar{y}_{1a}$ is an unbiased estimate of both μ_{0a} and μ_{1a}. These estimates lead to the following estimate of the CACE, which is consistent under ER or ER_μ because the numerator and denominator are unbiased estimates of their respective estimands:

$$\hat{\delta}_{\text{IV}} = (\bar{y}_{1+} - \bar{y}_{0+})/(1 - \hat{\pi}_a - \hat{\pi}_n), \tag{8}$$

where $\hat{\pi}_a = m_{0a}/m_{0+}$ and $\hat{\pi}_n = m_{1n}/m_{1+}$ estimate the proportions of always-takers and never-takers. Equation (8) is sometimes termed the IV estimate, because it is has the form of an IV estimate with the randomization indicator as the instrument. Because under ER the treatment effect is zero for the always-takers and never-takers, $\hat{\delta}_{\text{IV}}$ inflates the ITT estimate $\bar{y}_{1+} - \bar{y}_{0+}$ by the estimated proportion $1 - \hat{\pi}_a - \hat{\pi}_n$ of compliers (Baker and Lindeman, 1994; AIR).

Suppose we assume NCEC_μ, NCET_μ, and ER_μ simultaneously:

$$\begin{aligned} &\text{NCEC}_\mu + \text{ER}_\mu: \mu_{0n} = \mu_{1n} = \mu_{0c} = \mu_0, \\ &\text{NCET}_\mu + \text{ER}_\mu: \mu_{1a} = \mu_{0a} = \mu_{1c} = \mu_1, \end{aligned} \tag{9}$$

or the corresponding conditional independence assumptions NCEC + ER, NCET + ER. The natural estimates of μ_0 and μ_1 pool the data for all cases according to treatment received, yielding the AT estimator of the CACE:

$$\begin{aligned} &\hat{\delta}_{\text{AT}} = \bar{y}_1 - \bar{y}_0, \;\; \bar{y}_1 = \frac{m_{1(c+a)}\bar{y}_{1(c+a)} + m_{0a}\bar{y}_{0a}}{m_{1(c+a)} + m_{0a}}, \\ &\bar{y}_0 = \frac{m_{0(c+n)}\bar{y}_{0(c+n)} + m_{1n}\bar{y}_{1n}}{m_{0(c+n)} + m_{1n}}. \end{aligned} \tag{10}$$

To summarize, the NCEC/NCET assumptions lead to $\hat{\delta}_{\text{PP}}$, the ER assumption leads to $\hat{\delta}_{\text{IV}}$ and the combined NCEC/NCET and ER assumptions lead to $\hat{\delta}_{\text{AT}}$. Note that all these estimators are moment-based estimators and the underlying assumptions for consistency only require assumptions about the cell means, not the full distributions. In particular, the IV estimate does not require an assumption of homogeneous treatment effects, under ER and the assumption of monotonicity discussed in Section 2.1. (cf. Brookhart and Schneeweiss, 2007) The choice between the estimators rests largely on the perceived validity of their underlying assumptions, although the precision of the estimates may also play a secondary role.

2.3 *Precision of the AT, PP, and IV Estimators under the ER Assumption*

Because the ER assumption is often plausible, we compare biases and variances of the estimates under that assumption. Under these conditions, the large-sample biases of the ITT, ER, PP, and AT estimates for the CACE are $B(\hat{\delta}_{\text{ITT}}) = -(1 - \pi_c)\text{CACE}$, $B(\hat{\delta}_{\text{IV}}) = 0$, $B(\hat{\delta}_{\text{PP}}) = \pi_n(\pi_n + \pi_c)^{-1}\Delta_0 - \pi_a(\pi_a + \pi_c)^{-1}\Delta_1$, and $B(\hat{\delta}_{\text{AT}}) = \pi_n\left((1-\alpha)\pi_c + \pi_n\right)^{-1}\Delta_0 - \pi_a\left(\alpha\pi_c + \pi_a\right)^{-1}\Delta_1$, where $\Delta_0 = (\mu_{0c} - \mu_{0n})$ and $\Delta_1 = (\mu_{1c} - \mu_{1a})$. Thus IV is unbiased, ITT is attenuated even when NCEC and NCET hold, and the bias of PP and AT depend on the relative sizes of Δ_0 and Δ_1.

Large sample variances of the estimates are obtained by expressing them in the form $g(\bar{y}_{1(c+a)}, \bar{y}_{1n}, \bar{y}_{0(c+n)}, \bar{y}_{0a}, \hat{\pi}_n, \hat{\pi}_a)$, where $(\bar{y}_{1(c+a)}, \bar{y}_{1n}, \bar{y}_{0(c+n)}, \bar{y}_{0a})$ are asymptotically independent, and applying the delta method. Because the resulting expressions are complex and not very insightful, we examine variance and root MSE in more detail for the case when $\pi_a = 0$, that is, there are no always-takers, as in our example. Assuming a constant within-cell variance σ^2, we obtain

$$\text{Var}(\hat{\delta}_{\text{ITT}}) = \frac{\sigma^2 + \pi_c(1-\pi_c)\left(\alpha\Delta_0^2 + (1-\alpha)(\text{CACE} + \Delta_0)^2\right)}{m\alpha(1-\alpha)}, \tag{11}$$

$$\text{Var}(\hat{\delta}_{\text{IV}}) = \frac{1}{m\alpha(1-\alpha)\pi_c^2}\left(\sigma^2 + \pi_c(1-\pi_c)\Delta_0^2\right), \tag{12}$$

$$\text{Var}(\hat{\delta}_{\text{PP}}) = \frac{1}{m\alpha(1-\alpha)\pi_c}\left((\alpha\pi_c + 1 - \alpha)\sigma^2 + \alpha\pi_c^2(1-\pi_c)\Delta_0^2\right), \tag{13}$$

$$\text{Var}(\hat{\delta}_{\text{AT}}) = \frac{1}{m}\left(\frac{\sigma^2}{1-\alpha\pi_c} + \frac{\sigma^2}{\alpha\pi_c} + \frac{(1-\alpha)\pi_c(1-\pi_c)\Delta_0^2}{(1-\alpha\pi_c)^4} \times \left(1 - 2\alpha\pi_c + \alpha\pi_c^2\right)\right). \tag{14}$$

It is easy to show that $\text{Var}(\hat{\delta}_{\text{IV}}) \geq \text{Var}(\hat{\delta}_{\text{PP}}) \geq \text{Var}(\hat{\delta}_{\text{AT}})$ when $\Delta_0 = 0$. When $\Delta_0 > 0$, $\text{Var}(\hat{\delta}_{\text{IV}}) \geq \text{Var}(\hat{\delta}_{\text{PP}})$ and $\text{Var}(\hat{\delta}_{\text{IV}}) \geq \text{Var}(\hat{\delta}_{\text{AT}})$ when $\alpha\pi_C \leq 0.5$; it is theoretically possible for $\text{Var}(\hat{\delta}_{\text{AT}})$ to exceed $\text{Var}(\hat{\delta}_{\text{IV}})$ or $\text{Var}(\hat{\delta}_{\text{PP}})$ for large Δ_0, but we expect it to be smaller in most realistic settings. Details are provided in a Web Appendix.

IV is markedly less efficient than PP and AT for small values of π_c. For example, when $\mu_{0n} = \mu_{1n} = 0$, $\mu_{0c} = 1$, $\mu_{1c} = 2$, $\sigma^2 = 1$, $\alpha = 0.5$, the asymptotic relative efficiencies $\text{Var}(\hat{\delta}_{\text{PP}})/\text{Var}(\hat{\delta}_{\text{IV}})$ and $\text{Var}(\hat{\delta}_{\text{AT}})/\text{Var}(\hat{\delta}_{\text{IV}})$ are about 0.8 when $\pi_c = 0.9$, but only about 0.4 when $\pi_c = 0.6$. Figures 1 and 2 plot the MSE ($\text{MSE} = B^2(\hat{\delta}) + \text{Var}(\hat{\delta})$) of IV, PP, and AT in units of σ^2 against the compliance rate π_c, for $\alpha = 0.52$, two

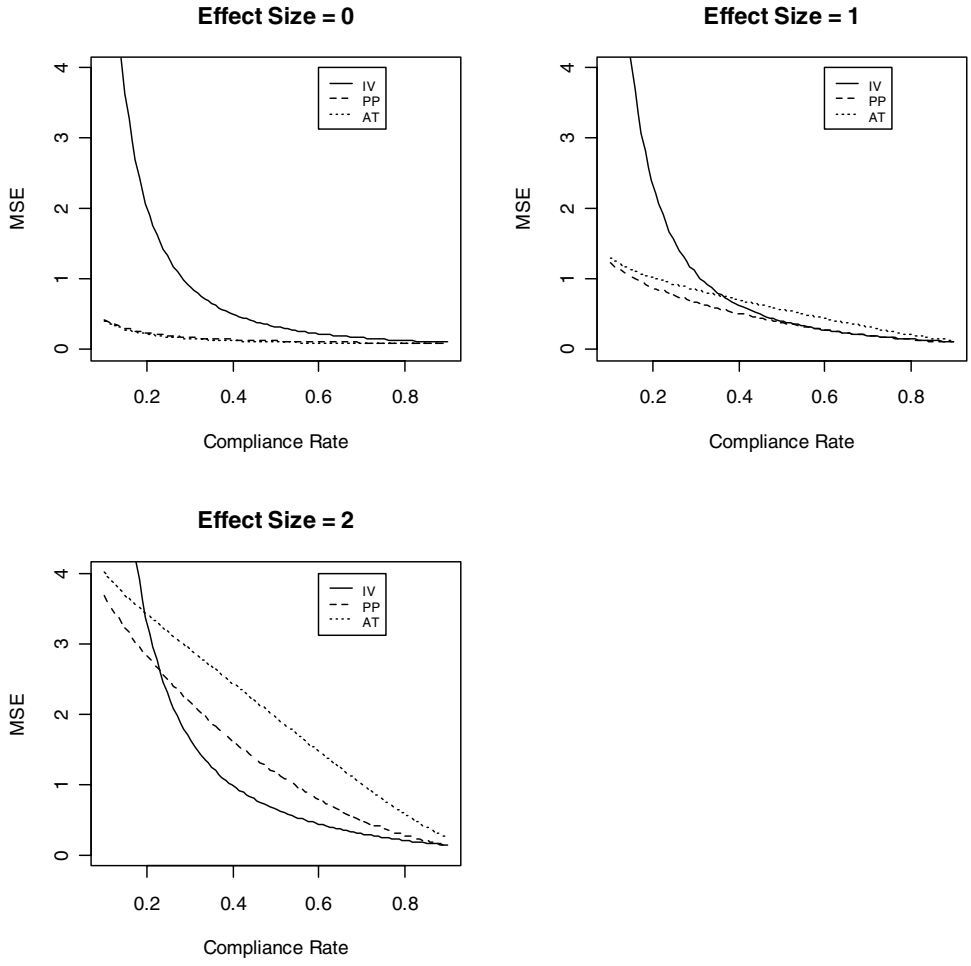

Figure 1. MSE of IV, PP, and AT estimates plotted against compliance rate with $\alpha = 0.522$, sample size $m = 50$, and effect size $\Delta = \Delta_0/\sigma = 0$, 1, 2, respectively.

sample sizes, and various choices of $\Delta = \Delta_0/\sigma$. In Figure 1, where the sample size is small ($m = 50$), AT has lower or comparable RMSE to IV unless Δ is high ($\Delta = 2$). In Figure 2, where the sample size is large ($m = 250$), IV tends to have superior RMSE unless the compliance rate is low or Δ is small or zero. The implication is that the IV estimate becomes an increasingly attractive alternative to the PP and AT estimates as the sample size increases.

If one of either NCEC or ER is assumed true, the other assumption can be tested empirically by comparing $\bar{y}_{10} - \bar{y}_{0+}$ with zero; because IV and AT both assume ER, one might increase the efficiency of the CACE estimate by choosing AT over IV if this test is not rejected, or the difference $\bar{y}_{10} - \bar{y}_{0+}$ is "small." This approach has most appeal when the compliance rate is low, because in this case $\hat{\delta}_{\text{IV}}$ has substantially higher variance and the power of the test may be reasonable; when compliance is high the utility of the test is compromised by low power. An indirect approach to checking the ER assumption using covariates is discussed in Section 3.

What fraction α of cases should be assigned to the treatment group for optimal efficiency? Differentiating equation (12) with respect to α, the variance of $\hat{\delta}_{\text{IV}}$ is minimized when $\alpha = 0.5$, that is, an equal allocation of treated and control cases. This is contrary to the intuition that given noncompliance, more cases should be assigned to the treatment group. On the other hand, the PP variance (13) is minimized when $\alpha = (1 + \sqrt{\pi_c(1 + \Delta^2\pi_c(1 - \pi_c))})^{-1}$. This equals $1/(1 + \sqrt{\pi_c})$ when $\Delta = 0$, which does assign more cases to the treatment group when there is noncompliance.

2.4 *Model-Based Estimators of the CACE*

The high variance of the IV estimator when the compliance rate is low motivates a search for more efficient estimators. Technically all the estimators considered so far can be viewed as method-of-moment estimators under the various assumptions. Another approach to inference is maximum likelihood (ML) estimation based on a model for the joint distribution of Y, R, and C. For example,

$$(y_i \,|\, r_i = r, c_i = j) \sim_{\text{ind}} N(\mu_{rj}, \sigma^2);\ (c_i \,|\, r_i = r) \sim_{\text{ind}} \text{MNOM}(\pi),$$

where $N(\mu_{rj},\ \sigma^2)$ denotes a normal distribution with mean μ_{rj} and variance σ^2, and MNOM (π) denotes a multinomial distribution with probabilities $\pi = (\pi_n,\ \pi_c,\ \pi_a)$ for never-takers, compliers, and always-takers, respectively. The

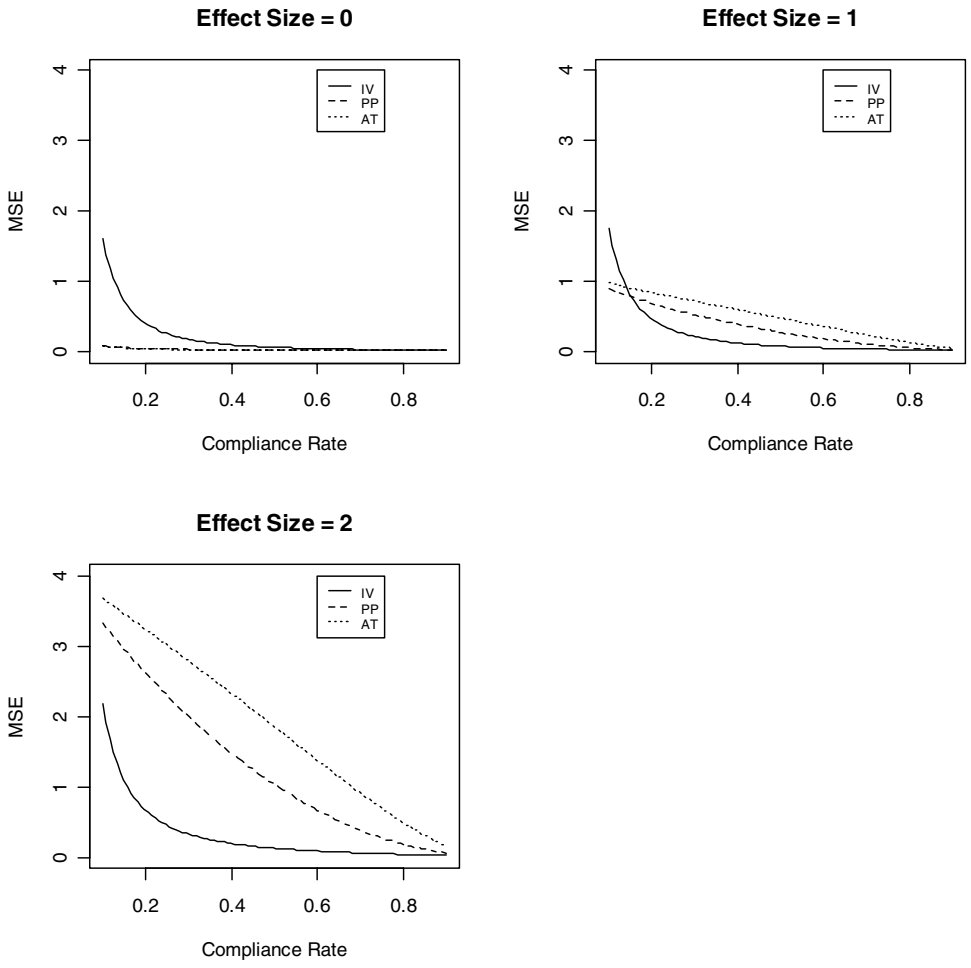

Figure 2. MSE of IV, PP, and AT estimates plotted against compliance rate with $\alpha = 0.522$, sample size $m = 250$, and effect size $\Delta = \Delta_0/\sigma = 0$, 1, 2, respectively.

parameters of this model are identified by restrictions on the means. Specifically, it is easily shown that $\hat{\delta}_{\mathrm{PP}}$ is ML for this model under the NCEC_μ and NCET_μ assumptions (3), and $\hat{\delta}_{\mathrm{AT}}$ is ML under the NCEC_μ, NCET_μ, and ER_μ assumptions (9).

The ML estimate under the ER_μ restriction (7) (ML-ER, denoted as $\hat{\delta}_{\mathrm{MLER}}$) differs from $\hat{\delta}_{\mathrm{IV}}$; it does not have an explicit form, but can be computed using the expectation–maximization algorithm, treating partially observed values of C as missing data (AIR). The estimate $\hat{\delta}_{\mathrm{MLER}}$ is more efficient than $\hat{\delta}_{\mathrm{IV}}$ (Imbens and Rubin, 1997a), but makes stronger distributional assumptions, and is potentially sensitive to violations of assumptions like constant σ^2 (Imbens and Rubin, 1997a, 1997b; Abadie, 2002). For binary outcomes Y with a Bernoulli distribution, $\hat{\delta}_{\mathrm{IV}}$ is the ML estimate of the CACE, providing the resulting means in Table 1B, which are estimated probabilities for a binary Y, all lie between zero and one (Baker and Lindeman, 1994).

3. CACE Estimation of Treatment Efficacy with Covariates

We discuss in this section the choice and role of covariates in improving the AT, PP, and IV estimators. We show that covariates can reduce bias in the AT and PP estimators, but do not reduce bias in the IV estimator, although they can increase precision. We also propose two compliance propensities, and discuss their role in dimension reduction when the set of covariates is extensive.

The AT and PP estimates with no covariates are the estimated coefficients of T in a regression of Y on T, computed using all the cases for AT and cases that take their assigned treatment for PP. Covariates X can be incorporated as covariates in these regressions, and as with covariate adjustments in observational settings, can reduce bias and increase precision. Concerning bias, a correctly specified regression adjustment with covariates X weakens NCEC and NCET in equation (4) to

$$\begin{aligned} &\mathrm{NCECX}: [Y \wedge C \,|\, C = n \text{ or } c, R = 0, X], \\ &\mathrm{NCETX}: \ [Y \wedge C \,|\, C = a \text{ or } c, R = 1, X], \end{aligned} \tag{15}$$

that is, to an assumption of no compliance effects within strata defined by X. Covariates that are predictive of the outcome also increase precision by reducing the residual variance of the regression of Y on X and T. Effect modification can be modeled by including interactions between X and T.

In observational studies, a popular strategy for robust modeling when there are a number of covariates is to stratify on the propensity to take the treatment given covariates X. In randomized studies with full compliance this strategy is not needed, because the assignment propensity is unrelated to X because assignment is randomized. With noncompliance, randomization is compromised, and propensity methods again have a role. We propose here the use of *compliance propensities* in adjusting the AT, PP, and IV estimators. Specifically, to limit bias, a function is sought such that if equation (15) is true, then it remains valid with X replaced by coarsened functions of X. By Rosenbaum and Rubin's (1983) theory of propensity scores, NCECX and NCETX imply

$$\begin{aligned} &\text{NCECP}: [Y \wedge C \,|\, C = n \text{ or } c, R = 0, p_n(X)], \\ &\text{NCETP}: \ [Y \wedge C \,|\, C = a \text{ or } c, R = 1, p_a(X)], \end{aligned} \tag{16}$$

respectively, where $p_n(X) = p(C = n | C = n \text{ or } c,\ X)$ and $p_a(X) = p(C = a | C = a \text{ or } c,\ X)$ are propensities to be never-takers or always-takers. These compliance propensities are the coarsest functions of X for which equation (16) holds. These propensities can be estimated by computing (a) $\hat{p}^*_a(X) = p(C = a \,|\, X)$ from a logistic regression of the binary indicator for always-takers among controls, (b) $\hat{p}^*_n(X) = p(C = n \,|\, X)$ from a logistic regression of the binary indicator for never-takers among those with $R = 1$; and (c) estimating $p_n(X)$ by $\hat{p}_n(X) = \hat{p}^*_n(X)/\left(1 - \hat{p}^*_a(X)\right)$ and $p_a(X)$ by $\hat{p}_a(X) = \hat{p}^*_a(X)/\left(1 - \hat{p}^*_n(X)\right)$. A propensity adjustment then stratifies on both $\hat{p}_n(X)$ and $\hat{p}_a(X)$. Note that the same propensity scores apply for both AT and PP, because the NCEC and NCET assumptions are shared by both methods. AT also requires the ER assumption, but that is not weakened by the propensity adjustment, as we shall see below.

Covariates do not play a role in bias reduction for the IV method. To see this, suppose the ER assumption (6) is assumed conditional on X, that is

$$\text{ER}(X): [Y \wedge R \,|\, X, C = n], \ \ [Y \wedge R \,|\, X, C = a]. \tag{17}$$

Then in terms of densities, for never-takers:

$$\begin{aligned} &p(Y \,|\, R = 1, C = n) \\ &\quad = \int p(Y \,|\, X, R = 1, C = n) p(X \,|\, R = 1, C = n)\, dX \\ &\quad = \int p(Y \,|\, X, R = 0, C = n) p(X \,|\, R = 0, C = n)\, dX \\ &\quad = p(Y \,|\, R = 0, C = n), \end{aligned}$$

where the first and last equalities are by definition, and the middle is implied by ER(X) and the randomization of treatments. Similarly $p(Y|R = 1,\ C = a) = p(Y|R = 0,\ C = a)$. Thus ER($X$) implies ER, and bias from failure of the ER assumption is not reduced by conditioning on the covariates X.

Covariates can be used to increase the *precision* of $\hat{\delta}_{\text{IV}}$, however. For a single categorical X with J categories, let $\hat{\delta}_{\text{IV},k} = (\bar{y}_{1+k} - \bar{y}_{0+k})/(1 - \hat{\pi}_{ak} - \hat{\pi}_{nk})$ be the IV estimator within the stratum $X = k$. Let p_k be the proportion of cases in stratum k, estimated from the pooled sample. Then a stratified estimate of the proportion of principal compliers in stratum k is

$$p_k(1 - \hat{\pi}_{ak} - \hat{\pi}_{nk}) \Big/ \sum_{l=1}^{K} p_l(1 - \hat{\pi}_{al} - \hat{\pi}_{nl}).$$

Weighting $\hat{\delta}_{\text{IV},k}$ by this estimated proportion and summing yields the following stratified IV estimator:

$$\bar{y}_{\text{IV}\,|\,X} = \sum_{k=1}^{K} p_k(\bar{y}_{1+k} - \bar{y}_{0+k}) \Big/ \sum_{k=1}^{K} p_k(1 - \hat{\pi}_{ak} - \hat{\pi}_{nk}), \tag{18}$$

where the numerator is a stratified version of the ITT estimator, with improved precision when X is predictive of the outcome, and the denominator is a stratified version of the overall compliance rate, with improved precision when X is predictive of compliance. We conjecture that the former of these two components has the greater potential for variance reduction. A natural generalization of $\bar{y}_{\text{IV}\,|\,X}$ for a set of categorical and/or continuous X's is

$$\hat{\delta}_{\text{IV}\,|\,X} = \sum_i (\hat{y}_{1i} - \hat{y}_{0i}) \Big/ \sum_i \hat{\pi}_i, \tag{19}$$

where the summation is over all individuals i in the sample; $\hat{y}_{ri}$ is the predicted outcome for unit i if randomized to treatment r, computed from a regression of Y on X and R; and $\hat{\pi}_i$ is the predicted true compliance for unit i, computed from a regression of C on X estimated from the cases assigned to treatment. The latter should be of a form appropriate for a binary outcome, for example logistic or probit regression. Comparisons of equation (18) or (19) with the structural equations approach of Nagelkerke et al. (2000) would be of interest.

An alternative to equation (18) or (19) is to compute ML or Bayes' estimates of the CACE given covariates X, using a full model for the distribution of Y and C, given R and X, and treating the unknown principal compliance indicators as missing data (Imbens and Rubin, 1997a; Little and Yau, 1998). For example, one might assume

$$\begin{aligned} &(y_i \,|\, x_i, r_i = r, c_i = j, \theta_Y, \sigma^2) \sim_{\text{ind}} N(\mu(x_i, r_i, c_i; \theta_Y), \sigma^2); \\ &(c_i \,|\, r_i = r, \theta_C) \sim_{\text{ind}} \text{MNOM}(\pi(x_i, \theta_C)), \end{aligned} \tag{20}$$

where the compliance model for C given R and X is a multinomial logistic model with parameters θ_C, and the outcome model for Y given C, R, and X is a linear regression model with regression parameters θ_Y and variance σ^2. In equation (20), $\pi\ (x_i,\ \theta_C)$ excludes effects involving r_i, because treatment assignment is randomized. If the ER (17) is assumed to hold within all subpopulations defined by the covariates, the mean $\mu\ (x_i,\ r_i,\ c_i;\ \theta_Y)$ is subject to restrictions $\mu\ (x_i,\ r_i,\ c_i = j;\ \theta_Y) = \mu\ (x_i,\ c_i = j;\ \theta_Y)$ when $j = a$ or n. The CACE in the subpopulation defined by x_i in this model is

$$\mu(x_i, r_i = 1, c_i = c;\ \theta_Y) - \mu(x_i, r_i = 0, c_i = c;\ \theta_Y). \tag{21}$$

Interactions between C, R, and baseline covariates can be included in the mean function, but then the CACE (21) is no longer unique but varies according to the value of x_i. See also Bond, White, and Walker (2007). The modeling approach yields gains of efficiency over equation (19), but is more vulnerable to model misspecification (Imbens and Rubin, 1997a, 1997b; Abadie, 2002). More simulations comparing these methods, and extensions of equation (20) that allow the variance σ^2 to vary across the principal strata, would be of interest.

Table 2

WTP study: (A) sample means (sample sizes) for outcome 6-minute walk in control and group treatment subgroups, and (B) observed and predicted means under PP, IV, and AT models

A		Compliance C		
	N	C		
Randomized treatment R	0	?	?	748.90 (122)
	1	694.12(16)	866.01 (105)	843.28 (121)

B

	IV Compliance		PP Compliance		AT Compliance		
Treatment R	N	C	N	C	N	C	ITT
0	*694.12*	*757.25*	*748.90*	*748.90*	*742.55*	*742.55*	748.90
1	*694.12*	866.01	694.12	866.01	*742.55*	866.01	843.28
CACE		108.76		117.11		123.45	94.38
(SE)		(65.53)		(58.97)		(57.37)	(57.08)

Wenoted in Section 2 that the data do not provide direct evidence of the validity of ER or NCEC/NCET. The same comment applies within strata defined by the covariates X. However, if the covariates are good predictors of never-takers in the treatment group, the relationship between propensity to be a never-taker and outcome can be assessed in the control group, and lack of evidence of a relationship might be construed as indirect evidence in favor of the NCEC assumption, subject to the caveat that this analysis misses effects of unmeasured confounders. Specifically, transform the covariates X into the propensity score $\hat{p}_n(X)$ and covariates Z orthogonal to $\hat{p}_n(X)$, and regress Y on $\hat{p}_n(X)$ and Z in the control group. If the coefficient of $\hat{p}_n(X)$ in this regression is small, this provides some justification for the NCEC assumption, suggesting that PP or AT analysis may be reasonable options. On the other hand if the regression coefficient of $\hat{p}_n(X)$ is large, estimates like $\hat{\delta}_{\text{IV}}$ or $\hat{\delta}_{\text{MLER}}$ that do not require the NCEC assumption may be preferable. A similar analysis can be applied to the NCET assumption for compliers and always-takers.

4. Application to the Women Take Pride Study

We illustrate the various efficacy estimates with data from the WTP study. As discussed in Section 1, we restrict attention to randomized subjects, and compare women assigned to the group behavioral intervention ($R = 1$), where intervention classes are taken in a group, with the control "usual care" treatment ($R = 0$). In this example, individuals assigned to control do not have access to the group treatment, so there are no always-takers, and C reduces to a two-category variable consisting of compliers and never-takers.

Table 2A shows the observed counts and means for the outcome "6-minute walk" taken at month 12, measuring the distance in feet an individual can walk in 6 minutes. We exclude 69 of the 190 subjects in the intervention group and 62 of 184 subjects in the control group who drop out before month 12. We assume that drop out is random within each treatment group, after conditioning on any covariates included in the analysis. In the treatment group $\hat{\pi}_c = 105/121 = 86\%$ complied with treatment, where compliance is defined here as completion of at least one of the treatment modules. We compare the IV, PP, and AT estimators of intervention efficacy without and with covariates.

As discussed in Section 2, the IV estimator requires the ER assumption, while the PP analysis requires the NCEC assumption and the AT analysis requires the ER + NCEC assumption. In pharmaceutical clinical trials, where treatments are blinded and the efficacy is based on the properties of the drugs, the ER assumption often seems more reasonable than NCEC, so the IV methods are attractive. For behavioral intervention trials, such as the WTP study, it is not clear that ER is superior to NCEC, particularly if outcomes are thought to be affected by the psychological effects of failing to comply with an assigned treatment, because these effects are in play when an individual is assigned to the treatment but not when an individual is assigned to the control. Faced with uncertainty about which assumption is appropriate, one strategy is to compute both PP and IV estimators and assess whether treatment effects are robust to these alternative assumptions.

We first present analyses without covariates. Table 2B shows the estimated cell means from the IV, PP, AT, and ITT methods, and associated standard errors (SEs); the estimated means in italics are those implied by the assumptions of each method. Note that the ITT estimate is smaller than the others because it is attenuated, and it has the smallest SE. The estimates of the CACE from the other three methods differ somewhat, although not in relation to the SEs, and the PP and AT estimates are statistically significant at the 0.05 alpha level. The SEs order as described in Section 2.3, with relatively small differences because the compliance rate is quite high in this application. The t-test for the combined NCEC + ER assumptions compares the mean for controls (748.90) with the mean for treatment noncompliers (694.12), not statistically significant even though the difference in means is quite substantial. This illustrates the low power of the test when the compliance rate is high.

We next include the following covariates in our analysis of the WTP data: age, employment, and symptom impact

Table 3
WTP study: ML estimates of regression coefficients of models for the outcome 6*-minute walk and compliance with covariates*

Parameter	Estimate (B)	SE(B)	P-value
Outcome model (1)			
Intercept	707.97	278.89	<0.001
Age	−6.763	3.404	0.032
Baseline 6-minute walk	0.707	0.048	<0.001
Employment	28.71	55.18	0.59
Treatment compliance	−29.72	84.87	0.83
Compliance × treatment	97.20	43.73	0.043
Compliance model (2)			
Intercept	2.259	0.497	<0.001
Employment	−0.464	0.792	0.526
Baseline SIP physical	−0.040	0.041	0.282

Table 4
WTP study: IV, ML-ER, PP, AT, and ITT estimates for the outcome 6*-minute walk, with and without covariate adjustment*

Covariates		IV	ML-ER	PP	AT	ITT
Unadjusted	Estimate	108.76	109.78	117.11	123.45	94.38
	SE	65.53	61.28	58.97	57.37	57.08
Adjusted	Estimate	100.05[a]	97.20	95.58	90.13	86.39
	SE	44.25[a]	43.73	40.30	38.74	38.39

[a]The IV estimate with covariate adjustments is obtained using (19), with baseline variables including employment status and SIP physical scores used to estimate propensity scores $\hat{\pi}_i$ for each subject. Its SE is computed using the bootstrap.

profile (SIP) physical score, which measures a subject's physical functioning (Janevic et al., 2003). The ML estimate of the CACE for the data without adjusting for covariates is $\hat{\delta}_{\text{MLER}} = 109.78$, closeto the IV estimate of $\hat{\delta}_{\text{IV}} = 108.76$ in Table 2. It has a slightly lower SE (61.28 versus 65.53). Table 3 shows the results of fitting a model of form (20) to the data from the WTP study. Block (1) shows the coefficients for the outcome model, with the CACE being the compliance × treatment interaction. Block (2) shows coefficients from the compliance model. The covariate-adjusted ML estimate of the CACE from this model is 97.20 with a reduced SE, namely 43.73, indicating some improvement in precision from adjusting for the covariates. In addition, women with higher baseline measure tend to have higher 6-minute walk measure at month 12, and older women tend to have lower 6-minute walk measure at month 12.

Table 4 summarizes the IV, ML-ER, PP, AT, and ITT estimates both with and without covariate adjustments. ITT, PP, and AT estimates with covariate adjustments are obtained using linear regression adjusting for age, baseline 6-minute walk measurement, and employment status. The IV estimate with covariate adjustments is obtained using equation (19), and baseline variables including employment status and SIP physical scores are used to estimate propensity scores $\hat{\pi}_i$ for each subject. Table 4 shows that covariate adjustment generally improves precision and reduces differences between the methods, except ITT which remains attenuated. The large gain in efficiency may be due to the significant effects of age and baseline measurements on the outcome of interest. We note in particular that covariate adjustment of IV based on equation (18) results in 50% reduction in estimated SE, and a significant treatment effect. The results suggest that compliers performed better under treatment rather than under control for this outcome.

5. Discussion

We have compared a variety of methods for estimating efficacy in randomized trials for a control and active treatment, when there is all-or-nothing noncompliance. In practice, the choice of methods depends on various factors, effect sizes relative to between-subject variability of the outcome measure, sample size, and differences in characteristics of compliers and noncompliers. The choice also depends on the plausibility of the different modeling assumptions and the trade-off between efficiency and robustness. If NCEC/NCET or NCEC/NCET + ER given a set of covariates can be believed, and regressions on the covariates are correctly specified, then AT can be dramatically more efficient than IV. On the other hand, the IV estimate of the CACE under ER is robust against misspecification of our regression model or false belief in NCEC. Thus, it may be wise to compute and compare all the estimates to assess sensitivity of answers to the choice of method.

Many generalizations and extensions of these methods are possible. Extensions that require more restrictions to identify the parameters include the case where partial compliance is modeled (Goetghebeur and Molenberghs, 1996), or there are more than two treatments, such as two active treatments and a control treatment that are assumed to apply to noncompliers. CACE estimation in trials involving a control group and more than one treatment groups is more complicated, with more principal compliance categories and more complicated identifiability assumptions. Results on this case will be reported elsewhere. Another extension is to joint models for noncompliance and missing data—for simplicity we confined our analyses of 12-month WTP data to completers. For some approaches to this issue see Frangakis and Rubin (1999) and Peng, Little, and Raghunathan (2004).

6. Supplementary Materials

The Web Appendix referenced in Section 2.3 is available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

This research was supported by grant R01CA76404 from the National Cancer Institute. We thank Noreen Clark for kindly providing the data from the WTP Study that illustrate the methods, and an associate editor and two referees for very helpful comments.

References

Abadie, A. (2002). Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American Statistical Association* **97,** 284–292.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion and

rejoinder). *Journal of the American Statistical Association* **91,** 444–472.

Baker, S. G. (1997). Compliance, all-or-none. In *The Encyclopedia of Statistical Science*, S. Kotz, C. R. Read, and D. L. Banks (eds), Update Volume 1, 134–138. New York: John Wiley and Sons.

Baker S. G. (2000). Analyzing a randomized cancer prevention trial with a missing binary outcome, an auxiliary variable, and all-or-none compliance. *Journal of the American Statistical Association* **95,** 43–50.

Baker, S. G. and Kramer, B. S. (2005). Simple maximum likelihood estimates of efficacy in randomized trials and before-and-after studies, with implications for meta analysis. *Statistical Methods in Medical Research* **14,** 605 and Correction (Vol. 14, p. 349, 2005).

Baker, S. G. and Lindeman, K. S. (1994). The paired availability design: A proposal for evaluating epidural analgesia during labor. *Statistics in Medicine* **13,** 2269–2278.

Bang, H. and Davis, C. E. (2007). On estimating treatment effects under non-compliance in randomized clinical trials: Are intent-to-treat or instrumental variables analyses perfect solutions? *Statistics in Medicine* **26,** 954–964.

Bond, S. J., White, I. R., and Walker, A. S. (2007). Instrumental variables and interactions in the causal analysis of a complex clinical trial. *Statistics in Medicine* **26,** 1473–1496.

Brookhart, M. A. and Schneeweiss, S. (2007). Preference-based instrumental variable methods for the estimation of treatment effects: Assessing validity and interpreting results. *International Journal of Biostatistics* **3,** 14.

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

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

Goetghebeur, E. and Molenberghs, G. (1996). Causal inference in a placebo-controlled clinical trial with binary outcome and ordered compliance. *Journal of the American Statistical Association* **91,** 928–934.

Greenland S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology* **29,** 722–729.

Imbens, G. W. and Rubin, D. B. (1997a). Estimating outcome distributions for compliers in instrumental variables models. *Review of Economic Studies* **64,** 555–574.

Imbens, G. W. and Rubin, D. B. (1997b). Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics* **25,** 305–327.

Janevic, M. R., Janz, N. K., Dodge, J. A., Lin, X., Pan, W., Sinco, B. R., and Clark, N. M. (2003). The role of choice in health education intervention trials: A review and case study. *Social Science and Medicine* **56,** 1581–1594.

Little, R. J. A. and Yau, L. (1998). Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. *Psychological Methods* **3,** 147–159.

Nagelkerke, N., Fidler, V., Bernsen, R., and Borgdor, M. (2000). Estimating treatment effects in randomized clinical trials in the presence of non-compliance. *Statistics in Medicine* **19,** 1849–1864.

Peng, Y., Little, R. J., and Raghunathan, T. (2004). An extended general location model for causal inferences from data subject to non-compliance and missing values. *Biometrics* **60,** 598–608.

Robins, J. M. (1989). The analysis of randomized and nonrandomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service Research Methodology: A Focus on AIDS*, L. Sechrest, H. Freeman, and A. Mulley (eds), 113–159. Washington, DC: U.S. Public Health Service.

Robins, J. M. and Greenland, S. (1996). Discussion of "Identification of causal effects using instrumental variables" by J. D. Angrist, G. W. Imbens, and D. B. Rubin. *Journal of the American Statistical Association* **91,** 456–458.

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

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology* **66,** 688–701.

Rubin, D. B. (1977). Assignment to a treatment group on the basis of a covariate. *Journal of Educational Statistics* **2,** 1–26.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics* **6,** 34–58.

White I. R. (2005). Uses and limitations of randomization-based efficacy estimators. *Statistical Methods in Medical Research* **14,** 327–347.

Received June 2006. *Revised March* 2008.
Accepted March 2008.