Received: 8 November 2016 | Revised: 24 July 2017 | Accepted: 3 August 2017

DOI: 10.1002/sim.7447

RESEARCH ARTICLE

WILEY Statistics in Medicine

A synthetic estimator for the efficacy of clinical trials with all-or-nothing compliance

Joseph Antonelli[1] | Bing Han[2] | Matthew Cefalu[2]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health, 655 Huntington Avenue, Boston, MA 02115, U.S.A.

[2]RAND Corporation, 1776 Main Street, Santa Monica, CA 90401, U.S.A.

Correspondence
Joseph Antonelli, Department of Biostatistics, Harvard T.H. Chan School of Public Health, 655 Huntington Avenue, Boston, MA 02115, U.S.A.
Email: jla538@mail.harvard.edu

Funding information
National Institutes of Health, Grant/Award Number: ES000002, ES024332, ES007142, ES026217 P01CA134294, R01GM111339, R35CA197449 and P50MD010428, R01DA040721; National Institute of Drug Abuse

A critical issue in the analysis of clinical trials is patients' noncompliance to assigned treatments. In the context of a binary treatment with all or nothing compliance, the intent-to-treat analysis is a straightforward approach to estimating the effectiveness of the trial. In contrast, there exist 3 commonly used estimators with varying statistical properties for the efficacy of the trial, formally known as the complier-average causal effect. The instrumental variable estimator may be unbiased but can be extremely variable in many settings. The as treated and per protocol estimators are usually more efficient than the instrumental variable estimator, but they may suffer from selection bias. We propose a synthetic approach that incorporates all 3 estimators in a data-driven manner. The synthetic estimator is a linear convex combination of the instrumental variable, per protocol, and as treated estimators, resembling the popular model-averaging approach in the statistical literature. However, our synthetic approach is nonparametric; thus, it is applicable to a variety of outcome types without specific distributional assumptions. We also discuss the construction of the synthetic estimator using an analytic form derived from a simple normal mixture distribution. We apply the synthetic approach to a clinical trial for post-traumatic stress disorder.

KEYWORDS
causal inference, clinical trials, model averaging, noncompliance, principle stratification

1 | INTRODUCTION

Patients in randomized controlled clinical trials (RCTs) may not comply with their assigned treatments, where compliance can be defined and measured in different ways.[1] Often times, a dichotomized compliance status is used to simplify the actual compliance level of a patient, eg, whether a patient has taken enough dosage of a medication. Such a dichotomized compliance status is customarily referred to as "all-or-nothing," although the actual compliance may be more complicated than full compliance versus no compliance. Throughout this paper, we focus on the dichotomous compliance among patients. The proportion of patients who have positive compliance status, ie, the *observed compliance rate*, can vary widely among studies even under similar settings. For example, in 3 comparable clinical trials of coordinated care interventions for post-traumatic stress disorder (PTSD),[2-4] the observed compliance rates were 73%,[2] 89%,[3] and 96%.[4] In another RCT for PTSD with 3 therapeutic intervention arms and a control arm,[5] the observed compliance rate was between 43% and 65% in the treatment arms.

In the presence of noncompliance, *effectiveness* is defined as the average treatment effect of the treatment assignment, and *efficacy* usually refers to the treatment effect given that the treatment is indeed taken. Estimating the overall

effectiveness of a trial is straightforward by the intent-to-treat (ITT) analysis.[6] The effectiveness of a trial can also be estimated for subgroups of patients defined by pretreatment characteristics by implementing an ITT analysis within each subgroup. The ITT analysis is the gold standard in drug approval processes and is arguably the most appropriate for analyzing pragmatic trials. Patients' compliance is ignored in an ITT analysis.

In contrast, conceptualization and analysis for the efficacy of a trial are far more complicated. Three estimation approaches are frequently applied in the scientific literature: as-treated (AT) analysis, per-protocol (PP) analysis, and instrumental variable (IV) estimation. Roughly speaking, the AT analysis regroups patients according to the treatment actually received and ignores the original randomization.[7] The PP analysis uses the original random assignments to group patients but ignores patients who do not comply with the randomization.[7] Both the AT and PP analyses can be biased, because compliance decisions can be due to various sample selection processes. The IV approach considers random assignments as an instrument to the actual treatment received. In econometric terms, an instrument is a variable that relates to the outcome only through the treatment variable.[8] Hence, an instrument must be related to the treatment and independent of the outcome conditioning on the treatment and perhaps other covariates. In an RCT, randomization may be seen as an instrument, whose effect is assumed to realize only through the actual treatment status.

Little et al reported that no single estimator among IV, AT, and PP can achieve the best statistical efficiency under all possible scenarios.[9] On the contrary, each method may have poor statistical efficiency in a given scenario relative to the estimator with the smallest mean squared error (MSE). Under its own assumptions, the IV approach is free from selection biases but can yield inefficient estimates compared with AT and PP analyses. The lack of efficiency of the IV estimator is particularly remarkable when the compliance rate is relatively low or the selection bias is relatively light, resulting in MSEs multiple-fold higher than those of the AT and PP estimators. Nevertheless, when the selection bias is severe and the compliance rate is high, the IV approach can be the most efficient one among all 3 approaches. The relative efficiency among the 3 approaches varies dramatically in different scenarios.

In this paper, we propose a synthetic estimator to combine the AT, PP, and IV estimators. Hereafter, we call the original approaches (AT, PP, and IV) as *candidate estimators*. Our synthetic estimator is a special case of the well-known linear model averaging estimator.[10-14] However, we avoid using the term "model averaging" because the 3 candidate estimators considered in this paper and the proposed synthetic estimator are moment estimators without explicit modeling assumptions. Averaging across multiple candidates avoids choosing the "best" candidate, which depends on specific settings, but also avoids choosing the worst candidate. In Section 2, we review the basic assumptions of the 3 candidate estimators, all of which are essentially untestable, making it difficult to locate the "best" candidate estimator without making further assumptions in practice. The basic assumption of the IV estimator is usually considered more plausible than the other 2 candidate estimators in many applications; therefore, we present the synthetic estimator using the IV estimator's assumption. Importantly, the proposed synthetic estimator is equally applicable under an alternative candidate estimator's assumption. We use numerical studies to show that the synthetic estimator has nearly the best efficiency compared to the 3 candidates under a wide range of scenarios. The proposed synthetic estimator seeks to minimize the finite-sample MSE, which is particularly relevant for trials with small sample sizes.

This paper is organized as follows. In Section 2, we review the conceptual framework of potential outcomes and principal stratification in the context of RCTs.[9] Under this framework, efficacy is defined as the complier-average causal effect (CACE), which is the average treatment effect among principal compliers. All three candidate estimators are estimating the CACE based on different moment assumptions. In Section 3, we present the proposed composite estimator and its implementation. We also utilize a simple, normal mixture model to illustrate the weighted average rationale in Section 3. In Section 4, we report extensive numerical studies to examine the performance of the proposed synthetic estimator. In Section 5, we apply the synthetic estimator to estimate the efficacy of the Violence and Stress Assessment (ViStA) study, a randomized controlled clinical trial for PTSD.[2,15] Finally, we conclude in Section 6 with a discussion.

2 | REVIEW OF THE CANDIDATE ESTIMATORS IN THE PRINCIPAL COMPLIANCE FRAMEWORK

We adopt the notation in Little et al.[9] Let R be the treatment assignment ($R = 1$ for active treatment and 0 for control). The *actual treatment* received is subject to all-or-nothing compliance, denoted as $T(R) \in \{0, 1\}$, where $T(R)$ can be different from R. The principal compliance framework assumes that the patient population consists of 3 principal strata, denoted by C: never-takers ($C = n, T(R) = 0$), who take the control condition regardless of their assigned conditions; compliers ($C = c, T(R) = R$), who take the assigned conditions; and always-takers ($C = a, T(R) = 1$), who take the active

TABLE 1 Notation for population means, sample means, and sample sizes under principal stratification

		Principle compliance C			
(A) Population proportions		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					
Randomized treatment R	0	μ_{0a}	μ_{0c}	μ_{0n}	μ_{0+}
	1	μ_{1a}	μ_{1c}	μ_{1n}	μ_{1+}
	ALL	μ_{+a}	μ_{+c}	μ_{+n}	
(C) Observed means (sample counts)					
Randomized treatment R	0	$\bar{Y}_{0a}(m_{0a})$	$\bar{Y}_{0(c+n)}(m_{0(c+n)})$		$\bar{Y}_{0+}(m_{0+})$
	1	$\bar{Y}_{1(c+a)}(m_{1(c+a)})$		$\bar{Y}_{1n}(m_{1n})$	$\bar{Y}_{1+}(m_{1+})$
	ALL	?	?	?	

treatment regardless of their assigned conditions. As in Little et al,[9] we adopt the monotonicity assumption and the stable unit-treatment value assumption (SUTVA). The former assumes no existence of defiers, who always take a condition opposite to the assigned condition, and the latter assumes that compliance and outcomes for individuals are not affected by other individuals in the sample.

By the principal stratification framework,[16] the observed compliance strata consists of principal compliers ($C = c$), always-takers in the treatment arm ($C = a, R = 1$), and never-takers in the control arm ($C = n, R = 0$). Table 1 is the same table as in Little et al[9] that presents the notation for population means, sample means, and sample size under the different principal compliance strata when the proportion of subjects assigned to treatment is α. Let μ_{rj} denote the mean of the outcome Y among those assigned $R = r$ within the compliance stratum $C = j$. Additionally, let $\bar{Y}_{rj}$ and m_{rj} denote the corresponding sample mean and observed sample size. Frequently, we use notation such as $\bar{Y}_{1(c+a)}$, which represents the sample mean among those assigned to treatment who are either compliers or always takers, or $\bar{Y}_{1+}$, which represents the sample mean among those assigned treatment across all principal strata. Finally, let π_j represent the probability of being in compliance stratum $C = j$.

One reasonable way to define efficacy is the CACE, ie, the population mean difference between treated and controls among principal compliers, denoted as θ:

$$\theta = \mu_{1c} - \mu_{0c}. \tag{1}$$

The following assumptions can be used to identify the CACE.

1. No compliance effects on means (NCEM): $\mu_{0c} = \mu_{0n}, \mu_{1c} = \mu_{1a}$.
2. Exclusion restriction for means (ERM): $\mu_{0n} = \mu_{1n}, \mu_{0a} = \mu_{1a}$.

The NCEM assumption implies that principal compliance has no effect given the assigned and actual treatment condition. In contrast, the ERM assumption implies no effect of randomization for always-takers and never-takers. The NCEM assumption is particularly fragile because compliers can be systematically different from always-takers and never-takers in various observed and unobserved characteristics. Such a source for bias is usually referred to as selection bias[8] and is a very legitimate concern in many studies, although White[17] argued that NCEM may hold in some special cases such as double-blind prevention trials. The ERM assumption, on the other hand, can be violated if randomization alone can affect the outcome measure not through the actual treatment conditions. While both assumptions can be violated in reality, the NCEM assumption is generally considered too strong to be plausible. These 2 assumptions can be tested with further assumptions. For example, if we are willing to accept either ERM or NCEM, the other assumption can be empirically tested. If we can accept some distributional assumptions, such as an equal-variance normal distribution for every cell in Table 1 and a known order relationship among μ_{0n}, μ_{0c}, μ_{1a}, and μ_{1c}, a standard normal mixture model suffices to estimate all cell means in Table 1. Thus, the NCEM and ERM assumptions can both be tested. However, without further assumptions, the validity of these two assumptions is largely based on researchers' perception.

Under NCEM, the CACE can be written as:

$$CACE = \frac{\pi_a}{\pi_c + \pi_a}\mu_{1a} + \frac{\pi_c}{\pi_c + \pi_a}\mu_{1c} - \frac{\pi_c}{\pi_c + \pi_n}\mu_{0c} - \frac{\pi_n}{\pi_c + \pi_n}\mu_{0n},$$

which yields the PP estimator:

$$\widehat{\theta}_{pp} = \bar{Y}_{1(c+a)} - \bar{Y}_{0(c+n)}. \tag{2}$$

Under ERM, the CACE can be written as

$$E(\bar{Y}_1) - E(\bar{Y}_0) = \pi_c(\mu_{1c} - \mu_{0c}),$$

which suggests the IV estimator:

$$\widehat{\theta}_{iv} = (\bar{Y}_{1+} - \bar{Y}_{0+})/(1 - \widehat{\pi}_a - \widehat{\pi}_n). \tag{3}$$

Note that the IV estimator has the exact form as if the estimator is constructed using the standard two-stage least squares procedure in the econometric literature.[8] If we assume both the ERM and NCEM assumptions, namely,

$$\mu_{1a} = \mu_{1c} = \mu_{0a}, \mu_{1n} = \mu_{0n} = \mu_{0c},$$

then the CACE is equal to the mean difference between the actual treatment conditions, resulting in the AT estimator

$$\widehat{\theta}_{at} = \frac{m_{1(c+a)}\bar{Y}_{1(c+a)} + m_{0a}\bar{Y}_{0a}}{m_{1(c+a)} + m_{0a}} - \frac{m_{0(c+n)}\bar{Y}_{0(c+n)} + m_{1n}\bar{Y}_{1n}}{m_{0(c+n)} + m_{1n}}. \tag{4}$$

All 3 candidate estimators are moment-based estimators, and their validity only depends on the assumptions governing the means (NCEM, ERM, or both), as well as the fundamental assumptions in the principal compliance framework (monotonicity and SUTVA). Despite relying on different assumptions to obtain unbiasedness, each estimator is estimating the same quantity, the CACE, which allows us to combine estimators without changing the estimand of interest.

3 | THE SYNTHETIC ESTIMATOR FOR ESTIMATING THE CACE

3.1 | A brief review of the synthetic estimator

Let $\widehat{\theta}_0, \widehat{\theta}_1, \ldots, \widehat{\theta}_{k-1}$ be k candidate estimators for the same parameter of interest θ. Denoting the vector of candidate estimators as $\widehat{\boldsymbol{\theta}}$, we can write

$$E\left(\widehat{\boldsymbol{\theta}}\right) = \theta + \mathbf{B}, \quad var\left(\widehat{\boldsymbol{\theta}}\right) = \mathbf{V}. \tag{5}$$

Here, $\boldsymbol{B}$ and $\boldsymbol{V}$ represent the bias and the covariance matrix of the estimators, respectively. A synthetic estimator is a linear convex combination[14] of $\widehat{\boldsymbol{\theta}}$, $\widetilde{\theta} = \sum_{i=0}^{k-1} b_i\widehat{\theta}_i = \mathbf{b'}\widehat{\boldsymbol{\theta}}$, where $0 \leqslant b_i \leqslant 1$, $\sum_{i=0}^{k-1} b_i = 1$. The MSE of the synthetic estimator $\widetilde{\theta}$ is

$$MSE(\widetilde{\theta}) = \mathbf{b'Vb} + (\mathbf{b'B})^2, \tag{6}$$

The ideal synthetic estimator is the one that minimizes the MSE in (6). Longford[14] derived the conditions under which a wide range of synthetic estimators can be more efficient than any candidate estimators when 2 candidate estimators are available. Therefore, not only does a synthetic estimator circumvent the trouble of selecting one candidate but it may also improve the efficiency by combining candidate estimators in an optimal fashion.

However, since **B** and **V** are usually unknown, their corresponding estimates are substituted in (6), yielding an empirical synthetic estimator. The empirical synthetic estimator does not achieve the optimal efficiency due to errors in estimating the unknown moments, but its MSE may still be small compared to all candidates. For example, the empirical best linear unbiased predictor (EBLUP) in linear mixed-effect models is a well-known example following this rationale.[18,19] The empirical synthetic estimator has also been widely applied in the small area estimation problem in the survey methodology literature.[20-23]

Our convex estimator shares much in common with a series of papers[24-26] that derived the synthetic estimator under an affine combination of the estimators. Under affine transformation Judge and Mittelhammer derived a closed-form solution for the synthetic estimator between 2 vector estimators.[24] Moreover, their closed form was further simplified under multivariate normal assumptions in regression settings[24] and then extended to asymptotically normally distributed

estimators.[25] All of these related works can trace their roots to the James-Stein estimator,[27] where a biased estimator can dominate the usual, best unbiased estimator in MSE. While in some settings the affine combination of candidate estimators could provide a smaller MSE, we pursue the convex combination for a number of reasons. The convex estimator is fully nonparametric, allowing us to apply it without parametric assumptions. It can also be extended to any number of candidate estimators without additional development, as we only require an estimate of the covariance between the candidate estimators and the identification of an unbiased estimator. Finally, the convex estimator is easily interpreted as a weighted average (see Table 3 of our case study).

3.2 | The proposed synthetic estimator

The proposed synthetic estimator is a linear convex combination of $\widehat{\theta}_{iv}$, $\widehat{\theta}_{pp}$, $\widehat{\theta}_{at}$. Since the ERM assumption is usually considered plausible, we identify the IV estimator as an unbiased estimator in the remainder of the paper. If, in some studies, the NCEM assumption is deemed more plausible, our development is equally applicable by simply swapping the role of the PP and IV estimators in all formulations below. Using the notation in the previous sections, we have

$$\begin{aligned}\mathbf{B} &= (0, bias(\widehat{\theta}_{pp}), bias(\widehat{\theta}_{at}))',\\ \mathbf{V} &= \begin{pmatrix} var(\widehat{\theta}_{iv}) & cov(\widehat{\theta}_{iv}, \widehat{\theta}_{pp}) & cov(\widehat{\theta}_{iv}, \widehat{\theta}_{at}) \\ cov(\widehat{\theta}_{pp}, \widehat{\theta}_{iv}) & var(\widehat{\theta}_{pp}) & cov(\widehat{\theta}_{pp}, \widehat{\theta}_{at}) \\ cov(\widehat{\theta}_{at}, \widehat{\theta}_{iv}) & cov(\widehat{\theta}_{at}, \widehat{\theta}_{pp}) & var(\widehat{\theta}_{at}), \end{pmatrix}. \end{aligned} \tag{7}$$

Not incurring further assumptions on the outcome measurement, we use the bootstrap method[28] to estimate the covariance matrix $\mathbf{V}$, denoted by $\widehat{\mathbf{V}}$. The bias term $\mathbf{B}$ is estimated by the observed difference between 2 candidate estimators, $\widehat{\mathbf{B}} = (0, \widehat{\theta}_{pp} - \widehat{\theta}_{iv}, \widehat{\theta}_{at} - \widehat{\theta}_{iv})'$. This yields the empirical synthetic estimator, denoted by $\widehat{\theta}_s$

$$\begin{aligned}&\widehat{\theta}_s = \widehat{b_0}\widehat{\theta}_{iv} + \widehat{b_1}\widehat{\theta}_{pp} + \widehat{b_2}\widehat{\theta}_{at},\\ &\text{where } 0 \leqslant \widehat{b_i} \leqslant 1, \widehat{b_0} + \widehat{b_1} + \widehat{b_2} = 1, \text{ and } \widehat{\boldsymbol{b}} \text{ minimizes}\\ &M = \mathbf{b'}\widehat{\mathbf{V}}\mathbf{b} + (\mathbf{b'}\widehat{\mathbf{B}})^2. \end{aligned} \tag{8}$$

The minimization in (8) is a regular quadratic programming problem, which has many fast numerical solvers, eg, the *constrOptim()* function and the *quadprog* package in R.[29]

To perform inference on $\widehat{\theta}_s$, there are 2 challenges to overcome: accounting for the uncertainty in estimation of the weights and accounting for the fact that the estimator is generally biased. To handle the first issue, we will perform a bootstrap of the entire synthetic estimation procedure, which itself contains a nonparametric bootstrap, leading to 2 layers of bootstrapping, ie, a double bootstrap. Double bootstrapping procedures have been used before[30,31] in various contexts such as bias correction and interval correction for confidence intervals. Here, we use it to account for all of the variation in our estimation procedure for the synthetic estimator, which includes a bootstrap. This algorithm can be described in the following way:

1. Bootstrap N_1 new datasets from the original data by sampling with replacement. Call the new data sets $D^{(1)}, \ldots, D^{(N_1)}$
2. For each new data set $D^{(i)}$, estimate the synthetic estimator $\widehat{\theta}_s^{(i)}$ using the following steps:
   (a) Bootstrap N_2 new data sets by sampling with replacement from $D^{(i)}$
   (b) Use the N_2 data sets to estimate $\widehat{\boldsymbol{V}}^{(i)}$, the covariance matrix for the i^{th} bootstrapped dataset
   (c) Use the difference between the IV estimate and the PP and AT estimates from $D^{(i)}$ as an estimate of the bias, $\widehat{\boldsymbol{B}}^{(i)}$
   (d) Estimate the weights $\widehat{\mathbf{b}}^{(i)}$ to be the vector that minimizes the MSE in Eq. 8 using $\widehat{\boldsymbol{V}}^{(i)}$ and $\widehat{\boldsymbol{B}}^{(i)}$
3. Estimate $\text{var}(\widehat{\theta}_s)$ as the variance of the N_1 bootstrapped synthetic estimates.

This bootstrap variance can be in conjunction with a normal approximation to perform inference. Alternatively, we can use the quantiles of the bootstrap distribution to create a confidence interval. While both of these approaches should account for the variation in estimating the weights, neither account for the fact that our estimator is biased. We expect the bias to be small; therefore, confidence intervals based solely on the variance should perform reasonably well. If this bias is not small, then the synthetic estimator may have poor interval coverage when based only on $var(\widehat{\theta}_s)$. One way to widen the confidence intervals to maintain nominal coverage is to use the MSE instead of the variance. One can estimate the MSE of $\widehat{\theta}_s$ by $(\widehat{\mathbf{b}}'\widehat{\mathbf{B}})^2 + var(\widehat{\theta}_s)$. R code to implement the synthetic estimator can be found in the Supporting Information.

3.3 | A normal-mixture example

Under parametric assumptions, the finite-sample moments in Section 3.2 can be expressed as functions of parameters. However, even with closed form expressions for these moments, the optimal weights of (8) may still fail to have simple, closed forms due to the constrained optimization. Therefore, we consider a simplified setting where (1) we focus our attention on $\widehat{\theta}_{iv}$ and $\widehat{\theta}_{pp}$; (2) we assume π_c, π_n, and π_a are known and that $\pi_a = 0$, ie, no always takers or equivalently the observed compliance rate in the control arm is always 100%; (3) we do not require that the weights be between 0 and 1, ie, an affine combination; and (4) the outcome measurement is assumed to follow a simple normal mixture distribution with an unknown but equal variance. Nevertheless, even under this simplified parametric setting, the closed form of the synthetic estimator is still fairly complex.

Using results of Longford,[14] it can be shown that the optimal weight in this simplified scenario is

$$b_1 = \frac{1 + \Delta_0^2 \pi_c (1 - \pi_c)\alpha}{(1 - \alpha\pi_c) + \Delta_0^2 \pi_c (1 - \pi_c)\alpha \left[1 - \pi_c + m\pi_c(1 - \alpha)\right]},$$

where $\Delta_0 = \frac{\mu_{0c} - \mu_{0n}}{\sigma}$, m is the overall sample size and α is the known proportion of subjects assigned treatment by design. Details of the derivation of the above result can be found in the Supporting Information. This weight provides some useful insight on our proposed synthetic estimator. First, if NCEM holds (ie, $\Delta_0 = 0$), then the optimal weight is $b_1 = \frac{1}{1-\alpha\pi_c}$, which shows that $\widehat{\theta}_{pp}$ receives a large positive weight regardless of sample size. This is expected as $\widehat{\theta}_{pp}$ is known to be more efficient than $\widehat{\theta}_{iv}$, and both estimators are unbiased when both ERM and NCEM hold. The reason that the weight is not equal to 1 in this setting is because of the covariance between the IV and PP estimators and because we did not constrain the weights to be between 0 and 1. Second, if NCEM does not hold, then $b_1 \to 0$ as the sample size increases, ie, the IV estimator receives all of the weight. This occurs because $\widehat{\theta}_{pp}$ is biased, and the MSE is dominated by bias for large samples. Third, as compliance goes to 0, the PP estimator receives all the weight (ie, $b_1 = 1$). This is also expected, as the variance of the IV estimator is known to increase as compliance decreases.

Under alternative specifications of the problem, such as allowing the outcome to have different variances in each of the principal strata or considering all 3 estimators simultaneously, the closed form expression for the optimal weight can be derived. However, we chose to bypass these derivations for the nonparametric bootstrap-based approach proposed in Section 3.2 because of its flexibility. The closed form optimal weights rely on either parametric assumptions or asymptotics, neither of which are appealing, while the nonparametric bootstrap-based approach is applicable to a wide range of outcomes without modification.

3.4 | Synthetic estimator with covariates

As described in Little et al,[9] all of the candidate estimators can be improved by including covariates. Adjusting for covariates may help improve the precision of the IV estimator and the biases of the AT and PP estimators. The covariate free AT and PP estimators are equivalent to a regression of Y on $T(R)$ with all cases and a regression of Y on $T(R)$ with all cases who complied, respectively. These regressions can readily include additional covariates to reduce selection bias. This is important for the proposed synthetic estimator, because in scenarios with a large amount of selection bias, the synthetic estimator assigns most of the weight to the IV estimator leading to a less efficient estimator. If the bias of the AT and PP estimators can be reduced, then the synthetic estimator would assign more weight to these estimators and would be more efficient. Similarly, covariates can be included to improve the efficiency of the IV estimator by stratifying on covariates that are predictive of the outcome of interest or by including them in the two-stage least squares approach to IV regression, which would further improve the synthetic estimator.

If the number of covariates is large and the sample size is also relatively large, then ideas similar to the propensity score[32] can be adopted to avoid model misspecification bias in the regression models. Compliance propensities, $p_n(X) = p(C = n|C = n \text{ or } c, X)$ and $p_a(X) = p(C = a|C = a \text{ or } c, X)$ can be estimated from the data to address selection bias in the AT and PP estimators. As described in Little et al,[9] if the NCEM assumption holds conditional on covariates, then the NCEM condition will hold conditional on the compliance propensities. First, one can estimate both $\widehat{P_n}^*(x) = P(C = n|X)$ from a logistic regression among those with $R = 1$, and $\widehat{p_a}^*(x) = P(C = a|X)$ from a logistic regression among those with $R = 0$. Then the compliance propensity estimates are $\widehat{p_a}(x) = \widehat{p_a}^*(x)/(1 - \widehat{p_n}^*(x))$ and $\widehat{p_n}(x) = \widehat{p_n}^*(x)/(1 - \widehat{p_a}^*(x))$. These estimates can be stratified on or included in a regression model to reduce bias in the AT and PP estimators. Similar to the

use of covariates above, this reduction in bias in the AT and PP estimators can improve both the bias and the efficiency of the synthetic estimator. Importantly, any of the estimators mentioned in this section can be included as a candidate estimator in the synthetic estimator without any additional work.

4 | SIMULATION STUDY

4.1 | Simulation setup

In the simulation study, we varied 4 key components to assess the synthetic estimator's ability to estimate the CACE: (1) the outcome distribution (Normal, T, Poisson, and Bernoulli); (2) the sample size within both the treated and controls ($m_{0+} = m_{1+}, m_{1+} \in \{100, 400\}$); (3) the compliance rate from 50% to 90%; and (4) the amount of selection bias in the sample ($\Delta_0 = \frac{\mu_{0c} - \mu_{0n}}{\sigma} \in \{0, 0.5, 0.8\}$ where σ is the standard deviation of the data). For brevity, we only present results for the Poisson distribution with a sample size of 100; however, the results for all distributions and sample sizes were qualitatively similar and the additional results can be found in the Supporting Information. To examine the performance of the synthetic estimator, we consider 2 distinct sets of operating characteristics. Section 4.2 highlights the ability of the proposed estimator to estimate the CACE by examining the bias and MSE of the synthetic estimator. Section 4.3 examines the ability to perform inference using the estimated standard errors of the synthetic estimator, measured in terms of 95% confidence interval coverages. All simulations use 5000 replications.

4.2 | Bias and MSE

The top 2 panels of Figure 1 show the MSE and bias of synthetic estimator compared to the 3 candidate estimators for the range of compliance rates and selection bias values that were considered. When $\Delta_0 = 0$ (Figure 1; first column), all of the estimators are unbiased, and the AT, PP, and synthetic estimators have smaller MSE than the IV estimator (0.28, 0.32, 0.56, and 0.94 for $\pi_c = 0.5$, respectively). This illustrates a situation where the synthetic estimator outperforms the IV estimator.

When $\Delta_0 \neq 0$, only the IV estimator is unbiased (Figure 1; second row). However, the synthetic estimator is substantially less biased than the AT or PP estimators. The bias in the synthetic estimator is due to the fact that it is a convex combination of all 3 candidate estimators, and the AT and PP estimators are biased. The synthetic estimator seeks to optimize the tradeoff between this bias and the reduction in variance. This is observed in Figure 1 when $\Delta_0 \neq 0$, as the synthetic estimator is able to obtain an MSE that is generally as good as or better than any of the candidate estimators across the compliance rates. Specifically, when $\Delta_0 = 0.8$, the synthetic estimator outperforms the AT and PP estimators.

Summarizing these results, when $\Delta_0 \neq 0$, the synthetic estimator places most weight on the IV estimator leading to improved performance over the AT and PP estimators, while if $\Delta_0 = 0$, the synthetic estimator assigns more weight to the AT and PP estimators leading to more efficient estimates.

4.3 | Inference

The bottom panel of Figure 1 shows the 95% confidence interval coverage for the 3 candidate estimators and the synthetic estimator. For the simulation, we used bootstrapped confidence intervals for all of the estimators. When there is no selection bias ($\Delta_0 = 0$), all of these estimators achieve the nominal coverage because they are all unbiased and the bootstrap is accurately estimating the standard errors. As the amount of selection bias increases, the bias of the PP and AT estimators grows substantially and the 95% confidence intervals do not obtain coverages near the desired level. Because of its unbiasedness, the IV estimator maintains nominal 95% coverage under any scenario. Importantly, the proposed synthetic estimator achieves coverages at or near 95% in all scenarios illustrating the ability of the bootstrap to estimate the standard errors of our proposed procedure. Note that although the synthetic estimator is generally biased, it still achieves reasonable coverages because it places most of the weight on the IV estimator in scenarios with large bias in the AT and PP estimators. Even in the scenario with the largest amount of bias in the AT and PP estimators ($\Delta_0 = 0.8$), the synthetic estimator obtains coverages between 92% and 94% depending on the amount of noncompliance. It is also important to note that, while not shown, the widths of the confidence interval for the synthetic estimator are always smaller than the IV estimate. When $\Delta_0 = 0.8$, most of the weight is assigned to the IV estimator and the interval widths of the synthetic estimator are similar to those of the IV estimator. However, when Δ_0 is small, much more weight is assigned to the PP and AT estimators, leading to much smaller confidence intervals for the synthetic estimator than the IV estimator.

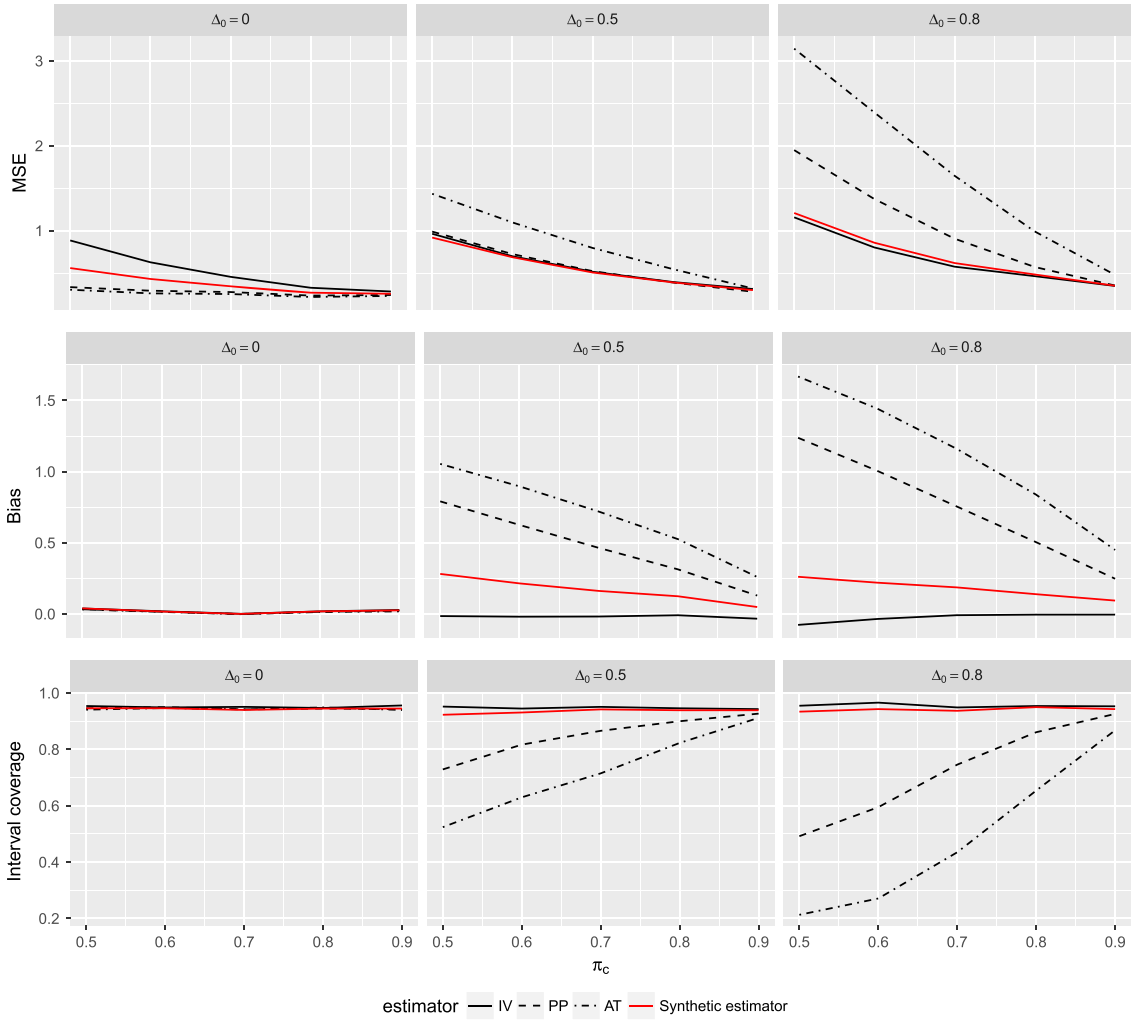

FIGURE 1 Mean squared error, bias, and interval coverage of the four estimators for a range of compliance rates (π_c) and values of Δ_0. Data were generated from a Poisson distribution with a sample size of 100 in both the treatment and control groups. Results are calculated across 5000 simulations [Colour figure can be viewed at wileyonlinelibrary.com]

5 | APPLICATION TO THE THE VIOLENCE AND STRESS ASSESSMENT (VISTA) STUDY

The ViStA study was a care management intervention for PTSD in six federally qualified health centers primarily serving low income, minority, and uninsured or underinsured patients.[2,15] The designed treatment was to provide an enhanced PTSD care management in these clinics with limited integrated care capacities before the study. The treatment was to be compared with a standard care arm, in which the PTSD patients received the minimum level of integrated care that the clinics originally had. In each clinic, enrolled patients were randomized to either the treatment or the control condition according to a predesigned randomization procedure.[15] In this paper, we examined a numeric outcome (PTSD severity score) and a binary outcome (any utilization of mental health care). The PTSD severity score was calculated using the Clinician-Administered PTSD Scale, or CAPS[33] and was the primary outcome, where higher values in severity scores corresponded to more severe conditions. Mental health care utilization was a process outcome, which the treatment was expected to directly influence. Both outcomes were measured between 6 and 9 months after patients enrolled in the study. In the treatment arm, 73% of patients received the minimum level of treatment exposure. In the control arm, due to the design of the intervention (each treated patient must have a dedicated care manager) no patients had opportunities to receive the treatment; therefore, there are no always-takers ($\pi_a = 0$).

Table 2 reports the descriptive statistics for the 2 outcomes. Note that the never-takers among those randomized to receive treatment had a lower mean severity score and a lower probability of using mental health care compared to the

TABLE 2 Descriptive statistics for the outcomes

	Compliance		
Randomization *R*	**N**	**C**	**N+C**
	Severity score		
0	—	—	50.3 (24.8)
1	38.5 (22.9)	50.1 (23.3)	46.9 (23.7)
	Utilization		
0	—	—	62.0
1	48.0	79.9	71.2

Note. The sample sizes are 171 for the control arm ($R = 0$), 50 non-takers in the treatment arm ($T(1) = 0$), and 134 compliers in the treatment arm ($T(1) = 1$).The upper table is the mean (SD) of the severity score, and the lower table gives the percentage of any utilization.

other 2 principal strata. This difference may suggest that NCEM and ECM do not hold together. However, the difference was not significant at the regular 5% significance level (p-values are 0.086 for severity scores and 0.109 for utilization), and, therefore, Little et al[9] suggest that the AT estimator may still be considered.

We conducted 2 sets of analyses. The first set did not use any baseline covariates, and the candidate estimators were in the exact form as in Section 2. The second set of analyses adjusted for a set of covariates in regression models as discussed in Section 3.4, where the AT and PP were fitted by regular ordinary least squares linear regressions and the IV estimator was fitted through the two-stage least squares regression as implemeted in the *AER* package in R.[29] The covariates included demographics (age, gender, race/ethnicity, whether English is the primary language, education level, marital status, and whether living with children), insurance status (Medicaid vs not), baseline mental health measurements (eg, total number of lifetime traumas, depression disorder diagnosis, anxiety severity, alcohol and other substance abuse diagnosis), total number of medical conditions, mental health utilization history (eg, prescription, counseling, community services, inpatient care, and emergency care), and interview mode (in person/over a phone, with/without assistance). The complete list of covariates as well as the substantive discussion of their roles were reported in Meredith et al.[2] It is worth noting that the baseline measurements of both outcomes were included as covariates, because they fell in the category of baseline mental health measurement and mental health utilization history, respectively. Using baseline values of the outcome as covariates may also improve the statistical efficiency in testing the main treatment effect[34] and may help reduce the bias in the PP and AT estimator.

Table 3 presents the 3 candidate estimates and the synthetic estimate, along with the ITT estimate for effectiveness. The variances of all estimators were based on the bootstrap method. The differences among the 3 candidate estimates were remarkable. Without adjusting for covariates, the PP and AT estimates suggested a significant treatment effect on utilization of mental health care. Although no estimates were significant for PTSD severity score, estimates by PP and AT gave less favorable results for the treatment because the severity was estimated as qualitatively unchanged (decreasing severity score by 0.2 for PP) or even increased (increasing severity score by 2.4 for AT). The IV estimate, although still insignificant, suggested that treatment decreased the severity score by −4.6. In contrast, the ITT estimate for effectiveness was not significant for either outcome. After adjusting for covariates, differences between the PP, AT, and IV estimates were smaller, perhaps due to reduced biases in the PP and AT estimates. The ITT estimate did not have a notable change because of randomization. Standard errors of all methods were smaller because many covariates were related to the outcome (eg, baseline measurements of the outcomes). However, the overall pattern among all estimates looked still very similar to the results without adjusting for covariates. In general, the synthetic estimate may have improved statistical efficiency, whose results suggested a significant treatment efficacy on utilization of mental health care (increasing utilization by roughly 15% according to the unadjusted analysis), but the efficacy was not significant for PTSD severity (decreasing 3.9 points in severity scores according to the unadjusted analysis).

As shown in Table 3, without adjusting for covariates, the synthetic estimate was essentially a weighted average between the IV and PP estimates, where more weights were given to the IV estimate. This assignment of the weights was not surprising. Given the notable difference between non-takers and others, we cannot comfortably accept both the NCEM and ECM assumptions. Consequently, it is reasonable to believe in the AT estimator the least. The synthesis was therefore between the IV and PP and the weights were estimated by the trade-off of bias and variance. After adjusting for covariates, possibly due to reduced biases in the AT and PP estimates, more weights were given to the AT and PP estimates. In particular, the synthetic estimate for the utilization of mental health care is almost entirely based on the AT estimate

TABLE 3 Estimates and standard errors for efficacy and effectiveness for the ViStA data

	Unadjusted		Adjusted	
Method	**Estimate (SE)**	**b**	**Estimate (SE)**	**b**
		Severity Score		
ITT	−3.4 (2.5)	–	−2.3 (2.2)	–
IV	−4.6 (3.4)	0.84	−3.1 (3.0)	0.76
PP	−0.2 (2.7)	0.16	−0.7 (2.5)	0.09
AT	2.4 (2.6)	0.00	0.7 (2.4)	0.15
Synthetic	−3.9 (3.5)	–	−2.3 (3.0)	–
		Utilization		
ITT	9.2 (4.9)	–	8.0 (4.3)	–
IV	12.6 (6.6)	0.61	10.8 (5.8)	0.05
PP	**17.9 (4.9)**	0.39	**13.0 (4.3)**	0.00
AT	**21.0 (9.2)**	0.00	**12.3 (4.1)**	0.95
Synthetic	**14.7 (6.7)**	–	**12.2 (5.0)**	–

Note: The ITT estimator is estimating effectiveness while the remaining estimators estimate the efficacy. Unadjusted estimates are the same as described in Section 2, while adjusted estimates further adjust for covariates in the ViStA data. Statistical significance using the bootstrapped standard errors and regular Wald's z-tests at the .05 level are in bold. The "b" column is the estimated weights for the synthetic estimator.

(weight = 0.95). This table exemplified the fact that the weights of the synthetic estimator is fully data driven and can vary drastically depending on the estimated bias and variance trade-off in analyzing an outcome.

6 | DISCUSSION

In this paper, we have shown that synthetic estimation can have a positive impact for estimating treatment efficacy when there is noncompliance in clinical trials. We have shown via simulation that the synthetic estimator protects against having an MSE as bad as the worst candidate estimator and has the potential to perform even better than the best candidate estimator. These results hold under a wide variety of scenarios and levels of selection biases. The synthetic estimator is nonparametric, which allows it to be applied to any data type without additional restrictive assumptions. Using weights that are derived from bootstrapping procedures also has advantages over theoretically driven weights, which rely on additional assumptions or asymptotics. Synthetic estimation can be applied to any group of candidate estimators, which allows for the incorporation of additional information through covariates or compliance propensities that improve efficiency without any additional work.

If the covariance and biases of the candidate estimators were known, then the synthetic estimator would be guaranteed to have MSE lower than or equal to all of the candidate estimators.[14] This holds because each candidate estimator is a special case of the synthetic estimator where all of the weight is placed on that estimator, and knowledge of the covariances and bias allows us to identify the optimal choice among all potential weights. In practice, these quantities are not known; in many cases, they can be estimated with enough precision for the synthetic estimator to perform nearly as well as the best candidate estimator without knowing a priori which estimator is the best. In particular, we used the observed difference between the IV estimator and the other candidate estimators as a surrogate for the unknown bias. This approach to estimating bias seems crude but yields very good performance in simulations. It is worth noting that, under the normal mixture model, the bias of the AT and PP estimators can be estimated. However, the performance of the synthetic estimator was not further improved using model-based bias estimates compared to the raw differences when the data were generated from such a normal mixture in our simulation studies.

The convex combination constraint is not necessary for forming the synthetic estimator, although it is useful in 2 ways. It gives an easy interpretation as a weighted average, which has been well exemplified in Table 3 of our case study. It can also be shown that this assumption guarantees the MSE of the synthetic estimator with any weights is no worse than the largest MSE among the candidate estimators. Presumably, if we relax the convexity assumption, then the MSE of the synthetic estimator would be reduced, but it may be difficult to interpret how the synthetic estimator combines the candidate estimators, and it is no longer guaranteed to outperform the worst candidate estimator.

Another drawback of the proposed approach is that the synthetic estimator will necessarily be biased except in the scenario when both ERM and NCEM hold. We do not find this to be a major drawback since the bias is generally small and interval coverage is still close to the desired level. If, however, the true CACE is very small, this small bias could still impact operating characteristics such as power, which are important in clinical research. Interestingly, the synthetic estimator is consistent as long as the assumptions for the unbiased estimator are satisfied, because all of the weight will be assigned to the unbiased estimator asymptotically. While asymptotics is not of great concern in clinical trials due to small sample sizes, this further justifies the use of the synthetic estimator as it will tend towards unbiasedness. If an unbiased estimator is desired, then one must choose the unbiased estimator a priori (eg, the IV estimator under ERM).

Our approach can be readily generalized to more than 2 discrete levels of compliance. In many trials, the all-or-nothing compliance is actually a dichotomization to the actual compliance level. There is usually either a clinical threshold or an important milestone in the treatment plan, both of which can define a binary clinical compliance. When more than one such threshold is available, one can discretize the compliance level to more than 2 levels. For example, when there is no always taker, the treatment group may consist of 3 compliance levels: never-taker, partial compliance, and full compliance. We can conduct 2 separate CACE analyses: First, eliminate partial compliance cases from the study sample and estimate the CACE of full compliance, and then eliminate the full compliance cases and study the CACE of partial compliance. The causal assumptions (ECM or NCEM) have to be assumed for each subset analysis separately. The synthetic estimator is still applicable.

As mentioned in Section 2, all 3 candidate estimators are interpreted as estimating the CACE; therefore, the synthetic estimator is also interpreted as estimating the CACE. The fact that the interpretation is shared among all estimators facilitates the use of synthetic estimation in randomized trials with dichotomous noncompliance. In other contexts, the candidate estimators may be estimating different quantities, and analysts must be explicit about the exact target of inference. Care should be taken to ensure interpretability of the results before applying synthetic estimation in contexts outside the scope of this paper.

Overall, we have shown that synthetic estimation is a promising approach when noncompliance is affecting a clinical trial. Synthetic estimation has many advantages in its ease of application, extension to a wide range of estimators and distribution types, and ability to produce estimators with substantially lower MSE than many of the candidate estimators.

ACKNOWLEDGEMENTS

Joseph Antonelli was supported by the National Institutes of Health (ES000002, ES024332, ES007142, ES026217 P01CA134294, R01GM111339, R35CA197449, and P50MD010428). Matthew Cefalu and Bing Han were supported by grant R01DA040721 from the National Institute of Drug Abuse.

ORCID

Joseph Antonelli http://orcid.org/0000-0001-7464-5766

REFERENCES

1. Besch CL. Compliance in clinical trials. *Aids*. 1995;9(1):1-10.
2. Meredith LS, Eisenman DP, Han B, et al. Impact of collaborative care for underserved patients with PTSD in primary care: a randomized controlled trial. *J Gen Intern Med*. 2016;31(5):509-517.
3. Schnurr PP, Friedman MJ, Oxman TE, et al. Respect-PTSD: re-engineering systems for the primary care treatment of PTSD, a randomized controlled trial. *J Gen Intern Med*. 2013;28(1):32-40.
4. Roy-Byrne P, Craske MG, Sullivan G, et al. Delivery of evidence-based treatment for multiple anxiety disorders in primary care: a randomized controlled trial. *JAMA*. 2010;303(19):1921-1928.
5. Marks I, Lovell K, Noshirvani H, Livanou M, Thrasher S. Treatment of posttraumatic stress disorder by exposure and/or cognitive restructuring: a controlled study. *Arch Gen Psychiatry*. 1998;55(4):317-325.
6. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ*. 1999;319(7211):670-674.
7. Higgins JPT, Green S. *Cochrane Handbook for Systematic Reviews of Interventions*, Vol. 5. Hoboken, NJ: Wiley Online Library; 2008.
8. Greene WH. *Econometric Analysis*. Delhi, Chennai, India: Pearson Education India; 2003.
9. Little RJ, Long Q, Lin X. A comparison of methods for estimating the causal effect of a treatment in randomized clinical trials subject to noncompliance. *Biometrics*. 2009;65(2):640-649.
10. Buckland ST, Burnham KP, Augustin NH. Model selection: an integral part of inference. *Biometrics*. 1997;53(2):603-618.

11. Hjort NL, Claeskens G. Frequentist model average estimators. *J Am Stat Assoc*. 2003;98(464):879-899.
12. Claeskens G, Croux C, Van Kerckhoven J. Variable selection for logistic regression using a prediction-focused information criterion. *Biometrics*. 2006;62(4):972-979.
13. Hansen BE. Least squares model averaging. *Econometrica*. 2007;75(4):1175-1189.
14. Longford NT. *Missing Data and Small-Area Estimation: Modern Analytical Equipment for the Survey Statistician*. New York, NY: Springer Science & Business Media; 2006.
15. Meredith LS, Eisenman DP, Green BL, et al. Design of the Violence and Stress Assessment (ViStA) study: a randomized controlled trial of care management for PTSD among predominantly latino patients in safety net health centers. *Contemp Clin Trials*. 2014;38(2):163-172.
16. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics*. 2002;58(1):21-29.
17. White IR. Uses and limitations of randomization-based efficacy estimators. *Stat Methods Med Res*. 2005;14(4):327-347.
18. Robinson GK. That blup is a good thing: the estimation of random effects. *Stat Sci*. 1991;6(1):15-32.
19. Searle SR, Casella G, McCulloch CE. *Variance Components*, Vol. 391. Hoboken, NJ: John Wiley & Sons; 2009.
20. Ghosh M, Rao JNK. Small area estimation: an appraisal. *Stat Sci*. 1994;9(1):55-76.
21. Pfeffermann D. Small area estimation-new developments and directions. *Int Stat Rev*. 2002;70(1):125-143.
22. Datta GS, Hall P, Mandal A. Model selection by testing for the presence of small-area effects, and application to area-level data. *J Am Stat Assoc*. 2011;106(493):362-374.
23. Han B. Conditional Akaike information criterion in the Fay–Herriot model. *Stat Methods*. 2013;11:53-67.
24. Judge GG, Mittelhammer RC. A semiparametric basis for combining estimation problems under quadratic loss. *J Am Stat Assoc*. 2004;99(466):479-487.
25. Mittelhammer RC, Judge GG. Combining estimators to improve structural model estimation and inference under quadratic loss. *J Econ*. 2005;128(1):1-29.
26. Judge GG, Mittelhammer RC. A minimum mean squared error semiparametric combining estimator. *Essays in Honor of Jerry Hausman*. Bingley, United Kingdom: Emerald Group Publishing Limited; 2012:55-85.
27. Stein C. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: Proceedings of the third berkeley symposium on mathematical statistics and probability, Vol. 1. Berkeley, CA; 1956:197-206.
28. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. Boca Raton, FL: CRC press; 1994.
29. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2014. http://www.R-project.org/.
30. Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. *J Am Stat Assoc*. 1983;78(382):316-331.
31. Martin MA. On the double bootstrap. *Computing science and statistics*. New York, NY: Springer; 1992:73-78.
32. Rosenbaum PR, Rubin D. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55.
33. Weathers FW, Keane TM, Davidson JRT. Clinician-administered PTSD scale: a review of the first ten years of research. *Depress Anxiety*. 2001;13(3):132-156.
34. Yang L, Tsiatis AA. Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The Am Stat*. 2001;55(4):314-321.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Antonelli J, Han B, Cefalu M. A synthetic estimator for the efficacy of clinical trials with all-or-nothing compliance. *Statistics in Medicine*. 2017;1–12. https://doi.org/10.1002/sim.7447