Received: 16 October 2018 | Accepted: 14 March 2019

DOI: 10.1111/biom.13062

BIOMETRIC PRACTICE

WILEY Biometrics
A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY

Analysis of covariance in randomized trials: More precision and valid confidence intervals, without model assumptions

Bingkai Wang | **Elizabeth L. Ogburn** | **Michael Rosenblum**

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland

Correspondence
Michael Rosenblum, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, MD 21205.
Email: mrosen@jhu.edu

Funding information
NIH, Grant/Award Numbers: R01AG048349, UL1TR001079

Abstract

"Covariate adjustment" in the randomized trial context refers to an estimator of the average treatment effect that adjusts for chance imbalances between study arms in baseline variables (called "covariates"). The baseline variables could include, for example, age, sex, disease severity, and biomarkers. According to two surveys of clinical trial reports, there is confusion about the statistical properties of covariate adjustment. We focus on the analysis of covariance (ANCOVA) estimator, which involves fitting a linear model for the outcome given the treatment arm and baseline variables, and trials that use simple randomization with equal probability of assignment to treatment and control. We prove the following new (to the best of our knowledge) robustness property of ANCOVA to arbitrary model misspecification: Not only is the ANCOVA point estimate consistent (as proved by Yang and Tsiatis, 2001) but so is its standard error. This implies that confidence intervals and hypothesis tests conducted as if the linear model were correct are still asymptotically valid even when the linear model is arbitrarily misspecified, for example, when the baseline variables are nonlinearly related to the outcome or there is treatment effect heterogeneity. We also give a simple, robust formula for the variance reduction (equivalently, sample size reduction) from using ANCOVA. By reanalyzing completed randomized trials for mild cognitive impairment, schizophrenia, and depression, we demonstrate how ANCOVA can achieve variance reductions of 4 to 32%.

KEYWORDS
imbalance, relative efficiency, robustness

1 | INTRODUCTION

Pocock et al. (2002), in a survey of 50 randomized trial reports, found that 36 used covariate adjustment, but only 12 emphasized adjusted over unadjusted estimators. They stated that "the statistical properties of covariate adjustment are quite complex and often poorly understood, and there remains confusion as to what is an appropriate statistical strategy." Austin et al. (2010), in a paper titled "A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals", surveyed 114 randomized trial articles and found that only 39 of them presented an adjusted analysis.

We focus on the analysis of covariance (ANCOVA) estimator, referred to as "ANCOVA I" by Yang and Tsiatis (2001). It involves fitting a linear regression model for the primary outcome with intercept and main terms for the treatment assignment and baseline variables. For

trials with continuous-valued or change score outcomes, covariate adjustment often involves the ANCOVA estimator. We use the term "covariate adjustment" to refer to the ANCOVA estimator.

Concerns have been raised about the validity of ANCOVA for analyzing randomized trial data when the linear model is misspecified. For example, Kraemer (2015) states "The linear model used for covariate adjusting (eg, analysis of covariance) assumes ... that there is no interaction between the covariates and the treatment effect." and "Given these risks for bias, ANCOVA should not generally be used for such adjustment." Ludvigsson et al. (2008) and Montalban et al. (2017) both checked whether data are normally distributed before applying ANCOVA and Ludvigsson et al. (2008) state that "ANCOVA involves the assumption of normally distributed response data and homogeneity of variances." These authors are wise to have the general concern about model misspecification, since in many contexts it can lead to biased or uninterpretable analyses. However, when the ANCOVA estimator is used to analyze randomized trial data, it has special robustness properties that obviate the above concerns.

Yang and Tsiatis (2001) proved that the ANCOVA estimator is consistent under arbitrary misspecification of the linear model. We build on this result by proving that the standard error, computed as if the linear model were correct, is also consistent. Therefore, not only estimates but also confidence intervals and hypothesis tests conducted as if the linear model were correct are asymptotically valid even when the linear model is arbitrarily misspecified, for example, when the true relationships between variables are nonlinear, and/or when there is treatment effect heterogeneity. This is important since it is not possible to rule out all types of model misspecification.

We attempt to provide intuition behind the precision gains from covariate adjustment by showing a direct analogy to ordinary least squares linear regression. We prove that the asymptotic variance reduction (i.e., precision gain) due to covariate adjustment equals the fraction of variance in the primary outcome explained by the baseline variables, beyond what is already explained by the main effect of treatment. This holds under arbitrary model misspecification and leads to a simple formula for estimating the variance reduction due to ANCOVA. This variance reduction is important since it equals the reduction in the required sample size to achieve a desired power.

The above results build on key ideas from Tsiatis et al. (2008); Rubin and van der Laan (2008); Moore and van der Laan (2009); Moore et al. (2011); Rubin and van der Laan (2011); Jiang et al. (2018); Tian et al. (2019). As in their work, our results are asymptotic, that is, they hold in the limit as sample size grows to infinity while the set of covariates is fixed. A special case of the results of Bugni et al. (2018) coincides with a special case of our result in Section 4.2 (explained in that section).

We present data analyses based on three completed randomized clinical trials for treatment of mild cognitive impairment (MCI; Petersen et al., 2005) schizophrenia, (Jarskog et al., 2013), and depression (Treatment for Adolescents With Depression Study (TADS) Team, 2004), respectively. By analyzing these datasets, we demonstrate how covariate adjustment can reduce variance, have greater added value in large trials in terms of reducing the required sample size to achieve a desired power, and increase power even when by chance there is perfect or near-perfect balance across arms in the baseline variables (due to the ANCOVA estimator's smaller standard error).

In the next section, we describe the three trials. In Section 3, we define the unadjusted estimator, ANCOVA estimator, and covariate imbalance. In Section 4, we present our main results. Illustrations are provided in Section 5, where trial analyses are presented. Some practical recommendations for applying covariate adjustment are given in Section 6.

2 | THREE COMPLETED RANDOMIZED CLINICAL TRIALS

2.1 | Mild cognitive impairment trial

The "Vitamin E and Donepezil for the Treatment of Mild Cognitive Impairment" phase 3 randomized trial was completed in 2004 (Petersen et al., 2005). The goal was to estimate the effect of a drug treatment on preventing progression from MCI to Alzheimer's disease. Participants were randomized to the following three arms: the drug Donepezil, Vitamin E, and placebo control. We compare the Donepezil arm (253 participants, 33% missing outcomes) to the placebo arm (259 participants, 28% missing outcomes). The primary outcome was time to progression to Alzheimer's disease. In order to apply the ANCOVA estimator, which requires a continuous or change score outcome, we instead use the change in Clinical Dementia Rating-sums of boxes score (CDR-SB) between baseline and 18 months. We use the following baseline variables for adjustment: age, gender, Alzheimer's Disease Assessment Scale (ADAS)-cognitive score, Mini-Mental State Examination (MMSE) score, Activities of Daily Living total score, Global Deterioration scale, and CDR-SB.

2.2 | Metformin for weight loss trial

The "Metformin for weight loss and metabolic control in overweight outpatients with schizophrenia and schizoaffective disorder" trial, referred to as "METS", is a phase 4 randomized trial completed in 2010 (Jarskog et al., 2013). Participants were randomly assigned to the following two arms: Metformin (treatment, 75 participants; 15% missing outcomes) and placebo (control, 71 participants; 14% missing outcomes). The primary outcome was weight loss over 16 weeks. We use this outcome and the following baseline variables: age, gender, Clinical Global Impressions (CGI) severity rating score, tobacco use, illicit drug use, alcohol use, weight, and body mass index (BMI).

2.3 | Treatment for adolescents with depression study

The "Treatment for Adolescents with Depression Study" (TADS) is a phase 3, four-arm, randomized trial completed in 2003 (Treatment for Adolescents With Depression Study (TADS) Team, 2004). The goal was to evaluate cognitive-behavioral therapy (CBT) and Fluoxetine (FLX), each alone and combined (CMB), for treating major depressive disorder in adolescents (age 12–17). Participants were randomized to the following four arms: FLX only (109 participants, 15% missing outcomes), CBT only (111 participants, 29% missing outcomes), combined (CMB, 107 participants, 16% missing outcomes), and placebo (112 participants, 20% missing outcomes). The coprimary outcomes were the change in Children's Depression Rating Scale-Revised (CDRS-R) score and improvement of Clinical Global Impressions (CGI) severity rating score at 12 weeks. We focus on the former outcome and adjust for the following baseline variables: age, gender, CDRS-R score, CGI severity rating score, Children's Global Assessment Scale score (CGAS), Reynolds Adolescent Depression Scale total score (RADS), suicide ideation score, current major depressive episode duration, and comorbidity (indicator of any other psychiatric disorder except dysthymia).

3 | DEFINITIONS

3.1 | Estimators of average treatment effect

We focus on randomized clinical trials where each participant contributes the generic data vector $(\boldsymbol{W}, A, Y)$, where $\boldsymbol{W}$ is a $k \times 1$ column vector of predefined baseline variables, A is the study arm assignment, and Y is the outcome. We assume that Y is continuous or a change score (difference between a score measured at follow-up and baseline). We assume the study arm assignment indicator A is binary ($A = 1$ for treatment and $A = 0$ for control). For trials with more than 1 treatment (eg, TADS), we consider each treatment arm versus control comparison separately. All variables are assumed to be bounded.

The components of the baseline vector $\boldsymbol{W}$ can be continuous, binary, ordinal, and/or categorical. We assume that the components of $(1, \boldsymbol{W}^t)$ are linearly independent, that is, no component is a linear combination of the others.

For each participant $i = 1,...,n$, we observe the data vector $(\boldsymbol{W}_i, A_i, Y_i)$, which we assume to be an independent, identically distributed draw from the unknown, joint distribution on the generic data vector $(\boldsymbol{W}, A, Y)$. The only assumption that we make about the joint distribution is that the study arm A is randomly assigned with equal probability to treatment or control independent of the baseline variables $\boldsymbol{W}$. This holds by design in trials using simple randomization with equal probability of assignment to each study arm, which is the type of trial design that we consider throughout. The assumptions in this paragraph do not hold in trials that use stratified block randomization or covariate-adaptive randomization, which are discussed in Section 6.

The goal is to estimate the population average treatment effect $\Delta = E[Y|A = 1] - E[Y|A = 0]$, that is, the difference between population means if everyone in the study population had been assigned to treatment versus control. We focus throughout on estimating the average treatment effect Δ, since that is the principal quantity of interest in the primary efficacy analysis of confirmatory randomized trials (Tsiatis et al., 2008). An estimator of the average treatment effect Δ is called robust to arbitrary model misspecification if it is consistent under the aforementioned assumptions. These assumptions do not put any restrictions on the joint distribution of $\boldsymbol{W}, A, Y$ other than A being randomly assigned with equal probability to each arm independent of the baseline variables $\boldsymbol{W}$.

Denote the unadjusted estimator (which ignores baseline variables) of the average treatment effect Δ by. The unadjusted estimator is consistent, that is, converges to Δ as the sample size goes to infinity.

$$\widehat{\Delta}^{\text{unadj}} = \sum_{i=1}^{n} Y_i A_i \Big/ \sum_{i=1}^{n} A_i - \sum_{i=1}^{n} Y_i (1 - A_i) \Big/ \sum_{i=1}^{n} (1 - A_i)$$

The ANCOVA estimator of the average treatment effect Δ adjusts for chance imbalance between study arms in $\boldsymbol{W}$. It is computed by fitting the following linear regression model:

$$E[Y|A, \boldsymbol{W}] = \beta_0 + \beta_A A + \boldsymbol{\beta}_{\boldsymbol{W}}^t \boldsymbol{W}, \quad (1)$$

using ordinary least squares (OLS). Denote the estimated coefficients by $\widehat{\beta}_0, \widehat{\beta}_A, \widehat{\boldsymbol{\beta}}_{\boldsymbol{W}}$. The ANCOVA estimator $\widehat{\Delta}^{\text{ancova}}$ of the average treatment effect Δ is the estimated coefficient $\widehat{\beta}_A$.

According to Huitema (2011), the ANCOVA model assumes: (i) a linear relationship between the outcome and the other variables, that is, $Y = \beta_0 + \beta_A A + \boldsymbol{\beta}_{\boldsymbol{W}}^t \boldsymbol{W} + \varepsilon$, where ε is the error term, and (ii) the distribution of the error ε is normal with mean 0 conditional on A and $\boldsymbol{W}$. These assumptions may fail to hold if there is an interaction between treatment and covariate (Kraemer, 2015), if there are unmeasured prognostic covariates that are correlated with $\boldsymbol{W}$ (Austin et al., 2010), or if the outcome is nonlinearly related to the covariates. Fortuitously, the key statistical properties of ANCOVA (consistency of the point estimate and standard error) hold under any of these types of model misspecification.

Yang and Tsiatis (2001) proved that the ANCOVA estimator is consistent for Δ, that is, $\widehat{\beta}_A$ converges to Δ in probability, even under arbitrary misspecification of the linear model Equation (1). Furthermore, the ANCOVA estimator is asymptotically normal and we denote its asymptotic variance as $Var^*(\widehat{\Delta}^{\text{ancova}})$, that is, $n^{1/2}(\widehat{\Delta}^{\text{ancova}} - \Delta)$ converges to a normal distribution with mean 0 and variance $Var^*(\widehat{\Delta}^{\text{ancova}})$. Yang and Tsiatis (2001) also proved that when the probability of being randomized to each study arm is equal (as assumed here), the ANCOVA estimator has asymptotic variance at most that of the unadjusted estimator; if any baseline variable is correlated with the outcome, then ANCOVA is strictly more precise.

We use the ANCOVA estimator $\widehat{\Delta}^{\text{ancova}} = \widehat{\beta}_A$ to estimate the average (also called marginal) treatment effect $\Delta = E[Y|A = 1] - E[Y|A = 0]$. The treatment effect is not assumed to be constant across strata of $\boldsymbol{W}$. We emphasize this to avoid confusion, since the conventional interpretation of the estimated coefficient $\widehat{\beta}_A$ is the conditional treatment effect. That interpretation does not apply when the model is misspecified. For example, when the treatment effect differs within strata of $\boldsymbol{W}$, then the conditional treatment effect is not a single number but instead is a function mapping each stratum of $\boldsymbol{W}$ to the corresponding effect. Though it is of independent interest to estimate the conditional treatment effect, this is often much more challenging and requires more assumptions than estimating the marginal treatment effect (since it involves estimating a function rather than a single number). The reason for considering baseline variables at all when estimating the marginal treatment effect Δ is that this can improve precision and power by accounting for chance imbalances across study arms (Yang & Tsiatis, 2001).

The imbalance $\boldsymbol{I}$ between study arms in the baseline variables $\boldsymbol{W}$, called chance imbalance or covariate imbalance, is the difference between sample means of $\boldsymbol{W}$ comparing treatment versus control arms:

$$\boldsymbol{I} = \sum_{i=1}^n A_i \boldsymbol{W}_i / \sum_{i=1}^n A_i - \sum_{i=1}^n (1 - A_i) \boldsymbol{W}_i / \sum_{i=1}^n (1 - A_i)$$

Although A is independent of $\boldsymbol{W}$ by design, in any realization the baseline variables can be imbalanced.

3.2 | Linear regression variance decomposition and definition of $R^2_{Y-\Delta A\sim \boldsymbol{W}}$

We review properties of OLS regression and define a quantity ($R^2_{Y-\Delta A\sim \boldsymbol{W}}$) that plays a key role in our main results in Section 4. All results below hold under arbitrary model misspecification.

Consider regressing a generic response variable Z on a covariate vector $\boldsymbol{X}$ using the linear model $E[Z|\boldsymbol{X}] = \beta_0 + \boldsymbol{\beta}_{\boldsymbol{X}}^t \boldsymbol{X}$. We assume that all variables are bounded and the components of $(1, \boldsymbol{X}^t)$ are linearly independent. If the model is misspecified, i.e., if for every possible $\beta_0, \boldsymbol{\beta}_{\boldsymbol{X}}$ we have $E[Z|\boldsymbol{X}] \neq \beta_0 + \boldsymbol{\beta}_{\boldsymbol{X}}^t \boldsymbol{X}$, then the OLS estimator $\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}_{\boldsymbol{X}}$ (based on independent, identically distributed vectors $(\boldsymbol{X}_i, Z_i)$: $i = 1, ..., n$) still converges to a limit, denoted $\underline{\beta}_0, \underline{\boldsymbol{\beta}}_{\boldsymbol{X}}$. The variance of Z decomposes as $Var(Z) = Var(\underline{\beta}_0 + \underline{\boldsymbol{\beta}}_{\boldsymbol{X}}^t \boldsymbol{X}) + Var(Z - \underline{\beta}_0 - \underline{\boldsymbol{\beta}}_{\boldsymbol{X}}^t \boldsymbol{X})$, where $\underline{\beta}_0 + \underline{\boldsymbol{\beta}}_{\boldsymbol{X}}^t \boldsymbol{X}$ is the predicted response and $Z - \underline{\beta}_0 - \underline{\boldsymbol{\beta}}_{\boldsymbol{X}}^t \boldsymbol{X}$ is the residual. In other words, the response variance is the sum of the prediction variance and residual variance. The fraction of the variance of Z explained by covariates $\boldsymbol{X}$, denoted $R^2_{Z\sim \boldsymbol{X}}$, is defined as $1 - Var(Z - \underline{\boldsymbol{\beta}}_{\boldsymbol{X}}^t \boldsymbol{X})/Var(Z)$ (where we omit the intercept $\underline{\beta}_0$ here and below since it does not impact the variance).

We apply the above variance decomposition to the linear regression model Equation (1) that is used in computing the ANCOVA estimator. Let $(\underline{\beta}_A, \underline{\boldsymbol{\beta}}_{\boldsymbol{W}})$ denote the limit in probability of the OLS estimator $(\widehat{\beta}_A, \widehat{\boldsymbol{\beta}}_{\boldsymbol{W}})$ for the linear model Equation (1) as sample size n goes to infinity. Our interest is in the variance in the outcome Y explained by baseline variables $\boldsymbol{W}$, beyond what is already explained by treatment A. Therefore, we set the response to be $Z = Y - \underline{\beta}_A A$ and regressor to be $\boldsymbol{X} = \boldsymbol{W}$. The following variance decomposition, analogous to the decomposition of $Var(Z)$ above, is proved in the Supporting Information:

$$Var(Y - \underline{\beta}_A A) = Var(\underline{\boldsymbol{\beta}}_{\boldsymbol{W}}^t \boldsymbol{W}) + Var(Y - \underline{\beta}_A A - \underline{\boldsymbol{\beta}}_{\boldsymbol{W}}^t \boldsymbol{W}). \quad (2)$$

The corresponding fraction of the variance in the outcome Y explained by the baseline variables $\boldsymbol{W}$, beyond what is already explained by (the main effect of) treatment A, is denoted by

$$R^2_{Y-\Delta A\sim \boldsymbol{W}} = 1 - Var(Y - \underline{\beta}_A A - \underline{\boldsymbol{\beta}}^t_{\boldsymbol{W}} \boldsymbol{W})/Var(Y - \underline{\beta}_A A). \quad (3)$$

The subscript in $R^2_{Y-\Delta A\sim \boldsymbol{W}}$ is to indicate that this R^2 represents the fraction of variance of $Y - \Delta A$ explained by $\boldsymbol{W}$, where we made the substitution $\underline{\beta}_A = \Delta$ which holds by the consistency result of Yang and Tsiatis (2001).

The importance of $R^2_{Y-\Delta A\sim \boldsymbol{W}}$ is that, as we show below, it is identical to the asymptotic variance reduction (equivalently, the sample size reduction) comparing the ANCOVA estimator to the unadjusted estimator, and that this holds under arbitrary misspecification of Equation (1). This result builds on fundamental ideas from Rubin and van der Laan (2008); Moore and van der Laan (2009) as described below.

4 | $R^2_{Y-\Delta A\sim \boldsymbol{W}}$ AND THE RELATIONSHIP AMONG UNADJUSTED ESTIMATOR, ANCOVA ESTIMATOR, AND COVARIATE IMBALANCE, UNDER MODEL MISSPECIFICATION

All results below hold under arbitrary model misspecification. Our first result, in Section 4.1, is an equivalence between the ordinary least squares variance decomposition Equation (2) and a variance decomposition relating the unadjusted estimator, ANCOVA estimator, and covariate imbalance. Second, in Section 4.2, we show that the variance estimator for ANCOVA computed by standard statistical software is consistent. Our third result, in Section 4.3, is a simple formula for the variance reduction (equivalently, the sample size reduction) due to covariate adjustment. These results build on ideas from prior work as described below.

4.1 | Connecting ordinary least squares regression to the relationship among unadjusted estimator, ANCOVA estimator, and covariate imbalance

Jiang et al. (2018) proved the following relationship among the unadjusted estimator $\widehat{\Delta}^{\text{unadj}}$, ANCOVA estimator $\widehat{\Delta}^{\text{ancova}}$, and chance imbalance $\boldsymbol{I}$:

$$\widehat{\Delta}^{\text{unadj}} \approx \underline{\boldsymbol{\beta}}^t_{\boldsymbol{W}} \boldsymbol{I} + \widehat{\Delta}^{\text{ancova}}. \quad (4)$$

(Formally, the difference between the left and right sides of the above display, after multiplying by $n^{1/2}$, converges to 0 in probability.) They also showed the following variance decomposition:

$$Var^*(\widehat{\Delta}^{\text{unadj}}) = Var^*(\underline{\boldsymbol{\beta}}^t_{\boldsymbol{W}} \boldsymbol{I}) + Var^*(\widehat{\Delta}^{\text{ancova}}), \quad (5)$$

where Var^* denotes asymptotic (i.e., large sample) variance. We show that the above variance decomposition among the unadjusted estimator, chance imbalance, and ANCOVA estimator is identical to the variance decomposition Equation (2) for OLS, under arbitrary model misspecification. Specifically, we prove in the Supporting Information that each term in Equation (5) equals 4 times the corresponding term in Equation (2), i.e., $Var^*(\widehat{\Delta}^{\text{unadj}}) = 4Var(Y - \underline{\beta}_A A)$, $Var^*(\underline{\boldsymbol{\beta}}^t_{\boldsymbol{W}} \boldsymbol{I}) = 4Var(\underline{\boldsymbol{\beta}}^t_{\boldsymbol{W}} \boldsymbol{W})$, and $Var^*(\widehat{\Delta}^{\text{ancova}}) = 4Var(Y - \underline{\beta}_A A - \underline{\boldsymbol{\beta}}^t_{\boldsymbol{W}} \boldsymbol{W})$. This is summarized in Figure 1, where the first row is the variance decomposition in OLS, the second row is the variance decomposition of $\widehat{\Delta}^{\text{unadj}}$ from Jiang et al. (2018), and our contribution is to connect them by proving equality of quantities in the same column. When model (3.1) is misspecified, all equalities in Figure 1 still hold. These relationships are used to prove the results in Sections 4.2 and 4.3.

4.2 | Robustness of the ANCOVA variance estimator to arbitrary model misspecification

Consider the ANCOVA model-based variance estimator for $\widehat{\Delta}^{\text{ancova}}$ that is output by standard statistical software such as 'summary.lm' in R or 'proc reg' in SAS, which we denote by $\widehat{Var}(\widehat{\Delta}^{\text{ancova}})$. The formula for $\widehat{Var}(\widehat{\Delta}^{\text{ancova}})$ is

$$\widehat{Var}(\widehat{\Delta}^{\text{ancova}}) = \frac{\widehat{Var}(Y - \widehat{\beta}_0 - \widehat{\beta}_A A - \widehat{\boldsymbol{\beta}}^t_{\boldsymbol{W}} \boldsymbol{W})}{(n-1)[\widehat{Var}(A) - \widehat{Cov}(\boldsymbol{W}, A)^t \widehat{Var}(\boldsymbol{W})^{-1} \widehat{Cov}(\boldsymbol{W}, A)]} \quad (6)$$

where on the right side $\widehat{Var}$, $\widehat{Cov}$ are the sample variance and sample covariance, respectively, where degrees of freedom are taken into account. (See the Supporting Information for precise definitions of these.) The following theorem shows that the above variance estimator is robust to arbitrary model misspecification. (See the Supporting Information for proof.)

Theorem 1. *Given the assumptions in Section 3.1, which do not assume that the linear model Equation (1) is*

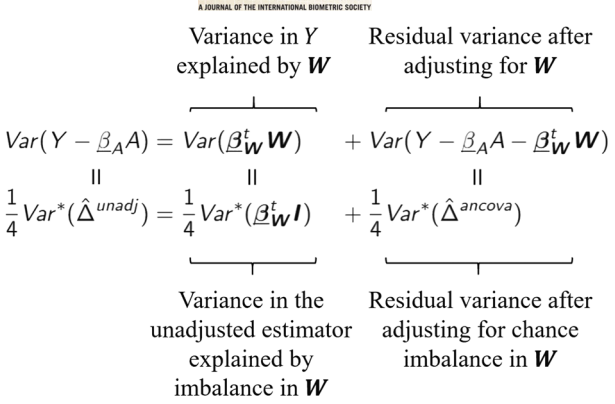

FIGURE 1 Variance decomposition equivalence between linear regression and estimators of average treatment effect. The variance decomposition in the first row is a result of ordinary least squares linear regression. The second row gives the asymptotic variance decomposition of the unadjusted estimator, which is a minor extension of key results from Jiang et al. (2018); Tian et al. (2019). Our contribution is to connect the two variance decompositions by showing their equivalence, that is, quantities in the same column are equal, under arbitrary model misspecification

correctly specified, n times the estimated variance $\widehat{Var}(\hat{\Delta}^{\text{ancova}})$ *converges in probability to the true asymptotic variance* $Var^*(\hat{\Delta}^{\text{ancova}})$ *of the ANCOVA estimator* $\hat{\Delta}^{\text{ancova}}$.

The above theorem implies that confidence intervals and Wald-type hypothesis tests conducted as if the linear model were correct are asymptotically valid even when the linear model is arbitrarily misspecified. The $1-\alpha$ confidence interval for the coefficient on the A term in Equation (1) that is output by the aforementioned, standard linear regression software is

$$(\hat{\Delta}^{\text{ancova}} - t_{n-p,\alpha/2}\sqrt{\widehat{Var}(\hat{\Delta}^{\text{ancova}})},\ \hat{\Delta}^{\text{ancova}} + t_{n-p,\alpha/2}\sqrt{\widehat{Var}(\hat{\Delta}^{\text{ancova}})}), \tag{7}$$

where $t_{n-p,\alpha/2}$ is the $1-\alpha/2$-quantile of the t-distribution with $n-p$ degrees of freedom where p is the number of coefficients in the linear model Equation (1). For large n and fixed p, the quantile $t_{n-p,\alpha/2}$ is approximately the $1-\alpha/2$-quantile of the standard normal distribution. It follows from the above theorem that the above display is an asymptotically valid confidence interval for the average treatment effect Δ, under arbitrary model misspecification.

Bugni et al. (2018) focused on trials using covariate-adaptive randomization, but their results also have implications for simple randomization as considered here. In particular, the special case of the above theorem where $\boldsymbol{W}$ is a single, categorical variable follows from Theorem 4.3 and Remark 4.6 of Bugni et al. (2018).

4.3 | $R^2_{Y-\Delta A\sim \boldsymbol{W}}$ Equals precision gain (and sample size reduction) due to adjustment, even under arbitrary model misspecification

Borm et al. (2007) and Rubin and van der Laan (2008) connect the R^2 from regressing Y on $\boldsymbol{W}$ to the variance reduction due to ANCOVA, while Moore and van der Laan (2009) and Moore et al. (2011) make a similar connection in the context of binary outcomes and estimators based on logistic regression models. Each of the aforementioned approaches requires conditions (such as the linear model being correctly specified or that $\Delta = 0$) or requires additional factors to connect the R^2 to the variance reduction due to covariate adjustment. (See the Supporting Information for more details.) Building on key ideas from their approaches, we prove that the R^2 $R^2_{Y-\Delta A\sim \boldsymbol{W}}$ equals the variance reduction due to ANCOVA without requiring these conditions or extra factors; this R^2 (which differs from the prior work above by incorporating A) is robust to arbitrary model misspecification.

It follows from the relationships in Figure 1 that the fraction $R^2_{Y-\Delta A\sim \boldsymbol{W}}$ of the variance in the outcome Y explained by the baseline variables $\boldsymbol{W}$, beyond what is explained by the treatment A, equals the asymptotic variance reduction due to ANCOVA, i.e.,

$$\begin{aligned} R^2_{Y-\Delta A\sim \boldsymbol{W}} &= 1 - \frac{Var(Y - \underline{\beta}_A A - \underline{\boldsymbol{\beta}}_{\boldsymbol{W}}{}^t \boldsymbol{W})}{Var(Y - \underline{\beta}_A A)} \\ &= 1 - \frac{Var^*(\hat{\Delta}^{\text{ancova}})}{Var^*(\hat{\Delta}^{\text{unadj}})}. \end{aligned} \tag{8}$$

The first equality is the definition of $R^2_{Y-\Delta A\sim \boldsymbol{W}}$, and the second shows that $R^2_{Y-\Delta A\sim \boldsymbol{W}}$ equals the variance reduction due to ANCOVA. The last expression, by definition, equals one minus the asymptotic relative efficiency (also called Pitman efficiency) comparing the unadjusted to the ANCOVA estimator.

In practice, $R^2_{Y-\Delta A\sim \boldsymbol{W}}$ can be estimated by $\hat{R}^2_{Y-\Delta A\sim \boldsymbol{W}} = 1 - \widehat{Var}(\hat{\Delta}^{\text{ancova}})/\widehat{Var}(\hat{\Delta}^{\text{unadj}})$, where $\widehat{Var}(\hat{\Delta}^{\text{ancova}})$ is the variance of the ANCOVA estimator output by standard statistical software as in Equation (6), and $\widehat{Var}(\hat{\Delta}^{\text{unadj}})$ is the variance of the unadjusted estimator estimated analogously (by regressing Y on A and an intercept).

The variance reduction Equation (8) due to ANCOVA is important since it equals the fractional sample size reduction that can be achieved through covariate adjustment when holding the desired power fixed, asymptotically. A variance reduction of $p\%$ means that the sample size required to achieve a desired power is also reduced by $p\%$. Therefore, $\hat{R}^2_{Y-\Delta A\sim \boldsymbol{W}}$ can be used to

estimate the benefits of covariate adjustment in terms of sample size reduction. The $\widehat{R}^2_{Y-\Delta A\sim \boldsymbol{W}}$ values from our datasets range from 4% to 32%, which can be translated into 4% to 32% sample size reductions.

5 | CLINICAL TRIAL APPLICATIONS

Our data analyses for each application (MCI, METS, and TADS) are summarized in Table 1 and described below. All baseline variables were standardized and missing baseline values were imputed by the median for continuous variables and the mode for binary and categorical variables. All participants with missing outcomes were removed from the analysis, for simplicity; in practice, missing outcome data would be handled as described in Section 6. Point estimates and standard errors are rounded to the nearest 0.01. "Confidence Interval" is abbreviated as "CI."

For the MCI trial, the unadjusted treatment effect estimate was $\widehat{\Delta}^{\text{unadj}} = -0.19$ CDR-SB points with standard error 0.15 and 95% CI (−0.49, 0.11), and the ANCOVA estimate was $\widehat{\Delta}^{\text{ancova}} = -0.18$ CDR-SB points with standard error 0.13 and 95% CI (−0.45, 0.08). Compared to the unadjusted estimator, the ANCOVA estimator has a 14% narrower confidence interval and 25% smaller variance, indicating that researchers planning to perform an adjusted analysis could achieve the same precision as the unadjusted analysis with approximately 25% fewer participants.

For the METS trial, the unadjusted treatment effect estimate is $\widehat{\Delta}^{\text{unadj}} = -3.66$ kg of weight change with standard error 1.62 and 95% CI (−6.83, −0.49), and the ANCOVA estimate is $\widehat{\Delta}^{\text{ancova}} = -3.60$ with standard error 1.58 and 95% CI (−6.71, −0.50). Adjustment resulted in an estimated 4% variance reduction.

For the TADS trial, as shown in Table 1, covariate adjustment results in substantial variance reduction for all three treatment arms. This stands out for the Fluoxetine arm, where we estimated that covariate adjustment reduced asymptotic variance by 32%. The ANCOVA estimator, unlike the unadjusted estimator, leads to a statistically significant treatment effect −4.36 CDRS-R points (p-value 0.01); the 95% CI of the ANCOVA estimator (−8.14, −0.58) excludes zero, but that of the unadjusted estimator (−6.02, 3.15) does not.

6 | PRACTICAL RECOMMENDATIONS

Consider the case where the primary outcome Y is a change score (difference between final score and baseline score). In some cases, adjusting for the baseline score alone brings substantial variance reduction. For example, for TADS(FLX) and TADS(CMB), adjusting for only the baseline CDRS-R score gives a similar variance reduction as adjusting for all of the baseline covariates. In other cases, the baseline score can have negligible impact while the other covariates provide substantial variance reduction. For example, in the MCI trial, the baseline score provided approximately 0% variance reduction while the other covariates led to an estimated 25% variance reduction. It is fine to adjust for correlated baseline variables as long as each adds some new prognostic information for the primary outcome.

When the trial has missing outcomes, under the assumption of missing at random (that the outcome distribution is the same for those with missing outcomes as for those with observed outcomes, conditional on treatment assignment and baseline covariates), the unadjusted estimator may no longer be consistent. This can happen if, for example, participants who benefit more from treatment are more likely to drop out than those who benefit less. The ANCOVA estimator remains consistent under missing at random if the ANCOVA model is correctly specified. To add robustness to model

TABLE 1 Summary of clinical trial data analyses: unadjusted estimator for average treatment effect, adjusted estimator (ANCOVA) for average treatment effect, 95% confidence intervals (CI), and estimated variance reduction due to adjustment

Trial Names	Unadjusted Estimator (95% CI)	ANCOVA Estimator (95% CI)	Variance Reduction ($\widehat{R}^2_{Y-\Delta A\sim W}$)
MCI	−0.19 (−0.49, 0.11)	−0.18 (−0.45, 0.08)	25%
METS	−3.66 (−6.83, −0.49)	−3.60 (−6.71, −0.50)	4%
TADS(FLX)	−1.44 (−6.02, 3.15)	−4.36 (−8.14, −0.58)	32%
TADS(CBT)	2.22 (−1.93, 6.38)	0.50 (−3.20, 4.20)	21%
TADS(CMB)	−6.64 (−10.97, −2.32)	−7.65 (−11.28, −4.03)	30%

Negative (positive) estimates are in the direction of clinical benefit (harm). ANCOVA: analysis of covariance; MCI: mild cognitive impairment trial; METS: metformin for weight loss and metabolic control trial; TADS: Treatment for Adolescents with Depression Study

misspecification, one can use a propensity score model for missing outcomes (modeling the probability of missingness given treatment assignment and covariates with, for example, a logistic regression model) as the inverse weight when fitting the ANCOVA model among those with observed outcomes. According to Robins et al. (2007), this estimator due to Marshall Joffe is doubly-robust, that is, consistent as long as one of the two models (propensity score model or ANCOVA model) is correctly specified, under the missing at random assumption. For the three trial examples in this paper, the ANCOVA estimator (which does not incorporate information from participants with missing outcomes) and the aforementioned doubly-robust extension (which incorporates information from all participants) gave similar estimates and confidence intervals. See the Supporting Information for details.

For large trials, for example, with total sample size at least 500, adjusting for prognostic baseline variables (if there are any) is highly recommended since it reduces the required sample size to achieve a desired power (EMA, 2015). This is counter to the (false) intuition that in large trials there is little to gain from covariate adjustment since randomization will likely leave little imbalance to adjust for. Adjustment can still be useful at large sample sizes, and arguably can be more useful since it leads to greater absolute reductions in the required sample size. For example, the METS trial involved 146 participants and our estimate of $R^2_{Y-\Delta A\sim \boldsymbol{W}}$ is 4%. If this were the true value of $R^2_{Y-\Delta A\sim \boldsymbol{W}}$, it would mean about six fewer participants required to achieve the same power as the unadjusted estimator; if this trial were 10 times larger, that is, 1460 participants, then covariate adjustment would lead to a sample size reduction of approximately 60 participants.

Even when a randomized trial ends up having negligible imbalance, covariate adjustment can still increase power when the hypothesis test is based on dividing the estimator by its standard error and rejecting the null hypothesis when this ratio exceeds a threshold. This results from the fact that power is related to the variance through the standard error in the denominator of the test statistic. For example, even though the MCI trial is well balanced, there is still an estimated 25% variance reduction from adjustment. When there are prognostic baseline variables, we recommend that covariate adjustment be preplanned as the primary efficacy analysis.

If the outcome is binary, count, ordinal or time-to-event, then covariate adjustment can be done using estimators of, for example, Moore and van der Laan (2009), Lu and Tsiatis (2011), Howard et al. (2012) and Díaz et al. (2018). However, robust variance estimators typically must be used, for example, when constructing confidence intervals or conducting hypothesis tests. The sandwich estimator could be used as described by Tsiatis et al. (2008); alternatively, the nonparametric bootstrap could be used. Because of these results for other outcome types, it was surprising that when using ANCOVA it is unnecessary to use these robust variance estimators (since as proved in Section 4.2 the standard, model-based variance estimator for ANCOVA is already robust to arbitrary model misspecification).

We assumed equal randomization probabilities to the two trial arms. If unequal probabilities are used, then a robust variance estimator is needed for ANCOVA.

We also assumed that the data vector for each participant is an independent, identically distributed draw from an unknown distribution. This assumption does not hold if stratified randomization or covariate-adaptive randomization is used. For stratified randomization and some types of covariate-adaptive randomization, Bugni et al. (2018) showed that if the covariates in the ANCOVA model are the indicators of the strata used in the randomization procedure, then the ANCOVA estimator is consistent; furthermore, its model-based variance estimator is consistent if the limiting probability of assignment to each arm is $1/2$ within each stratum. The previous sentence holds regardless of whether the true data generating distribution satisfies any of the ANCOVA model assumptions. In general, adjusting for stratification variables is recommended when using stratified randomization or covariate-adaptive randomization (Lachin et al., 1988; Kahan & Morris, 2012; EMA, 2015). It is an open question, to the best of our knowledge, as to what happens when more variables than the stratification indicators are included in the ANCOVA model under such randomization schemes, in terms of consistency of the ANCOVA estimator and how to compute its asymptotic variance under arbitrary model misspecification.

How to best pick the set of covariates to use in an adjusted estimator is a challenging problem. The methods of Moore and van der Laan (2009) and Moore et al. (2011) use cross-validation, Bloniarz et al. (2016) and Tian et al. (2019) use LASSO, and Wager et al. (2016) use a combination of regression and cross-validation. All aspects of the covariate adjustment method need to be prespecified in the study protocol (FDA & EMA, 2015).

ACKNOWLEDGMENTS

The research was funded by NIH grants UL1TR001079 and R01AG048349. MCI data were obtained from the Alzheimer's Disease Cooperative Study (National Institutes of

Health Grant U19 5U19AG010483) legacy database. TADS and METS data were obtained from the controlled access datasets distributed from the NIMH-supported National Database for Clinical Trials (NDCT) with identifiers TADS (NCT00006286) #2145 and METS (NCT00816907) #2156. This manuscript reflects the views solely of the authors and does not reflect the views of any of the above.

ORCID

Bingkai Wang http://orcid.org/0000-0002-9349-2336
Elizabeth L. Ogburn http://orcid.org/0000-0001-7036-7973
Michael Rosenblum http://orcid.org/0000-0001-7411-4172

REFERENCES

Austin, P., Manca, A., Zwarenstein, M., Juurlink, D. and Stanbrook, M. (2010). A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: A review of trials published in leading medical journals. Journal of Clinical Epidemiology, 63, 142–153.

Bloniarz, A., Liu, H., Zhang, C., Sekhon, J. and Yu, B. (2016). Lasso adjustments of treatment effect estimates in randomized experiments. Proceedings of the National Academy of Sciences of the United States of America, 113, 7383–7390.

Borm, G., Fransen, J. and Lemmens, W. (2007). A simple sample size formula for analysis of covariance in randomized clinical trials. Journal of Clinical Epidemiology, 60, 1234–1238.

Bugni, F.A., Canay, I.A. and Shaikh, A.M. (2018). Inference under covariate-adaptive randomization. Journal of the American Statistical Association, 113, 1784–1796.

Díaz, I., Colantuoni, E., Hanley, D.F. and Rosenblum, M. (2018). Improved precision in the analysis of randomized trials with survival outcomes, without assuming proportional hazards. Lifetime Data Analysis, 1–30. https://doi.org/10.1007/s10985-018-9428-5.

Treatment for Adolescents With Depression Study (TADS) Team (2004). Fluoxetine, cognitive-behavioral therapy, and their combination for adolescents with depression: Treatment for adolescents with depression study (TADS) randomized controlled trial. JAMA, 292, 807–820.

EMA. (2015). European medicines agency guideline on adjustment for baseline covariates in clinical trials. Reference number EMA/CHMP/295050/2013. Committee for Medicinal Products for Human Use (CHMP).

FDA, and EMA. (1998). E9 statistical principles for clinical trials. U.S. Food and Drug Administration: CDER/CBER. European Medicines Agency: CPMP/ICH/363/96.

Howard, G., Waller, J., Voeks, J., Howard, V., Jauch, E., Lees, K., ..., Hess, D.C. (2012). A simple, assumption-free, and clinically interpretable approach for analysis of modified Rankin outcomes. Stroke, 43, 664–669.

Huitema, B. (2011). The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies. Hoboken, NJ: Wiley.

Jarskog, L., Hamer, R., Catellier, D., Stewart, D., LaVange, L., Ray, N., ..., Stroup, T.S. (2013). Metformin for weight loss and metabolic control in overweight outpatients with schizophrenia and schizoaffective disorder. American Journal of Psychiatry, 170, 1032–1040.

Jiang, F., Tian, L., Fu, H., Hasegawa, T. and Wei, L.J. (2018). Robust alternatives to ANCOVA for estimating the treatment effect via a randomized comparative study. Journal of the American Statistical Association, 0, 1–37.

Kahan, B.C. and Morris, T.P. (2012). Improper analysis of trials randomised using stratified blocks or minimisation. Statistics in Medicine, 31, 328–340.

Kraemer, H. (2015). A source of false findings in published research studies: Adjusting for covariates. JAMA Psychiatry, 72, 961–962.

Lachin, J., Matts, J. and Wei, L. (1988). Randomization in clinical trials: Conclusions and recommendations. Controlled Clinical Trials, 9, 365–374.

Lu, X. and Tsiatis, A. (2011). Semiparametric estimation of treatment effect with time-lagged response in the presence of informative censoring. Lifetime Data Analysis, 17, 566–593.

Ludvigsson, J., Faresjö, M., Hjorth, M., Axelsson, S., Chŕamy, M., Pihl, M., ..., Casas, R. (2008). GAD treatment and insulin secretion in recent-onset type 1 diabetes. New England Journal of Medicine, 359, 1909–1920.

Montalban, X., Hauser, S., Kappos, L., Arnold, D., Bar-Or, A., Comi, G., ..., Wolinsky, J.S. (2017). Ocrelizumab versus placebo in primary progressive multiple sclerosis. The New England Journal of Medicine, 376, 209–220.

Moore, K. and van der Laan, M. (2009). Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. Statistics in Medicine, 28, 39–64.

Moore, K., Neugebauer, R., Valappil, T. and van der Laan, M. (2011). Robust extraction of covariate information to improve estimation efficiency in randomized trials. Statistics in Medicine, 30, 2389–2408.

Petersen, R., Thomas, R., Grundman, M., Bennett, D., Doody, R., Ferris, S., ..., Thal, L.J. (2005). Vitamin E and donepezil for the treatment of mild cognitive impairment. The New England Journal of Medicine, 352, 2379–2388.

Pocock, S., Assmann, S., Enos, L. and Kasten, L. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practice and problems. Statistics in Medicine, 21, 2917–2930.

Robins, J., Sued, M., Lei-Gomez, Q. and Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. Statistical Science, 22, 544–559.

Rubin, D., and van der Laan, M. (2008). Covariate adjustment for the intention-to-treat parameter with empirical efficiency maximization. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 229, https://biostats.bepress.com/ucbbiostat/paper229.

Rubin, D., and van der Laan, M. (2011). Targeted ANCOVA estimator in RCTs. In van der Laan, M., & Rose, S. (Eds.), Targeted Learning: Causal Inference for Observationaland Experimental Data. New York, NY: Springer.

Tian, L., Jiang, F., Hasegawa, T., Uno, H., Pfeffer, M. and Wei, L.J. (2019). Moving beyond the conventional stratified analysis to estimate an overall treatment efficacy with the data from a comparative randomized clinical study. Statistics in Medicine, 38, 917–932.

Tsiatis, A., Davidian, M., Zhang, M. and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. Statistics in Medicine, 27, 4658–4677.

Wager, S., Du, W., Taylor, J. and Tibshirani, R. (2016). High-dimensional regression adjustments in randomized experiments. Proceedings of the National Academy of Sciences of the United States of America, 113, 12673–12678.

Yang, L. and Tsiatis, A. (2001). Efficiency study of estimators for a treatment effect in a pretest-posttest trial. The American Statistician, 55, 314–321.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Wang B, Ogburn EL, Rosenblum M. Analysis of covariance inrandomizedtrials: Moreprecision and valid confidence intervals, without model assumptions. *Biometrics.* 2019;75:1391–1400. https://doi.org/10.1111/biom.13062