Johns Hopkins University, Dept. of Biostatistics Working Papers

10-19-2018

Analysis of Covariance (ANCOVA) in Randomized Trials: More Precision, Less Conditional Bias, and Valid Confidence Intervals, Without Model Assumptions

Bingkai Wang
Department of Biostatistics, Johns Hopkins University

Elizabeth Ogburn
Department of Biostatistics, Johns Hopkins University

Michael Rosenblum
Department of Biostatistics, Johns Hopkins University, m.a.rosenblum@gmail.com

Analysis of Covariance (ANCOVA) in Randomized Trials: More Precision, Less Conditional Bias, and Valid Confidence Intervals, Without Model Assumptions

BINGKAI WANG, ELIZABETH OGBURN, MICHAEL ROSENBLUM*

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, Maryland 21205, USA

mrosen@jhu.edu

Summary

"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, e.g., 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 ANCOVA estimator, which involves fitting a linear model for the outcome given the treatment arm and baseline variables, and trials 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

*To whom correspondence should be addressed.

the linear model were correct are still valid even when the linear model is arbitrarily misspecified, e.g., 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 re-analyzing completed randomized trials for mild cognitive impairment, schizophrenia, and depression, we demonstrate how ANCOVA can reduce variance, reduce bias conditional on chance imbalance, and increase power even when by chance there is perfect balance across arms in the baseline variables.

Key words: Imbalance, Relative Efficiency, Robustness

1. Introduction

Pocock *and others* (2002) surveyed 50 randomized trial reports published between July and September, 1997. They found that 36 used covariate adjustment, but only 12 emphasized adjusted over unadjusted estimators. They also 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 *and others* (2010), in a paper entitled "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 randomized trial articles published between January and June, 2007. Only 39 out of 114 trials in their sample presented an adjusted analysis.

We focus on the analysis of covariance (ANCOVA) estimator, referred to as "ANCOVA I" by Yang and Tsiatis (2001). ANCOVA 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 (if it is used) often involves the ANCOVA estimator. We use the terms "covariate adjustment" and "adjusted

estimator" 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 (e.g., 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 *and others* (2008) and Montalban *and others* (2017) both checked whether data is normally distributed before applying ANCOVA and Ludvigsson *and others* (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 valid even when the linear model is arbitrarily misspecified, e.g., when the true relationships between variables are non-linear, and/or when there is treatment effect heterogeneity. This is important since it is not possible to rule out all types of model misspecification. Also, it means that precision gains from ANCOVA, which in our data examples range from modest (4%) to substantial (32%), can be retained even if the linear model is incorrect.

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 precisely 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 and theory from Efron and Hinkley (1978); Robins and Morgenstern (1987); Rosenbaum (1987); Robins *and others* (1992); Tsiatis *and others* (2008); Rubin and van der Laan (2008); Moore and van der Laan (2009); Moore *and others* (2011); Rubin and van der Laan (2011); Jiang *and others* (2016); Tian *and others* (2016). As in their work, our results are asymptotic, i.e., they hold in the limit as sample size grows to infinity while the number of covariates is fixed.

We present data analyses based on three completed randomized clinical trials for treatment of mild cognitive impairment (MCI) (Petersen *and others*, 2005), schizophrenia (Jarskog *and others*, 2013), and depression (Treatment for Adolescents With Depression Study (TADS) Team, 2004), respectively. By analyzing these data sets, we demonstrate how covariate adjustment can reduce (unconditional) variance, reduce bias conditional on the observed chance imbalance (called conditional bias), 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 adjusted estimator having smaller standard error than the unadjusted estimator).

In the next section, we describe the three trials. In Section 3, we define the unadjusted estimator (which ignores baseline variables), the ANCOVA estimator, and the chance imbalance (also called covariate imbalance). In Section 4, we present our main results, which involve the close relationship among these three statistics. Illustrations are provided in Sections 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 (MCI) Trial*

The "Vitamin E and Donepezil for the Treatment of Mild Cognitive Impairment" (MCI) phase 3 randomized trial was completed in 2004 (Petersen *and others*, 2005). The goal was to estimate the effect of a drug treatment on preventing progression from MCI to Alzheimer's disease. Participants were randomly assigned to three arms: the drug Donepezil, Vitamin E, and placebo control. For simplicity, 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, MiniMental State Examination (MMSE) score, Activities of Daily Living total score, Global Deterioration scale, and CDR-SB.

2.2 *Metformin for Weight Loss (METS) 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 *and others*, 2013). Participants were randomly assigned to 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. In our analysis, we use the same primary outcome and the following baseline variables: age, gender, Clinical Global Impressions (CGI) severity rating score, tobacco use, illicit drug use, alcohol use, baseline weight and body mass index (BMI).

2.3 *Treatment for Adolescents with Depression Study (TADS)*

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 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 co-primary 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, co-morbidity (indicator of any other psychiatric disorder except dysthymia).

2.4 *Summary of Data Analysis*

Our analyses of the three trials are summarized in Table 1. For the MCI and METS trials, we focus on a single treatment arm versus control. For TADS, we compare each of three treatment arms (FLX, CBT, CMB) versus control. The analyses in Table 1 are not adjusted for multiplicity.

We describe a few key results from Table 1; a complete discussion is given in Section 5. The ANCOVA estimator leads to variance reductions (equivalently, sample size reductions to achieve a desired power) of 4% to 32%. We next consider the conditional bias reduction due to covariate adjustment (defined in Section 5.1). There is no such reduction in the MCI trial; in contrast, for TADS(FLX), the conditional bias is reduced by 2.9 points of CDRS-R total score (which exceeds 1 standard error for the unadjusted estimator, so is quite substantial). In the latter

case, the ANCOVA estimator leads to a statistically significant result (p-value 0.01), unlike the unadjusted estimator (p-value 0.73). This improvement is due to the substantial reductions in variance and conditional bias from covariate adjustment, explained in Sections 4 and 5.3.

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 (e.g., TADS), we consider each treatment arm vs. control comparison separately.

For each participant $i = 1, \ldots, n$, we observe the data vector $(\boldsymbol{W}_i, A_i, Y_i)$. Each data vector is assumed to be an independent draw from the unknown, joint distribution on generic data vector $(\boldsymbol{W}, A, Y)$. The goal is to estimate the population average treatment effect
$\Delta = E[Y|A = 1] - E[Y|A = 0]$, i.e., 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 randomized trials (Tsiatis *and others*, 2008).

The components of the baseline vector $\boldsymbol{W}$ can be continuous, binary, ordinal and/or categorical. All variables are assumed to be bounded. We assume the study arm A is randomly assigned with equal probability to treatment or control independent of the baseline variables $\boldsymbol{W}$, which holds by design in a randomized trial. We also assume that the components of $(1, \boldsymbol{W}^t)$ are linearly independent (since otherwise at least one is redundant and can be dropped from the corresponding design matrix). We do not assume any other relationships among the variables $\boldsymbol{W}, A, Y$. An

estimator of the average treatment effect Δ is called robust to arbitrary model misspecification if it is consistent under the aforementioned assumptions.

The unadjusted estimator of the average treatment effect Δ is the difference between sample means of the outcome in the treatment and control arms, denoted by

$$\hat{\Delta}^{unadj} = \frac{\sum_{i=1}^{n} Y_i A_i}{\sum_{i=1}^{n} A_i} - \frac{\sum_{i=1}^{n} Y_i (1 - A_i)}{\sum_{i=1}^{n} (1 - A_i)}.$$

The unadjusted estimator is consistent, i.e., converges to Δ as the sample size goes to infinity, but it ignores all information in the baseline variables.

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

$$E[Y|A, \boldsymbol{W}] = \beta_0 + \beta_A A + \boldsymbol{\beta}_{\boldsymbol{W}}^t \boldsymbol{W}, \tag{3.1}$$

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

According to Huitema (2011), the ANCOVA model assumes: (i) a linear relationship between the outcome and the other variables, i.e., $Y = \beta_0 + \beta_A A + \boldsymbol{\beta}_{\boldsymbol{W}}^t \boldsymbol{W} + \varepsilon$, 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 *and others*, 2010), or if the outcome is non-linearly 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, i.e., $\hat{\beta}_A$ converges to Δ in probability, even under arbitrary misspecification of the linear model (3.1). Furthermore, the ANCOVA estimator is asymptotically normal and we denote its asymptotic variance as $Var^*(\hat{\Delta}^{ancova})$, i.e., $n^{1/2}(\hat{\Delta}^{ancova} - \Delta)$ converges to a normal distribution with mean

0 and variance $Var^*(\hat{\Delta}^{ancova})$. 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 $\hat{\Delta}^{ancova} = \hat{\beta}_A$ to estimate the average (also called marginal) treatment effect $\Delta = E[Y|A = 1] - E[Y|A = 0]$, which 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 $\hat{\beta}_A$ is the conditional treatment effect. That interpretation does not apply when the model is misspecified in any way. 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 Δ (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 and 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:

$$\boldsymbol{I} = \frac{\sum_{i=1}^{n} A_i \boldsymbol{W}_i}{\sum_{i=1}^{n} A_i} - \frac{\sum_{i=1}^{n} (1 - A_i) \boldsymbol{W}_i}{\sum_{i=1}^{n} (1 - A_i)}. \tag{3.2}$$

The imbalance $\boldsymbol{I}$ is a vector with the same number of components as $\boldsymbol{W}$. A component being 0 represents perfect balance in that variable. Although A is independent of $\boldsymbol{W}$ by design, in any realization (i.e., any real trial) the baseline variables can be imbalanced.

3.2 *ANCOVA 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 $E[Z|\boldsymbol{X}] = \beta_0 + \boldsymbol{\beta}_{\boldsymbol{X}}^t \boldsymbol{X}$. 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 $\hat{\beta}_0, \hat{\boldsymbol{\beta}}_{\boldsymbol{X}}$ (based on independent, identically distributed vectors $(\boldsymbol{X}_i, Z_i) : i = 1, \ldots, n$) still converges to a limit, denoted $\underline{\beta}_0, \underline{\boldsymbol{\beta}}_{\boldsymbol{X}}$. Assuming the components of $(1, \boldsymbol{X}^t)$ are linearly independent (otherwise one or more is redundant and can be dropped), 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 (3.1) 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 $(\hat{\beta}_A, \hat{\boldsymbol{\beta}}_{\boldsymbol{W}})$ for the linear model (3.1) 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 Supplementary Material:

$$Var(Y - \underline{\beta}_A A) = Var(\underline{\boldsymbol{\beta}}_{\boldsymbol{W}}^t \boldsymbol{W}) + Var(Y - \underline{\beta}_A A - \underline{\boldsymbol{\beta}}_{\boldsymbol{W}}^t \boldsymbol{W}). \tag{3.3}$$

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

$$R^2_{Y-\Delta A\sim \boldsymbol{W}} = 1 - Var(Y - \underline{\beta}_A A - \underline{\boldsymbol{\beta}}_{\boldsymbol{W}}^t \boldsymbol{W})/Var(Y - \underline{\beta}_A A). \tag{3.4}$$

The subscript in $R^2_{Y-\Delta A\sim \boldsymbol{W}}$ is to indicate that this R-squared represents the fraction of variance of $Y - \Delta A$ explained by $\boldsymbol{W}$, where we made the substitution $\underline{\beta}_A = \Delta$ on the right side of (3.4), 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 (3.1). 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 (3.3) 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. We present visualizations of the relationship among the imbalance, unadjusted estimator, and ANCOVA estimator in the Supplementary Material. These results build on ideas from prior work as described below.

4.1 *Connecting OLS Regression to the Relationship Among Unadjusted Estimator, ANCOVA Estimator, and Covariate Imbalance*

Jiang *and others* (2016) proved the following relationship among the unadjusted estimator $\hat{\Delta}^{unadj}$, ANCOVA estimator $\hat{\Delta}^{ancova}$, and chance imbalance $\boldsymbol{I}$:

$$\hat{\Delta}^{unadj} \approx \underline{\boldsymbol{\beta}}_{\boldsymbol{W}}^{t} \boldsymbol{I} + \hat{\Delta}^{ancova}. \tag{4.5}$$

(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:

$$Var^*(\hat{\Delta}^{unadj}) = Var^*(\underline{\boldsymbol{\beta}}_{\boldsymbol{W}}^{t} \boldsymbol{I}) + Var^*(\hat{\Delta}^{ancova}), \tag{4.6}$$

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 (3.3) for OLS, under arbitrary model misspecification. Specifically, we prove in the Supplementary Material that each term in (4.6) equals 4 times the corresponding term in (3.3), i.e., $Var^*(\hat{\Delta}^{unadj}) = 4Var(Y - \underline{\beta}_A A)$, $Var^*(\underline{\boldsymbol{\beta}}_{\boldsymbol{W}}^{t} \boldsymbol{I}) = 4Var(\underline{\boldsymbol{\beta}}_{\boldsymbol{W}}^{t} \boldsymbol{W})$, and $Var^*(\hat{\Delta}^{ancova}) = 4Var(Y - \underline{\beta}_A A - \underline{\boldsymbol{\beta}}_{\boldsymbol{W}}^{t} \boldsymbol{W})$. This is summarized in Figure 1, where the first row is the variance decomposition in OLS, the second row is the variance decomposition of $\hat{\Delta}^{unadj}$ from Jiang *and others* (2016), 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 $\hat{\Delta}^{ancova}$ that is output by standard statistical software such as 'summary.lm' in R or 'proc reg' in SAS, which we denote by

$\widehat{Var}(\hat{\Delta}^{ancova})$. The formula for $\widehat{Var}(\hat{\Delta}^{ancova})$ is

$$\widehat{Var}(\hat{\Delta}^{ancova}) = \frac{\widehat{Var}(Y - \hat{\beta}_0 - \hat{\beta}_A A - \hat{\boldsymbol{\beta}}_{\boldsymbol{W}}^t \boldsymbol{W})}{(n-1)[\widehat{Var}(A) - \widehat{Cov}(\boldsymbol{W}, A)^t \widehat{Var}(\boldsymbol{W})^{-1} \widehat{Cov}(\boldsymbol{W}, A)]} \tag{4.7}$$

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 Supplementary Material for precise definitions of these.) The following theorem shows that the above variance estimator is robust to arbitrary model misspecification.

Theorem Given the assumptions in Section 3.1, which do not assume that the linear model (3.1) is correctly specified, n times the estimated variance $\widehat{Var}(\hat{\Delta}^{ancova})$ converges in probability to the true asymptotic variance $Var^*(\hat{\Delta}^{ancova})$ of the ANCOVA estimator $\hat{\Delta}^{ancova}$.

The above theorem implies that confidence intervals and Wald-type hypothesis tests conducted as if the linear model were correct are valid even when the linear model is arbitrarily misspecified. The $1-\alpha$ confidence interval for the coefficient on the A term in (3.1) that is output by the aforementioned, standard linear regression software is

$$\left(\hat{\Delta}^{ancova} - t_{n-p,\alpha/2}\sqrt{\widehat{Var}(\hat{\Delta}^{ancova})}, \hat{\Delta}^{ancova} + t_{n-p,\alpha/2}\sqrt{\widehat{Var}(\hat{\Delta}^{ancova})}\right), \tag{4.8}$$

where $t_{n-p,\alpha/2}$ is the $\alpha/2$-quantile of the t-distribution with $n-p$ degrees of freedom where p is the number of coefficients in the linear model (3.1). For large n and fixed p, the quantile $t_{n-p,\alpha/2}$ is approximately the $\alpha/2$-quantile of the standard normal distribution. It follows from the above theorem that the above display is a valid confidence interval for the average treatment effect Δ, under arbitrary model misspecification.

We next prove the above theorem. The term in square brackets in the denominator of (4.7) converges to 1/4 since (i) the variance of A is 1/4, and (ii) A and W being independent (by randomization) implies that they have zero covariance. The numerator of (4.7), i.e., the residual variance from fitting model (3.1), converges to $Var(Y - \underline{\beta}_A A - \underline{\boldsymbol{\beta}}_{\boldsymbol{W}}^t \boldsymbol{W})$. Therefore, the above

theorem follows from the equality $Var^*(\hat{\Delta}^{ancova}) = 4Var(Y - \underline{\beta}_A A - \underline{\boldsymbol{\beta}}_{\boldsymbol{W}}^t \boldsymbol{W})$ in the previous subsection.

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 *and others* (2007) and Rubin and van der Laan (2008) connect the R-squared from regressing Y on $\boldsymbol{W}$ to the variance reduction due to ANCOVA, while Moore and van der Laan (2009) and Moore *and others* (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-squared to the variance reduction due to covariate adjustment. (See the Supplementary Material for more details.) Building on key ideas from their approaches, we prove that the R-squared $R^2_{Y-\Delta A\sim \boldsymbol{W}}$ equals the variance reduction due to ANCOVA without requiring these conditions or extra factors; this R-squared (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.,

$$R^2_{Y-\Delta A\sim \boldsymbol{W}} = 1 - \frac{Var(Y - \underline{\beta}_A A - \underline{\boldsymbol{\beta}}_{\boldsymbol{W}}{}^t \boldsymbol{W})}{Var(Y - \underline{\beta}_A A)} = 1 - \frac{Var^*(\hat{\Delta}^{ancova})}{Var^*(\hat{\Delta}^{unadj})}. \tag{4.9}$$

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 (expression on the right). The rightmost 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 computing $1-\widehat{Var}(\hat{\Delta}^{ancova})/\widehat{Var}(\hat{\Delta}^{unadj})$, where $\widehat{Var}(\hat{\Delta}^{ancova})$ is the variance of the ANCOVA estimator output by standard statistical software

as in (4.7), and $\widehat{Var}(\hat{\Delta}^{unadj})$ is the variance of the unadjusted estimator estimated analogously (by regressing Y on A and an intercept).

The variance reduction (4.9) 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 $\hat{R}^2_{Y-\Delta A\sim \boldsymbol{W}}$ values from our data sets range from 4% to 32%, which can be translated into 4% to 32% sample size reductions.

5. Clinical Trial Applications

In each application (MCI, METS, TADS), 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 0.1.

5.1 *MCI Trial*

The unadjusted treatment effect estimate was $\hat{\Delta}^{unadj} = -0.2$ CDR-SB points with standard error 0.2 and 95% CI $(-0.5, 0.1)$, and the ANCOVA estimate was $\hat{\Delta}^{ancova} = -0.2$ CDR-SB points with standard error 0.1 and 95% CI $(-0.4, 0.1)$. 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.

Jiang *and others* (2016) proved that adjustment for baseline variables can reduce bias conditional on the chance imbalance $\boldsymbol{I}$ (called conditional bias). They showed that the unadjusted es-

timator has conditional bias $\underline{\boldsymbol{\beta}}_{\boldsymbol{W}}^t \boldsymbol{I}$ but the ANCOVA estimator is conditionally unbiased, asymptotically. In other words, ANCOVA approximately removes the conditional bias of the unadjusted estimator that is caused by chance imbalance, and this bias is approximately the dot product of the chance imbalance $\boldsymbol{I}$ with the regression coefficients $\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}$. This statistic $\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}^t \boldsymbol{I}$ distills the imbalance information typically presented in a large table (often "Table 1") of a clinical trial report into a single number that is directly relevant to interpreting the unadjusted estimator; it represents this estimator's bias due to chance imbalance across study arms of multiple baseline variables. In contrast, the ANCOVA estimator is immune to such conditional bias, asymptotically.

Table 2 shows the decomposition of the conditional bias in the MCI trial into contributions from each baseline variable. The "Coeff. $\hat{\beta}_j$" row displays the regression model coefficients $\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}$. The $\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}$ components combined with the correlations among the components of $\boldsymbol{W}$ determine the variance reduction due to adjustment for $\boldsymbol{W}$. (See Figure 1 and the Supplementary Material.) Roughly speaking, larger magnitudes of $\hat{\beta}_j$ generally correspond to larger contributions to the variance reduction, but correlations among the baseline variables need to be factored in as well. Strong correlations among components of $\boldsymbol{W}$ are not necessarily detrimental (though they may indicate that some variables are redundant); what matters in terms of the variance reduction due to adjusting for $\boldsymbol{W}$ is the fraction of variance explained $R^2_{Y-\Delta A\sim \boldsymbol{W}}$, which is defined in equation (4.9).

The "Imbal. I_j" row gives the components of the imbalance vector $\boldsymbol{I}$. Since we first standardized each covariate before fitting the regression model, the imbalance of each covariate is approximately normally distributed $N(0, \sqrt{1/n_1 + 1/n_0})$, where $n_1 = \sum_{i=1}^n A_i, n_0 = \sum_{i=1}^n (1 - A_i)$. For MCI, each variable's imbalance has standard error 0.1. The "Correction i.e. $\hat{\beta}_j I_j$" is given by the product of each component of the imbalance $\boldsymbol{I}$ and the corresponding regression coefficient $\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}$, which represents the contribution of each baseline variable toward correcting (i.e., removing the conditional bias of) the unadjusted estimator. In the MCI trial, due to the relatively small values

in $\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}$ and each variable's imbalance being approximately 1 standard error or less in magnitude, the overall conditional bias reduction ($\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}^{t}\boldsymbol{I} = -0.003$ CDR-SB points) is small.

5.2 *METS Trial*

The unadjusted treatment effect estimate is $\hat{\Delta}^{unadj} = -3.7$ kg of weight change with standard error 1.6 and 95% C.I. $(-6.8, -0.5)$, and the ANCOVA estimate is $\hat{\Delta}^{ancova} = -3.6$ with standard error 1.6 and 95% C.I. $(-6.7, -0.5)$. Adjustment resulted in a 4% variance reduction.

Table 2 shows that several of the baseline variables were imbalanced (each variable's imbalance is expected to have standard error 0.2). Because baseline weight and BMI evinced substantial imbalance and large absolute regression coefficient values, they contribute most to the conditional bias. Due to their opposite signs, however, their effects cancel out, yielding a relatively small overall conditional bias (as evidenced by the similar values of $\hat{\Delta}^{unadj}$ and $\hat{\Delta}^{ancova}$).

5.3 *TADS Trial*

As shown in Table 1, covariate adjustment results in substantial variance reduction and conditional bias 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 -4.4 CDRS-R points (p-value 0.01); the 95% C.I. of the ANCOVA estimator $(-8.1, -0.6)$ excludes zero, but that of the unadjusted estimator $(-6.0, 3.2)$ does not. This results from both conditional bias reduction and variance reduction. Bias reduction increases the estimated treatment effect magnitude and variance reduction narrows the confidence interval width.

Table 2 shows the contribution of each covariate to the conditional bias reductions. For each of the three TADS treatment arms, the covariate imbalance has approximate standard error 0.2. In each treatment comparison, the baseline score for the primary outcome, baseline CDRS-R score,

contributes most to the conditional bias reductions. This is the case even when baseline CDRS-R does not exhibit a relatively large imbalance, and is due to its high regression coefficient. For each treatment versus control comparison, the estimated conditional bias (Table 1, column 4) is non-negligible, and all are in the same direction, suggesting that there is conditional bias in the control arm (1.2 CDRS-R points, estimated by using methods of Jiang *and others* (2016)).

The conditional bias reduction due to a covariate can also be interpreted as a correction of the unadjusted estimator that is reflected in the ANCOVA estimator. For example, in the TADS(FLX) section of Table 2, the standardized baseline CDRS-R score has imbalance -0.2 CDRS-R points, indicating that, on average, the treatment (FLX) arm has lower baseline depression scores compared to the control arm. The estimated regression coefficient is -8.1, and its sign being negative means that lower baseline scores are correlated with smaller values of the change score Y. The unadjusted estimator (which ignores this imbalance) is biased against the treatment arm (which, roughly speaking, started at lower levels of depression and so has less room to improve compared to control) conditioned on the observed imbalance in baseline CDRS-R score. The ANCOVA estimator removes this bias by subtracting approximately 2.0 CDRS-R points. Similar adjustments were made for the other variables, whose combined impact is to subtract 2.9 points from the unadjusted estimator.

Sometimes in a trial, the ANCOVA and unadjusted estimators may be substantially different. For example, in the TADS(FLX) trial, the unadjusted estimator is -1.4 CDRS-R points and the ANCOVA estimator is -4.4 CDRS-R points, which yields a large difference between the two estimates. This is the type of case where adjustment adds the most value over the unadjusted estimator. The large difference is approximately the conditional bias of the unadjusted estimator. Covariate adjustment removes (approximately) this conditional bias. Due to having less conditional bias and lower (unconditional) variance than the unadjusted estimator, the ANCOVA estimator should be given more credence when these estimators differ substantially.

6. Practical Recommendations

Consider the case where the primary outcome Y is a change score (difference between final score and baseline score). As demonstrated in the 3 clinical trial examples in Section 5, the baseline score often has a substantial correlation with the change score and so we recommend adjusting for the baseline score (and possibly additional variables). In some cases, adjusting for the baseline score alone can bring 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. E.g., 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 participants who benefit more/less from treatment are more likely to drop out than those who do not. The ANCOVA estimator remains consistent under missing at random if the ANCOVA model is correctly specified. To add robustness to model misspecification, one can use a propensity score model for missing outcomes (modeling the probability of missingness given treatment assignment and covariates with, e.g., a logistic regression model) as the inverse weight when fitting the ANCOVA model among those with observed outcomes. According to Robins *and others* (2007), this estimator is doubly-robust, i.e., consistent as long as one of the two models (propensity score model or ANCOVA model) is correctly specified. 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 Supplementary Material for details.

For large trials, e.g., 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. 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 the sample size reduction due to covariate adjustment is 4%, which means 6 fewer participants are required to achieve the same power. If this trial were 10 times larger, i.e., 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 (s.d.) and rejecting the null hypothesis when this ratio exceeds a threshold. This results from the fact that power is related to the (unconditional) variance through the standard error in the denominator of the test statistic. For example, the MCI trial is well balanced, with estimated conditional bias -0.003 CDR-SB points. There is still an estimated 25% variance reduction from adjustment. When there are prognostic baseline variables, we recommend that covariate adjustment (e.g., with ANCOVA) 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, e.g., Moore and van der Laan (2009), Lu and Tsiatis (2011), Howard *and others* (2012) and Díaz *and others* (2018). However, robust variance estimators typically must be used, e.g., when constructing confidence intervals or conducting hypothesis tests. The sandwich estimator could be used as described by Tsiatis *and others* (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 such 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).

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 *and others* (2011) use cross-validation, Tian *and others* (2016) and Bloniarz *and others* (2016) use LASSO, and Wager *and others* (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 and EMA, 1998).

Supplementary Material

Supplementary Material, available at `http://people.csail.mit.edu/mrosenblum/ANCOVA.pdf` includes the following: (a) definitions of the sample variance and covariance used in Section 4.2; (b) simulations to generate visualizations of the relationship among the imbalance, unadjusted estimator, and ANCOVA estimator; (c) proofs of theoretical results; (d) relationship among different types of R-squared; (e) data analyses accounting for missing data; (f) link to the code for data analysis; and (g) information monitoring to achieve sample size reductions with covariate adjustment.

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.

References

Austin, P.C., Manca, A., Zwarenstein, M., Juurlink, D.N. and Stanbrook, M.B. (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. *J Clin Epidemiol* **63**(2), 142 – 153.

Bloniarz, A., Liu, H., Zhang, C., Sekhon, J.S. and Yu, B. (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proc Natl Acad Sci USA* **113**(27), 7383–7390.

Borm, G.F., Fransen, J. and Lemmens, W.A.J.G. (2007). A simple sample size formula for analysis of covariance in randomized clinical trials. *J Clin Epidemiol* **60**(12), 1234 – 1238.

Díaz, I., Colantuoni, E., Hanley, D. and Rosenblum, M. (2018, 02). Improved precision in the analysis of randomized trials with survival outcomes, without assuming proportional hazards. *Lifetime Data Anal*, 1–30.

Efron, B. and Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika* **65**(3), 457.

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.L., Voeks, J.H., Howard, V.J., Jauch, E.C., Lees, K.R., Nichols, F.T., Rahlfs, V.W. and Hess, D.C. (2012). A simple, assumption-free, and clinically interpretable approach for analysis of modified Rankin outcomes. *Stroke* **43**(3), 664–669.

Huitema, B. (2011). *The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, 2nd Edition*. WILEY.

Jarskog, L.F., Hamer, R.M., Catellier, D.J., Stewart, D.D., LaVange, L., Ray, N., Golden, L.H., Lieberman, J.A. and Stroup, T.S. (2013). Metformin for weight loss and metabolic control in overweight outpatients with schizophrenia and schizoaffective disorder. *American Journal of Psychiatry* **170**(9), 1032–1040.

Jiang, F., Tian, L., Fu, H., Hasegawa, T., Pfeffer, M.A. and Wei, L.J. (2016). Robust alternatives to ANCOVA for estimating the treatment effect via a randomized comparative study. *Harvard University Biostatistics Working Paper Series.* **Working paper 209**, https://biostats.bepress.com/harvardbiostat/paper209.

Kraemer, H.C. (2015). A source of false findings in published research studies: Adjusting for covariates. *JAMA Psychiatry* **72**(10), 961–962.

Lu, X. and Tsiatis, A.A. (2011). Semiparametric estimation of treatment effect with time-lagged response in the presence of informative censoring. *Lifetime Data Anal* **17**(4), 566–593.

Ludvigsson, J., Faresjö, M., Hjorth, M., Axelsson, S., Chŕamy, M., Pihl, M., Vaarala, O., Forsander, G., Ivarsson, S., Johansson, C., Lindh, A., Nilsson, N., Åman, J., Örtqvist, E., Zerhouni, P. *and others.* (2008). GAD treatment and insulin secretion in recent-onset type 1 diabetes. *New England Journal of Medicine* **359**(18), 1909–1920.

Montalban, X., Hauser, S.L., Kappos, L., Arnold, D.L., Bar-Or, A., Comi, G., de Seze, J., Giovannoni, G., Hartung, H.P., Hemmer, B., Lublin, F., Rammohan, K.W., Selmaj, K., Traboulsee, A., Sauter, A., Masterman, D., Fontoura, P., Belachew, S., Garren, H., Mairon, N., Chin, P. *and others.* (2017). Ocrelizumab versus placebo in primary progressive multiple sclerosis. *New England Journal of Medicine* **376**(3), 209–220.

Moore, K.L., Neugebauer, R., Valappil, T. and van der Laan, M.J. (2011). Robust extraction of covariate information to improve estimation efficiency in randomized trials. *Stat Med* **30**(19), 2389–2408.

Moore, K.L. and van der Laan, M.J. (2009). Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *Stat Med* **28**(1), 39–64.

Petersen, R.C., Thomas, R.G., Grundman, M., Bennett, D., Doody, R., Ferris, S., Galasko, D., Jin, S., Kaye, J., Levey, A., Pfeiffer, E., Sano, M., van Dyck, C.H. *and others*. (2005). Vitamin E and Donepezil for the Treatment of Mild Cognitive Impairment. *New England Journal of Medicine* **352**(23), 2379–2388. PMID: 15829527.

Pocock, S.J., Assmann, S.E., Enos, L.E. and Kasten, L.E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med* **21**(19), 2917–2930.

Robins, J.M., Mark, S.D. and Newey, W.K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* **48**(2), 479–495.

Robins, J.M. and Morgenstern, H. (1987). The foundations of confounding in epidemiology. *Computers & Mathematics with Applications* **14**(9), 869 – 916.

Robins, J., Sued, M., Lei-Gomez, Q. and Rotnitzky, A. (2007, 11). Comment: Performance of double-robust estimators when inverse probability weights are highly variable. *Statist. Sci.* **22**(4), 544–559.

Rosenbaum, P.R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association* **82**(398), 387–394.

Rubin, D.B. and van der Laan, M.J. (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.B. and van der Laan, M.J. (2011). Targeted ancova estimator in rcts. In: van der Laan, M.J. and Rose, S. (editors), *Targeted Learning: Causal Inference for Observational and Experimental Data*, Chapter 12. New York, NY: Springer, pp. 201–215.

Tian, L., Jiang, F., Hasegawa, T., Uno, H., Pfeffer, M.A. and Wei, L.J. (2016). Moving beyond the conventional stratified analysis to estimate an overall treatment efficacy with the data from a comparative randomized clinical study. *Harvard University Biostatistics Working Paper Series* **Working paper 208**, https://biostats.bepress.com/harvardbiostat/paper208.

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**(7), 807–820.

Tsiatis, A.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. *Stat Med* **27**(23), 4658–4677.

Wager, S., Du, W., Taylor, J. and Tibshirani, R.J. (2016). High-dimensional regression adjustments in randomized experiments. *Proc Natl Acad Sci USA* **113**(45), 12673–12678.

Yang, L. and Tsiatis, A.A. (2001). Efficiency study of estimators for a treatment effect in a pretest-posttest trial. *The American Statistician* **55**(4), 314–321.

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), estimated conditional bias reduction due to adjustment, and estimated variance reduction due to adjustment. Negative (positive) estimates are in the direction of clinical benefit (harm).

Trial Name	Unadjusted Estimator (95% CI)	ANCOVA Estimator (95% CI)	Conditional Bias Reduction	Variance Reduction ($\hat{R}^2_{Y-\Delta A\sim \boldsymbol{W}}$)
MCI	-0.2 (-0.5, 0.1)	-0.2 (-0.4, 0.1)	0.0	25%
METS	-3.7 (-6.8, -0.5)	-3.6 (-6.7, -0.5)	-0.1	4%
TADS(FLX)	-1.4 (-6.0, 3.2)	-4.4 (-8.1, -0.6)	2.9	32%
TADS(CBT)	2.2 (-1.9, 6.4)	0.5 (-3.2, 4.2)	1.7	21%
TADS(CMB)	-6.6 (-11.0, -2.3)	-7.7 (-11.3, -4.0)	1.0	30%

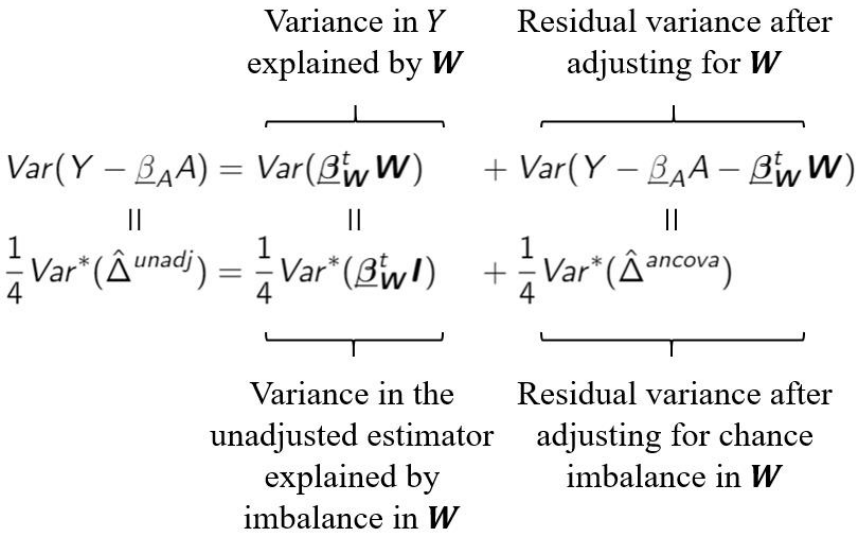

Fig. 1. Variance decomposition equivalence between linear regression and estimators of average treatment effect. The variance decomposition in the first row is a result of OLS linear regression. The second row gives the asymptotic variance decomposition of the unadjusted estimator, which is a minor extension of key results from Jiang *and others*, 2016; Tian *and others*, 2016. Our contribution is to connect the two variance decompositions by showing their equivalence, i.e., quantities in the same column are equal, under arbitrary model misspecification.

Table 2. Contribution of each baseline variable to overall conditional bias reduction (correction of the unadjusted estimate) in the MCI, METS and TADS trials. "Coeff.$\hat{\beta}_j$" displays the elements of $\hat{\boldsymbol{\beta}}_{\boldsymbol{W}}$. "Imbal.$I_j$" gives the imbalance vector $\boldsymbol{I}$. "Correction i.e., $\hat{\beta}_j I_j$" is the element-wise product of "Coeff.$\hat{\beta}_j$" and "Imbal.I_j". In MCI, baseline variables are age, gender, Alzheimer's Disease Assessment Scale-cognitive score (ADAS), MiniMental State Examination score (MMSE), Activities of Daily Living total score (ADLS), Global Deterioration scale (GDS) and baseline Clinical Dementia Rating-sums of boxes (CDR-SB). In METS, baseline variables are age, gender, Clinical Global Impressions severity rating scale (CGI), tobacco use (Tobacco), illicit drug use (Drug), alcohol use (Alcohol), baseline weight (Weight) and body mass index (BMI). In TADS, baseline variables are age, gender, baseline Children's Depression Rating Scale-Revised total score (CDRS-R), CGI, Children's Global Assessment Scale score (CGAS), Reynolds Adolescent Depression Scale total score (RADS), Suicide ideation, current major depressive episode duration (Depr. episode) and indicator of co-morbidity (Comor.).

		Age	Gender	ADAS	MMSE	ADLS	GDS	CDR-SB		
	Coeff. $\hat{\beta}_j$	0.1	0.1	0.4	-0.2	-0.4	0.1	-0.3		
MCI	Imbal.I_j	0.0	-0.1	0.0	-0.1	0.1	-0.1	-0.1		
	Correction i.e., $\hat{\beta}_j I_j$	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
		Age	Gender	CGI	Tobacco	Drug	Alcohol	Weight	BMI	
	Coeff. $\hat{\beta}_j$	0.0	0.7	1.6	-0.7	0.6	0.8	-3.4	2.5	
METS	Imbal.I_j	-0.3	0.1	0.0	0.1	0.0	-0.3	-0.2	-0.1	
	Correction i.e., $\hat{\beta}_j I_j$	0.0	0.1	0.0	-0.1	0.0	-0.3	0.5	-0.4	
		Age	Gender	CDRS-R	CGI	CGAS	RADS	Suicide ideation	Depr. episode	Comor.
	Coeff. $\hat{\beta}_j$	1.4	-0.1	-8.1	-0.5	0.0	-1.0	-1.6	0.8	-0.4
TADS	Imbal.I_j	0.1	0.0	-0.2	-0.3	-0.1	-0.3	-0.2	0.0	-0.2
(FLX)	Correction i.e., $\hat{\beta}_j I_j$	0.1	0.0	2.0	0.1	0.0	0.3	0.4	0.0	0.1
	Coeff. $\hat{\beta}_j$	1.4	-0.5	-4.2	-1.8	1.1	0.4	-2.5	2.1	1.3
TADS	Imbal.I_j	0.2	0.2	-0.2	-0.2	-0.3	-0.1	-0.2	0.1	0.0
(CBT)	Correction i.e., $\hat{\beta}_j I_j$	0.3	-0.1	1.0	0.3	-0.3	0.0	0.5	0.1	0.0
	Coeff. $\hat{\beta}_j$	1.3	-0.2	-8.0	-0.3	0.4	-1.0	0.5	1.0	0.3
TADS	Imbal.I_j	0.2	0.2	-0.1	-0.2	-0.1	-0.1	0.0	0.1	0.0
(CMB)	Correction i.e., $\hat{\beta}_j I_j$	0.2	0.0	0.6	0.1	0.0	0.1	0.0	0.2	0.0