A Comparison of Covariate Adjustment Approaches Under Model Misspecification In Individually Randomized Trials

Mia S. Tackney (✉ Mia.Tackney@lshtm.ac.uk)

London School of Hygiene and Tropical Medicine <https://orcid.org/0000-0003-3868-0550>

Tim Morris

MRC Clinical Trials Unit at UCL, London

Ian White

MRC Clinical Trials Unit at UCL, London

Clemence Leyrat

London School of Hygiene and Tropical Medicine

Karla Diaz-Ordaz

London School of Hygiene and Tropical Medicine

Elizabeth Williamson

London School of Hygiene and Tropical Medicine

Research Article

Keywords: Covariate adjustment, Randomized controlled trials, Misspecification, ANCOVA, G-computation, IPTW, AIPTW, TMLE

Posted Date: January 4th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1053600/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

RESEARCH

A Comparison of Covariate Adjustment Approaches under Model Misspecification in Individually Randomized Trials

Mia S. Tackney^{1*}
 , Tim Morris²
 , Ian White²
 , Clemence Leyrat¹
 , Karla Diaz-Ordaz¹
 and Elizabeth Williamson¹

*Correspondence:

Mia.Tackney@lshtm.ac.uk

¹Department of Medical Statistics,
 London School of Hygiene and
 Tropical Medicine, London, UK
 Full list of author information is
 available at the end of the article

Abstract

Adjustment for baseline covariates in randomized trials has been shown to lead to gains in power and can protect against chance imbalances in covariates. For continuous covariates, there is a risk that the form of the relationship between the covariate and outcome is misspecified when taking an adjusted approach. Using a simulation study focusing on small to medium-sized individually randomized trials, we explore whether a range of adjustment methods are robust to misspecification, either in the covariate–outcome relationship or through an omitted covariate–treatment interaction. Specifically, we aim to identify potential settings where G-computation, Inverse Probability of Treatment Weighting (IPTW), Augmented Inverse Probability of Treatment Weighting (AIPTW) and Targeted Maximum Likelihood Estimation (TMLE) offer improvement over the commonly used Analysis of Covariance (ANCOVA). Our simulations show that all adjustment methods are generally robust to model misspecification if adjusting for a few covariates, sample size is 100 or larger, and there are no covariate–treatment interactions. When there is a non-linear interaction of treatment with a skewed covariate and sample size is small, all adjustment methods can suffer from bias; however, methods that allow for interactions (such as G-computation with interaction and IPTW) show improved results compared to ANCOVA. When there are a high number of covariates to adjust for, ANCOVA retains good properties while other methods suffer from under- or over-coverage. An outstanding issue for G-computation, IPTW and AIPTW in small samples is that standard errors are underestimated; development of small sample corrections is needed.

Keywords: Covariate adjustment; Randomized controlled trials; Misspecification; ANCOVA; G-computation; IPTW; AIPTW; TMLE

Background

Whether to adjust for baseline covariates in the analysis of pragmatic randomized clinical trials is a question that has attracted controversy. In pragmatic trials, the aim is to estimate the marginal effect of the treatment. While unadjusted analyses in individually randomized trials are unbiased on average, there are several reasons

why covariate adjusted approaches are attractive. Firstly, if covariates are used in the randomization procedure by, for example, permuted blocks or minimization, it is necessary to adjust for the covariate (Kahan and Morris, 2012). Secondly, adjusting for covariates that are not used for randomization can lead to statistical advantages. Adjustment for covariates that are correlated with the outcome (prognostic covariates), such as the outcome measured at baseline, typically leads to increases in power. Kahan *et al.* (2014) showed that adjustment for prognostic covariates leads to substantial increases in power for moderate to large trials for continuous, binary and time-to-event outcomes. Covariate adjustment can offer protection against chance imbalance in the distribution of the covariates between treatment groups, which is particularly relevant for smaller to medium sized trials (Senn, 2013). Guidelines for clinical trials generally advocate adjustment of a select few, pre-specified covariates to avoid potential bias in estimation due to post-hoc selection of covariates and “fishing” for covariates that impact the statistical significance of the treatment effect (Tsiatis *et al.*, 2008).

Covariate adjustment is often achieved by a regression approach by modeling the effects of the treatment and covariates. We refer to this in the continuous outcome case as the Analysis of Covariance (ANCOVA) and in the binary outcome case as direct regression adjustment. The marginal treatment effect of interest may be a parameter of the model, or it may be a derived quantity of the model. For estimands that are collapsible, such as the difference in means for a continuous outcome or the risk difference for a binary outcome, the marginal effect of treatment is a parameter of the model. For non-collapsible estimands such as the odds ratio for a binary outcome, adjusting for covariates changes the estimand (Daniel *et al.*, 2021; Hauck *et al.*, 1998), so the marginal treatment effect may be a derived quantity. When interactions exist between treatments and covariates, a regression-based approach does not allow the direct estimation of the marginal effect, so the marginal effect is a derived quantity.

Practitioners may be reluctant to adopt a covariate-adjusted approach (Kahan and Morris, 2012), due to the potential for mis-specifying the model relating the outcome, treatment and covariates. This issue is particularly pronounced when covariates are continuous, since the functional form of the relationship between the covariate and outcome needs to be specified. Misspecification of this functional form could potentially lead to reduced power, and could lead to bias for non-continuous outcomes, or for continuous outcomes where sample size is small (Yang and Tsiatis, 2001). There may also be reluctance to adopt an adjusted approach due to the loss in degrees of freedom. The European Medicines Agency (2015) recommend using a simple functional form (e.g. linear or categorization) if the relationship between a continuous covariate and outcome is unknown, and discourage the inclusion of covariate–treatment interactions. Recent draft guidelines from the Food and Drug Administration (2021) suggest that interactions may be included, but the primary analysis should still estimate the average treatment effect. Kahan *et al.* (2016) studied the impact of several adjustment methods, including categorization of continuous variables, modeling the effect of the covariate with a linear effect, with

fractional polynomials and cubic splines. They investigated the effect on power, bias and type I error of moderate to large trials ($n = 200$ to 600). Their recommendation is to use fractional polynomials or restricted cubic splines.

In addition to ANCOVA, we consider covariate adjustment methods that are less commonly used in the analysis of randomized trials: G-computation, also known as standardization or marginalization, which requires a model for the covariate–outcome relationship but targets the marginal estimand; Inverse Probability of Treatment Weighting (IPTW), which does not require modeling of the covariate–outcome relationship but instead models the treatment allocation mechanism in order to balance covariates between arms; and two approaches, Augmented Inverse Probability of Treatment Weighting (AIPTW) and Targeted Maximum Likelihood Estimation, which involve specification of both types of models but require only one to be consistently estimated.

In randomized trials, both unadjusted and a range of adjusted estimators of treatment effect can be shown to belong to a class of methods which produce consistent and asymptotically normal treatment effect estimators, irrespective of whether the covariate adjustment is correctly specified (Davidian *et al.*, 2005). White *et al.* (2021) cautioned against using non-canonical link functions (as might be done to estimate a non-standard estimand in a direct regression approach) as it can lead to bias under the null hypothesis. While there are a range of estimators that are protected against the risks of misspecification in sufficiently large samples, the properties of adjustment methods in small to medium sized trials have received limited attention.

In this study, we focus on the question of whether adjusting for continuous baseline covariates is beneficial in small to medium sized trials where there is risk of misspecification of the covariate–outcome relationship. We consider the specific case of a trial with a binary treatment where randomization is 1:1 on the individual level (no blocking/stratification is used), and the marginal treatment effect is of interest. We use a simulation study to explore the extent to which the known benefits of adjustment in large trials - gain in power while estimates remain unbiased and coverage remains at the nominal level - are retained in small to medium-sized trials in the presence of model misspecification. In particular, we wish to identify whether any of the lesser-known adjustment approaches offer improvement over the commonly used ANCOVA. As this study is designed to identify corner cases that tease out differences between these related approaches, our simulation study explores a number of extreme settings that are unlikely to be encountered in practice, but can provide insight into the properties of these methods.

Methods

We consider continuous or binary outcomes, Y . We denote the potential outcome when a participant is given treatment z by Y^z , where $z = 0$ is the control and $z = 1$ is the active treatment. We denote a baseline covariate by X . For a continuous outcome, the marginal treatment effect is defined by taking the difference between

the marginal mean of the outcomes under the active treatment, and the marginal mean of the outcomes under the control:

$$\mathbb{E}(Y^1) - \mathbb{E}(Y^0). \quad (1)$$

For a binary outcome, we consider two estimands of interest, the risk difference (RD):

$$\mathbb{P}(Y^1 = 1) - \mathbb{P}(Y^0 = 1), \quad (2)$$

and the marginal odds ratio (OR),

$$\frac{\mathbb{P}(Y^1 = 1)/\mathbb{P}(Y^1 = 0)}{\mathbb{P}(Y^0 = 1)/\mathbb{P}(Y^0 = 0)}. \quad (3)$$

For a continuous outcome, an unadjusted analysis involves fitting the following model, and taking the estimated coefficient $\hat{\beta}$ as the treatment effect estimate, which is the difference between the sample mean of the outcomes under the active treatment and the sample mean of the outcome under the control:

$$\mathbb{E}(Y \mid Z) = \alpha + \beta Z. \quad (4)$$

For a binary outcome, we consider two unadjusted models. Firstly, a binomial model with an identity link function to estimate the risk, where the left-hand side of Equation (4) is $\mathbb{P}(Y = 1 \mid Z)$, then the coefficient β is the risk difference. Secondly, a binomial model with a logit link function can be used to estimate the log-odds, where the left-hand side of Equation (4) is $\log \frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)}$ and β represents the marginal log odds ratio.

Regression approaches

The most common approach to covariate adjustment in trials is through an Analysis of Covariance (ANCOVA), where the expectation of the outcome given the treatment and covariate is specified by a linear model:

$$\mathbb{E}(Y \mid Z, X) = \alpha + \beta_x Z + \gamma X. \quad (5)$$

The treatment effect estimate is given by $\hat{\beta}_x$, where the subscript emphasizes that the coefficient for treatment is adjusted for the covariate value x . This model can be extended to include additional covariates and/or non-linear functions of covariates, in which case $\mathbf{X}$ is a vector including functions of the covariate values. The ANCOVA treatment estimate has very desirable robustness properties in large samples; it is consistent (Yang and Tsiatis, 2001), and its standard error is consistent where randomization is 1:1, even when the model is misspecified (Wang et al., 2019).

There are two ways in which the adjusted model in Equation (5) could be misspecified. Firstly, ANCOVA assumes that the relationship between the covariate and outcome is linear; in other words, the effect of a one-unit increase in the covariate on the outcome is constant for all values across the range of the covariate. The true underlying covariate–outcome relationship could be a more complex non-linear relationship. To address this issue of potential non-linearity, the model can be adapted to allow a more flexible specification involving splines, which can capture non-linearities in the covariate–outcome relationship. The range of the covariate is split into m sections and, within each section, the covariate–outcome relationship is specified by a cubic polynomial. The $m - 1$ resulting curves are joined at knots to create a smooth function. The addition of splines lead to an additional $m + 1$ degrees of freedom required to fit the model. In this study, we place knots at equally spaced quantiles of the covariate. Secondly, there may be interactions between the treatment and covariate that are not reflected in the model. While ANCOVA will lead to consistent estimators even if the model is misspecified in large samples (Tsiatis et al., 2008), the properties of estimators for smaller sample sizes are less known.

For a binary outcome, an analogous adjusted model can be specified for the risk or log-odds. As discussed earlier, covariate adjustment changes the estimand in the case of the odds ratio, so an adjusted regression model may not be a pragmatic approach. With the risk difference as the estimand, smaller sample sizes can lead to well-known convergence problems (Williamson et al., 2013).

G-computation

G-computation is a standardization approach which can be used to obtain an adjusted estimate of the marginal treatment effect. A model for the mean outcome in terms of Z and X is specified:

$$m(Z, X) = \mathbb{E}(Y \mid Z, X), \quad (6)$$

and used to predict the expected value of both potential outcomes for each individual. The mean outcome $\mathbb{E}(Y^1)$, under a possibly counterfactual assignment to treatment, is then estimated by the sample average of the predicted outcomes $\hat{Y}^1$, and analogously, the mean outcome under the control arm $\mathbb{E}(Y^0)$ is computed:

$$\hat{\mathbb{E}}(Y^z) = \frac{1}{n} \sum_{i=1}^n \hat{m}(Z_i = z, x_i). \quad (7)$$

The treatment effect estimate is the difference between the estimated mean outcomes under the two treatments. If $m(Z, X)$ is Equation (5), the resulting treatment effect estimate is equal to the ANCOVA estimate. However, the covariate–outcome relationship can be modelled separately in each treatment group, which is equivalent to including a main effect and interaction between the treatment and covariate in Equation (5), and marginalising (as described above) to obtain an overall estimate of treatment effect. A nonlinear covariate–outcome relationship could also be

specified, for example by the use of splines. An advantage of this approach is that it separates the final estimation of the treatment effect from the modeling of the outcome.

For binary outcomes, a binomial model with logit link can be used to predict the potential outcomes on the probability scale. The sample averages can be attained to estimate $\mathbb{P}(Y^1 = 1)$ and $\mathbb{P}(Y^0 = 1)$, and the odds ratio or risk difference can be computed. There are particular advantages to using G-computation for the binary outcome case. Firstly, if the summary measure of interest is the risk difference, convergence problems that affect the direct regression approaches can be avoided. Secondly, if the odds ratio is the estimand of interest, G-computation achieves covariate adjustment while retaining the marginal estimand; the issue of adjustment changing the estimand is avoided.

G-computation can be written as an M-estimator, which relies on large-sample approximations to derive standard errors and confidence intervals (Stefanski and Boos, 2002). The standard errors are underestimated when sample sizes are small (Fay and Graubard, 2001), which translate to undercoverage and false gains in power. Bartlett (2018) showed that the estimates of marginal means $\mathbb{E}(Y^z)$ are consistent for canonical Generalized Linear Models, even if the model is mis-specified. Therefore, in large samples, we expect the difference in marginal means (for the continuous outcome case) and the risk difference (in the binary outcome case) to be consistently estimated, even if the model is mis-specified.

IPTW

Propensity score-based methods have largely been used in observational studies to address confounding and selection bias; however, Williamson et al. (2014) demonstrated they lead to similar large-sample properties as ANCOVA, such as increases in power, when applied to randomized controlled trials. Inverse Probability-of-Treatment Weighting (IPTW) involves specifying a model for the propensity score, which is the probability that a participant is assigned the active treatment, given values of their covariates: $e(X) = P(Z = 1|X)$. It may seem counter-intuitive to estimate the propensity score in a simple trial setting, since randomization implies that the true propensity score is 0.5. However, chance imbalance of covariates will be reflected in estimated propensity scores, which can then be re-balanced using a weighting approach. The propensity score can be estimated using logistic regression, by modeling Z as a binomial distribution where:

$$\text{logit}\{e(X)\} = \delta + \kappa X. \quad (8)$$

For a continuous outcome, the adjusted effect of treatment is obtained by fitting a model for the mean outcome, such as in Equation (4), where the outcomes are weighted by the inverse probability of being assigned the treatment. Outcomes for participants that received the active treatment are weighted by $\frac{1}{e(x_i)}$, and outcomes

for participants that received the placebo are weighted by $\frac{1}{1-\hat{e}(x_i)}$. The estimated weighted mean is given by:

$$\hat{\mathbb{E}}(Y^z) = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{\hat{e}(x_i)^{z_i} (1 - \hat{e}(x_i))^{(1-z_i)}}, \quad (9)$$

and the treatment effect estimate is the difference between the estimated weighted mean outcomes under the two treatments. For binary outcomes, a binomial model is specified for the mean outcome instead, with a linear link function for the risk difference, or a logistic link function for the marginal odds ratio.

A major attraction of this approach is that it avoids modeling the covariate–outcome relationship, and the potential for covariate–treatment interactions does not need to be considered. Further, similarly to G-computation, a feature of IPTW for binary outcomes is that the marginal estimand for the odds ratio can be estimated. IPTW also belongs to the class of M-estimators whose variance estimators rely on large sample properties (Stefanski and Boos, 2002) which have been found to perform poorly in some small sample settings (Raad et al., 2020); small sample corrections have not yet been established.

AIPTW and TMLE

Finally, we consider two approaches, Augmented Inverse Probability-of-Treatment Weighting (AIPTW) and Targeted Maximum Likelihood Estimation (TMLE), that require a model for the covariate–treatment relationship as well as a model for the treatment assignment. They are known as *doubly robust* estimators as only one of the two models need to be correctly specified to be consistent for the treatment effect (van der Laan and Rubin, 2006).

Augmented Inverse Probability-of-Treatment Weighting (AIPTW) requires a model for the mean outcome, which is then used to obtain predictions of the potential outcomes, as in G-computation. It also requires a model for the propensity score so that inverse probability of treatment weights can be calculated. These weights are then used to add an *error-correcting* term to the G-computation estimator, which is the sum of weighted differences between the observed outcomes and predicted outcomes:

$$\hat{E}(Y^z) = \underbrace{\frac{1}{n} \sum_{i=1}^n \hat{m}(Z_i = z, x_i)}_{\text{G-computation estimator}} - \underbrace{\frac{1}{n} \sum_{i=1}^n \frac{\hat{m}(Z = z, x_i) - y_i}{\hat{e}(x_i)^{z_i} (1 - \hat{e}(x_i))^{(1-z_i)}}}_{\text{error-correcting term}}. \quad (10)$$

Similarly to G-computation and IPTW, AIPTW belongs to the class of M-estimators which rely on large sample properties for the variance estimator (Stefanski and Boos, 2002).

Targeted Maximum Likelihood Estimation (TMLE) requires an initial model of the covariate–outcome relationship, which could be a regression model as in G-computation, or it could be a flexible machine learning model (van der Laan and Rubin, 2006). A model for treatment assignment, such as Equation (8), is then specified to obtain propensity scores. The propensity scores are required to compute so-called *clever covariates* for each individual, which are then used to estimate the *fluctuation parameter* for the efficient influence function using a maximum likelihood procedure (Luque-Fernandez et al., 2018). The fluctuation parameter corrects the initial estimate for $\mathbb{E}(Y | Z, X)$. This *targeting step* optimizes the bias-variance trade-off for the treatment effect (Schuler and Rose, 2017). The difference between the average of predicted potential outcomes under the treatment and the average of predicted potential outcome under the control is then computed to obtain the marginal treatment effect estimate. Standard errors can be estimated using the efficient influence function evaluated for the empirical distribution, or through non-parametric approaches such as the bootstrap (Petersen et al., 2014). TMLE is asymptotically efficient if both the propensity score model and the model for the outcome are correctly specified (van der Laan and Rose, 2011).

A comparison of the models required in these methods are illustrated in Figure 1.

Simulations

We performed simulation studies to compare covariate adjustment methods where the covariate–outcome model has been misspecified in small- to medium-sized parallel design two-arm trials with 1:1 randomization. We note that our simulation settings include a number of extreme settings which are unlikely to be encountered or implemented in practice, such as the splines with 20 degrees of freedom, adjusting for 17 or more covariates, or the harmonic relationship between covariate and outcome. While unrealistic, exploring these settings allow us to pinpoint the settings

where one method of covariate adjustment may offer advantages over another.

In our [Main Simulation](#), we explored the setting with one covariate, no covariate–treatment interaction, and a continuous outcome. Three smaller simulation studies vary these three design features in turn. [Extension 1](#) expands the main simulation to consider multiple covariates, with a continuous outcome and no interaction. [Extension 2](#) adds a covariate–treatment interaction to the main simulation, with one covariate and a continuous outcome. [Extension 3](#) considers the setting of a binary outcome where there is a single covariate and no interaction. In each setting, we were interested in estimating the effect of treatment and comparing the following performance measures for a number of analysis approaches:

- Bias
- Coverage of the 95% confidence interval
- Model-based and empirical standard error
- Power
- Type I error rate.

In each setting, total sample sizes of 25, 50, and 100 were considered where possible, and 1000 repetitions of the simulation were performed. The simulation was performed in R ([R Core Team, 2021](#)). We provide an overview of the data generating mechanism, estimand and analysis approaches in each of the four settings. Full details of the data generating mechanisms are provided in Additional File 1, and R code is provided in Additional Files 2–4.

Main Simulation

In this setting, we generated a continuous outcome from the model,

$$Y_i = \alpha + \beta Z_i + f(X_i) + \epsilon_i, \quad (11)$$

where $\epsilon_i \sim N(0, 42)$, and the binary treatment Z_i takes value 1 for the active arm and 0 for the placebo arm. The treatment was allocated randomly with a 0.5 probability of a participant receiving the active treatment. We considered the case with a treatment effect ($\beta = 40$) and without treatment effect ($\beta = 0$). The covariate is generated according to $f(X_i)$, where X_i is drawn from a standard normal distribution and the function $f(\cdot)$ denotes five possible covariate–outcome relationships: *linear*, *two-tier*, *flattening*, *quadratic* and *harmonic*, as illustrated in [Figure 2](#). These relationships range from those which may realistically be encountered in trials, through to difficult distributions unlikely to be encountered in practice.

The average treatment effect, β , is the estimand of interest. We consider the following methods for estimating β :

- **Unadjusted analysis**, equivalent to a t-test as in [Equation \(4\)](#);
- **ANCOVA** as in [Equation \(5\)](#) using an F-test;
- **G-computation**, implemented using `stdGlm()` in the `stdReg` package ([Sjolander and Dahlqvist, 2020](#)), where a single model, as in [Equation \(5\)](#), is fitted to both arms;

- **G-computation with interaction** where separate models assuming linear effects of covariates are fitted for each arm;
- **IPTW** where the model for treatment assignment is as in Equation (8), implemented using `psw()` in the PSW package (Mao and Li, 2018). The standard errors are corrected to account for the propensity score estimation;
- **AIPTW**, where the model for treatment assignment is as in Equation (8) and the model for the outcome is as in Equation (5), implemented using `psw()` in the PSW package (Mao and Li, 2018);
- **TMLE** where the model for treatment assignment is as in Equation (8) and the model for the outcome is as in Equation (5), implemented using `tmle()` in the TMLE package (Gruber and van der Laan, 2012). The standard error is computed using the efficient influence function evaluated for the empirical distribution.

We explored the addition of splines in a selection of these methods for sample size 50 and 100: splines with 4 or 20 degrees of freedom in the regression approach, splines with 4 degrees of freedom in G-computation and splines with 4 degrees of freedom in IPTW. In all uses of splines in this study, knots are placed at equally spaced quantiles of the covariate. Splines are implemented using the `ns()` function in the `splines` package (R Core Team, 2021). For IPTW, splines are implemented with `PSweight()` in the `PSWeight` package (Zhou et al., 2021).

Extension 1: Multiple Covariates

This setting has a continuous outcome, multiple covariates and no interaction. We consider a scenario where 21 covariates are measured for each individual, of which 17 are continuous and 4 are binary. The covariates are generated to mimic typical demographic and health-related covariates in a trial setting. Briefly, 17 covariates (9 continuous and 4 binary) are predictive of the outcome, of which three continuous covariates are highly predictive of the outcome. There are four additional noise covariates. The outcome is generated with a number of linear and non-linear effects from the covariates and some covariate-covariate interactions, but no covariate-treatment interactions. We considered the case with a treatment effect ($\beta = 40$) and without treatment effect ($\beta = 0$). The estimand of interest is the the marginal difference in means. We explored adjusting for:

- the three highly predictive covariates only
- a larger selection of 17 potentially predictive covariates
- all 21 covariates, which include noise variables.

Due to the high number of parameters in the models for adjustment, we considered sample sizes of $n = 50$ and $n = 100$ only. For each of the three cases, the used the following analysis methods:

- Unadjusted analysis,
- ANCOVA,
- G-computation,
- G-computation with interaction,
- IPTW,
- AIPTW,
- TMLE.

Extension 2: Interaction

This setting has a continuous outcome, one covariate and a covariate–treatment interaction. Four different interaction settings were considered, illustrated in Figure 3. A single covariate was generated from a $N(0, 1)$ distribution. In the first setting, this covariate has a small interaction with the treatment. In the second setting, the covariate has a larger interaction in which the treatment effect changes direction. In the third setting, the covariate–outcome relationship has different shapes in each arm (exponential under the active treatment and linear under the placebo). Finally, in the last setting, the covariate is the square of a standard normal distribution, and therefore has a skewed distribution. There is no effect of the covariate on the outcome for the active treatment, but there is an effect under the placebo. We demonstrate in the [Appendix](#) that the bias due to misspecification in a model including a covariate–treatment interaction is likely to be pronounced where there is a skewed covariate with different types of misspecification in each arm, prompting the addition of this last scenario in our simulation.

Figure 3: Extension 2. True relationships between the continuous covariate and outcome with interaction are shown: small interaction, large interaction, different shapes, and effect absent in one group. The dots depict the datapoints in a single simulated dataset of 100 participants, where the red dots indicate the active treatment and the blue dots indicate the control.

We consider the following methods of estimating the treatment effect:

- Unadjusted analysis,

- ANCOVA,
- G-computation,
- G-computation with interaction,
- IPTW,
- AIPTW,
- TMLE.

Extension 3: Binary outcome

This setting has a binary outcome, one covariate and no interaction. We generate the covariate X from a standard normal distribution. The outcomes are generated using Equation 11 on the logit scale, where the function $f(\cdot)$ denotes five possible covariate–outcome relationships: *linear*, *two-tier*, *flattening*, *quadratic* and *harmonic*. Settings with a treatment effect (with a conditional odds ratio of 0.2) and without treatment effect were considered. Due to potential convergence issues in smaller sample sizes, we considered only the sample size $n = 100$.

We considered the following estimands of interest: the risk difference, the marginal odds ratio (for all methods except direct RA with logistic link), and the data generating conditional odds ratio (for direct RA with logistic link). We consider the following methods for estimating the effect of interest:

- Unadjusted binomial regression with linear link for the risk difference or logistic link for the marginal odds ratio;
- Direct regression adjustment (RA) with logistic link for the data generating conditional odds ratio. An adjusted binomial model with a linear link for the risk difference leads to convergence issues so is omitted;
- G-computation for the risk difference or marginal odds ratio;
- IPTW for the risk difference or marginal odds ratio;
- AIPTW is included for the risk difference, but omitted for the marginal odds ratio as it is not readily available in the software used;
- TMLE for the risk difference or marginal odds ratio.

Results

Main Simulation

Figure 4 displays the results for the Main Simulation for the analytic methods for $n = 100$. In all plots in the Main Simulation and the Extensions, Type I error rate mirrors the coverage levels so are omitted from figures. All analytic methods considered produce unbiased estimates in the Main Simulation, but the variability around the estimated effect is larger for the non-linear relationships. At this sample size, we observe that the adjustment methods generally achieve nominal coverage, except for G-computation with interaction, IPTW and AIPTW, which lead to undercoverage in the highly non-linear *quadratic* relationship. All adjustment methods produce gains in power for the *linear*, *two-tier* and *flattening* relationships. For the *quadratic* and *harmonic* relationships, which are poorly approximated by a linear relationship, adjustment methods which assume a linear covariate–outcome relationship provide no increase in power, but do not lead to any loss in power.

Figure 5 displays results for $n = 25$, and results for $n = 50$ are provided in Additional File 1. For these smaller sample sizes, we observe that standard errors produced by IPTW, G-computation and AIPTW are smaller than the empirical standard error, leading to undercoverage, high type I error and false gains in power.

In Figure 6, the results for ANCOVA, G-computation and IPTW with and without the use of splines are shown for $n = 100$ for the *linear*, *flattening* and *quadratic* relationships. Additional File 1 displays these results for $n = 50$. We observe that the use of splines leads to considerable gains in power for the non-linear relationships, and generally does not incur bias or affect type I error. When the relationship is linear, the addition of splines generally does not affect bias, type I error or power. We observe that the discrepancy between the model-based and empirical standard errors is particularly pronounced for the IPTW approach with splines and the G-computation approach with splines when $n = 50$ due to the additional parameters

in the model from the splines.

Figure 5: Main Simulation results (continuous outcome, single covariate, no interaction) for sample size 25. The performance of analytic methods in terms of bias, coverage and power for the five different covariate–outcome relationships are displayed. The effect of treatment is 40. Model-based standard errors are indicated in black and the empirical standard error is shown in red. Estimates are indicated with $\pm 1.96 \times$ Monte Carlo standard error bars. Note that the error bars are too small to be seen for power, due to the scale of the plots.

Extension 1: Multiple covariates

Figure 7 displays the results for the continuous outcome case with multiple covariates when $n = 50$. All methods are unbiased regardless of the number of predictors, and all adjustment methods lead to gains in power when there are three predictors. When there are a large number of predictors, ANCOVA retains nominal coverage and Type I error, but there are concerns with the other approaches. For IPTW, the model-based standard errors overestimate the empirical standard error, leading to slight overcoverage. With AIPTW, G-computation and TMLE, the model-based standard errors underestimate the empirical standard errors, leading to undercoverage and high type I error. Underestimation of the standard error based on the efficient influence function for the TMLE has been reported in the literature (Petersen et al., 2014, Roy et al., 2016). The results for AIPTW and IPTW

when there are a high number of covariates should be interpreted with caution as they lead to convergence issues; see Table 1 in Additional File 1 for more details. Results for $n = 100$ are provided in Additional File 1, where we observe similar patterns, although the issues are alleviated at the larger sample size.

Extension 2: Interaction

Figures 8 and 9 display the results for *large interaction* and *absent in one group* settings for Extension 2. Results for the *small interaction* and *absent in one group* settings are shown in Additional File 1. For the *small interaction* and *large interaction* scenarios, all methods are unbiased, and standard errors are reduced as sample size increases. As before, the falsely small standard errors for IPTW, AIPTW and G-computation lead to undercoverage and false gains in power when sample size is small.

In the *Different shapes* and *Absent in one group* settings, the effect of the covariate is different in each arm. In the *Absent in one group* setting, the covariate is very skewed and is associated with the outcome only in one arm. In these settings, we see large biases with some methods, and biases are present even when sample size

is 100. In particular, in the extreme setting of *Absent in one group* setting, the we see that the biases are particularly pronounced for the ANCOVA, the spline and G-computation. It appears that G-computation with interaction, IPTW and AIPTW do not suffer from bias as much as G-computation without interaction and the regression-based methods. The undercoverage in the unadjusted approach in the *absent in one group* setting is due to unequal variances between the treatment and control groups; at sample size of 1000, the unadjusted approach achieves nominal coverage.

Extension 3: Binary outcome

Figure 10 displays results for the binary outcome case with the odds ratio as the estimand of interest when sample size is 100. For all methods except direct RA, bias in the estimate of the marginal log odds ratio is shown. For direct RA, bias in the estimated conditional log odds ratio compared to the data generating conditional log odds ratio is shown. For all relationships, we observe an apparent bias for direct RA. In the case of the *linear* relationship, this is due to small sample bias in logistic regression (Nemes et al., 2009); for all other relationships, direct RA assumes a linear effect of the covariate on the log-odds, and therefore estimates a different parameter compared to the data generating log odds ratio, which is reflected in

the bias. Coverage and Type I error appear to be reasonable for all methods except the adjusted approach. We observe that, as in the continuous outcome case, adjustment leads to gain in power when the covariate–outcome relationship can be approximated by a linear relationship. For the highly nonlinear *quadratic* and *harmonic* relationships, there are no gains in power.

Figure 11 displays results for the binary outcome case with the risk difference as the estimand of interest when sample size is 100. Convergence issues occur in the adjusted approach (in over 90% of simulations for the *linear*, *two-tier*, and *flattening* relationships, over 60% of simulation for the *quadratic* relationship, and over 3% of simulations for the *harmonic* relationship), so these results are omitted. All other adjustment methods produce unbiased estimates. Coverage and type I error appear reasonable although there is evidence of slight undercoverage for nonlinear relationships; for the *two-tier* relationship, AIPTW, G-computation and TMLE lead to slight undercoverage, and for the *flattening* relationship, G-computation and TMLE lead to slight undercoverage. The standard error is particularly underestimated for TMLE. Similarly to the odds ratio, we observe that gains in power for adjustment are strongest when the covariate–outcome relationship is approximately linear.

Discussion

Adjustment for baseline covariates in clinical trials have been shown to be beneficial where sample size is moderate to large. We investigated whether the benefits of adjustment - gain in power while estimates remain unbiased and coverage remains at the nominal level - are retained when there is potential for misspecification of the covariate–outcome relationship, and where sample size is small. We considered a wide range of adjustment methods including lesser-used methods such as IPTW, AIPTW, G-computation and TMLE, and considered whether they offer any ad-

Figure 10: Extension 3 results (binary outcome, single covariate, no interaction) for sample size 100 and odds ratio as the estimand of interest. The performance of analytic methods in terms of bias, coverage and power are shown for the five different covariate–outcome relationships. Bias calculated for the marginal log-odds for all methods except for direct regression adjustment (RA). Model-based standard errors are shown in black and the empirical standard error is shown in red. The conditional odds ratio is 0.2.

* For direct regression adjustment (RA), the bias is the difference between the estimated conditional odds ratio, and the data generating conditional odds ratio. The standard errors for direct RA cannot be compared to the other approaches.

vantages over the commonly used ANCOVA. See Table 1 for a summary of the analytic methods. We note that our simulations considered parallel design trial settings where randomization is 1:1 at the individual level, and our findings may not hold under other randomization schemes (Wang et al., 2021).

Should you adjust?

Our simulations showed that, for the continuous outcome case where sample size is at least 100 and adjustment is for a few covariates, and there are no strong covariate–treatment interactions, all methods have desirable properties.

In the continuous outcome case, where there is non-linearity in the covariate–outcome relationship, methods that allow for non-linearities lead to greater gain in power than the common ANCOVA approach; these include the ANCOVA with spline, G-computation with interaction, G-computation with spline and IPTW. The recommendation by Kahan et al. (2016) to use splines to model the covariate–

Figure 11: Extension 3 results (binary outcome, single covariate, no interaction) for sample size 100 and risk difference the estimand of interest. The performance of analytic methods in terms of bias, coverage and power are shown for the five different covariate–outcome relationships. Model-based standard errors are shown in black and the empirical standard error is shown in red.

outcome relationship is relevant in smaller trials.

When covariate–treatment interactions exist, adjustment methods can suffer from bias in small samples. The bias reduces with increasing sample size. If the interaction is non-linear and the covariate distribution is skewed, bias can present even in large samples. Adjustment methods which allow for the presence of an interaction (including IPTW and G-computation with interaction) achieve an unbiased estimate at smaller sample sizes than the ANCOVA and other methods which do not allow for an interaction. If interactions are likely to exist, it is essential to ensure that sample sizes are sufficiently large to protect against bias.

For the binary outcome case, an adjusted binomial model with a logistic link function will change the estimand. Further, small sample sizes are likely to lead to convergence issues. For the adjusted binomial model with linear link function, convergence issues are present even with $n = 100$.

Which covariates should be adjusted for?

While the focus of our simulation study was not in the selection of covariates for adjustment in the design/analysis stage of a trial, our simulations have shown that increase in power due to adjustment occurs when covariates are prognostic of outcome and the covariate–outcome relationship is linear or approximately linear, as shown previously by [Kahan et al. \(2014\)](#). If the relationship is non-linear but the adjustment approach allows for flexibility in the modeling of this relationship via

splines, gains in power can also be achieved. However, the choice of degrees of freedom for the splines should be sensible relative to the sample size.

If a large number of covariates are adjusted for, we found that the statistical properties of ANCOVA were similar when the number of covariates was a select few. However, for other adjustment methods, adjusting for a high number of covariates led to over- or under-coverage and high type I error.

Is it possible to improve on the ANCOVA?

We found that all methods performed well when sample sizes are moderate and covariate–treatment interactions are absent. We identified two settings where alternative adjustment methods to ANCOVA provide improvement. Firstly, where there are covariate–treatment interactions, G-computation with interaction, IPTW and AIPTW are potentially promising approaches. Secondly, our simulations showed that adding non-linearities by splines with a suitable number of degrees of freedom help to gain power when the true covariate–outcome relationship is non-linear. If the true relationship approximately linear, the addition of spline terms generally do not lead to loss of power.

In addition, for binary outcomes with the odds ratio as the estimand of interest, an advantage of using AIPTW, IPTW, G-computation and TMLE is that these adjustment approaches retain the marginal estimand, whereas regression-based approaches such as the ANCOVA and spline change the estimand. However, for smaller sample sizes, G-computation, AIPTW and IPTW lead to underestimation of the standard error. It is essential that small sample corrections are developed for these approaches. Underestimation of standard errors can also occur for TMLE if efficient influence function based variance estimators are used; bootstrapping and stratified TMLE have been recommended as alternative approaches, which were not explored in this study ([Petersen et al., 2014](#), [Roy et al., 2016](#)).

Practitioners can be reassured that covariate adjustment in settings commonly encountered in clinical trials generally leads to gains in power while estimates remain unbiased and coverage is at nominal level. The choice of method, ideally made in the planning stages of the trial, should take into account whether covariate–treatment interaction is likely and whether the sample size is sufficient for the use of methods that rely on large-sample properties.

Table 1: Summary of analytic methods and their properties

Method	Properties
Unadjusted	Unbiased in all settings. Typically reduced power compared to adjusted approaches.
ANCOVA/ Adjusted	Typically leads to increases in power. Retains good properties if many covariates adjusted for. No issues with estimation of standard errors in small samples. Bias in non-linear interaction setting. Marginal odds ratio cannot be targeted. Convergence issues if risk ratio is of interest.
G-computation	Undercoverage and high type I error in small sample sizes. Bias in non-linear interaction setting; alleviated by allowing for interaction.
IPTW	covariate–outcome relationship need not be specified. Undercoverage and high type I error in small sample sizes and adjusting for a few covariates. Overcoverage if adjusting for many covariates. Convergence issues if there are many covariates. Slight bias in non-linear interaction setting.
AIPTW	Either covariate–treatment or covariate–outcome relationship needs to be correct. Undercoverage and high type I error in small sample sizes. Convergence issues if there are many covariates. Slight bias in non-linear interaction setting.
TMLE	Either covariate–treatment or covariate–outcome relationship needs to be correct. Standard errors can be underestimated if efficient influence function based estimators are used. Slight bias in non-linear interaction setting.

Appendix

We demonstrate that bias can be incurred for the ANCOVA when there is a covariate–treatment interaction, the covariate has a skewed distribution, and the covariate–outcome relationship is misspecified.

For the simple ANCOVA, as in Equation (5), the least squares/maximum likelihood parameter estimates of the effect of treatment β_x and effect of covariate γ are given by:

$$\hat{\beta}_x = (\bar{y}_1 - \bar{y}_0) - \hat{\gamma}(\bar{x}_1 - \bar{x}_0),$$

$$\hat{\gamma} = \frac{\sum z_i(y_i - \bar{y}_1)(x_i - \bar{x}_1) + \sum (1 - z_i)(y_i - \bar{y}_0)(x_i - \bar{x}_0)}{\sum z_i(z_i - \bar{x}_1)^2 + \sum (1 - z_i)(x_i - \bar{x}_0)^2}.$$

We first show that $\hat{\gamma}$ can be expressed as an inverse-weighted average of the effect of the covariate for each arm. To do this, we fit a regression model for the effect of

the covariate in each arm:

$$\begin{aligned} Y &= \alpha^{(0)} + \gamma^{(0)}X + \epsilon^{(0)} & \text{if } Z &= 0 \\ Y &= \alpha^{(1)} + \gamma^{(1)}X + \epsilon^{(1)} & \text{if } Z &= 1, \end{aligned}$$

where $\epsilon^k \sim N(0, \sigma_{(k)}^2)$.

The least squares/maximum likelihood estimates for the effect of the covariate in each arm, $\gamma^{(1)}$ and $\gamma^{(0)}$ are given by:

$$\begin{aligned} \hat{\gamma}^{(1)} &= \frac{\sum z_i(y_i - \bar{y}_1)(x_i - \bar{x}_1)}{\sum z_i(x_i - \bar{x}_1)^2} \\ \hat{\gamma}^{(0)} &= + \frac{\sum (1 - z_i)(y_i - \bar{y}_0)(x_i - \bar{x}_0)}{\sum (1 - z_i)(x_i - \bar{x}_0)^2} \end{aligned}$$

Further, the estimates of their variances are given by:

$$\begin{aligned} Var(\hat{\gamma}^{(1)}) &= \frac{\sigma_{(k)}^2}{\sum z_i(x_i - \bar{x}_1)^2}, \\ Var(\hat{\gamma}^{(0)}) &= \frac{\sigma_{(k)}^2}{\sum (1 - z_i)(x_i - \bar{x}_0)^2}. \end{aligned}$$

We assume that $\sigma_{(1)}^2 = \sigma_{(0)}^2$ and define inverse variance weights w_0 and w_1 :

$$\begin{aligned} w_1 &= (\sigma^2)^{(-1)} \sum z_i(x_i - \bar{x}_1)^2, \\ w_0 &= (\sigma^2)^{(-1)} \sum (1 - z_i)(x_i - \bar{x}_0)^2 \end{aligned}$$

We can then write the the ANCOVA estimate of the treatment effect as an inverse-weighted average of the slopes within each arm:

$$\hat{\gamma} = \frac{w_1 \hat{\gamma}^{(1)} + w_0 \hat{\gamma}^{(0)}}{w_1 + w_0}.$$

We can then write the effect of treatment as follows:

$$\begin{aligned} \hat{\beta}_x &= (\bar{y}_1 - \bar{y}_0) - \hat{\gamma}(\bar{x}_1 - \bar{x}_0), \\ &= (\bar{y}_1 - \bar{y}_0) - \frac{w_1 \hat{\gamma}^{(1)} + w_0 \hat{\gamma}^{(0)}}{w_1 + w_0}(\bar{x}_1 - \bar{x}_0). \end{aligned}$$

The weights w_1 and w_0 tend in probability to 0.5. To simplify our argument below, we replace the weights with these limits. Thus we have that the expectation of the marginal estimator $\hat{\beta}$ is:

$$\begin{aligned}
\mathbb{E}[\hat{\beta}] &= \mathbb{E}[\bar{y}_1] - \mathbb{E}[\bar{y}_0] - \mathbb{E}\left[\frac{w_1\hat{\gamma}^{(1)} + w_0\hat{\gamma}^{(0)}}{w_1 + w_0}(\bar{x}_1 - \bar{x}_0)\right] \\
&= \beta - \mathbb{E}\left[\left(0.5\hat{\gamma}^{(1)} + 0.5\hat{\gamma}^{(0)}\right)(\bar{x}_1 - \bar{x}_0)\right] \\
&= \beta - 0.5\mathbb{E}\left[\hat{\gamma}^{(1)}\bar{x}_1\right] + 0.5\mathbb{E}\left[\hat{\gamma}^{(1)}\bar{x}_0\right] - 0.5\mathbb{E}\left[\hat{\gamma}^{(0)}\bar{x}_1\right] \\
&\quad + 0.5\mathbb{E}\left[\hat{\gamma}^{(0)}\bar{x}_0\right] \\
&= \beta - 0.5\mathbb{E}\left[\hat{\gamma}^{(1)}\bar{x}_1\right] + 0.5\mathbb{E}\left[\hat{\gamma}^{(1)}\right]\mathbb{E}[\bar{x}_0] - 0.5\mathbb{E}\left[\hat{\gamma}^{(0)}\right]\mathbb{E}[\bar{x}_1] \\
&\quad + 0.5\mathbb{E}\left[\hat{\gamma}^{(0)}\bar{x}_0\right] \tag{12}
\end{aligned}$$

$$\begin{aligned}
&= \beta - 0.5\mathbb{E}\left[\hat{\gamma}^{(1)}\bar{x}_1\right] + 0.5\gamma^{(1)}\mathbb{E}[\bar{x}_0] - 0.5\gamma^{(0)}\mathbb{E}[\bar{x}_1] \\
&\quad + 0.5\mathbb{E}\left[\hat{\gamma}^{(0)}\bar{x}_0\right] \tag{13}
\end{aligned}$$

$$\begin{aligned}
&= \beta - 0.5\mathbb{E}\left[\hat{\gamma}^{(1)}\bar{x}_1\right] + 0.5\gamma^{(1)}\mathbb{E}[\bar{x}_1] - 0.5\gamma^{(0)}\mathbb{E}[\bar{x}_1] \\
&\quad + 0.5\mathbb{E}\left[\hat{\gamma}^{(0)}\bar{x}_0\right] \tag{14}
\end{aligned}$$

$$= \beta - 0.5\left[\text{Cov}(\hat{\gamma}^{(1)}\bar{x}_1) - \text{Cov}(\hat{\gamma}^{(0)}, \bar{x}_0)\right] \tag{15}$$

where Equation (12) follows since we have that $\hat{\gamma}^{(1)}$ is independent of $\bar{x}_0$, and $\hat{\gamma}^{(0)}$ is independent of $\bar{x}_1$. Equation (13) follows since $\mathbb{E}[\hat{\gamma}^{(1)}] = \gamma^{(1)}$, and $\mathbb{E}[\hat{\gamma}^{(0)}] = \gamma^{(0)}$. Equation (14) follows since for a randomized experiment, $\mathbb{E}[\bar{x}_1] = \mathbb{E}[\bar{x}_0]$.

The bias due to adjustment by ANCOVA in a trial setting with 1:1 randomisation can therefore be written as $0.5[\text{Cov}(\hat{\gamma}^{(1)}, \bar{x}_1) - \text{Cov}(\hat{\gamma}^{(0)}, \bar{x}_0)]$. This bias is therefore approximately zero in expectation for sufficiently large sample sizes. For finite samples, if the effect of the treatment is different in each arm, the covariate has skewness, and further, the ANCOVA leads to misspecification of the covariate–outcome relationship, this bias may be particularly pronounced.

Acknowledgements

Not applicable

Funding

EW is supported by the MRC Network of Hubs for trials methodology research (MR/ L004933/2-N96). MT is supported by Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome. CL is supported by the UK Medical Research Council (Skills Development Fellowship MR/T032448/1). IW and TM are supported by the Medical Research Council Programme MC_UU_00004/07. KDO is funded by a Royal Society-Wellcome Trust Sir Henry Dale fellowship 218554/Z/19/Z.

Abbreviations

AIPTW: Augmented Inverse Probability-of-Treatment Weighting
 ANCOVA: Analysis of Covariance
 Direct RA: Direct Regression Adjustment
 IPTW: Inverse Probability-of-Treatment Weighting
 OR: Odds ratio TMLE: Targeted Maximum Likelihood Estimation

Availability of data and materials

All data generated or analysed during this study are included in this published article and its supplementary information files.

Ethics approval and consent to participate

Not applicable

Competing interests

TM consults for Kite Pharma, Inc. IW has provided consultancy services or courses to Exelixis, AstraZeneca, GSK and Novartis, for which his employer has received funding.

Consent for publication

Not applicable.

Authors' contributions

EW conceived the research idea, led the investigation and performed analyses. MT led the writing of the paper and performed analyses. TM, IW, CL and KDO provided methodological contributions. All authors read and approved the final manuscript.

Authors' information

Not applicable.

Author details

¹Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK. ²MRC Clinical Trials Unit at UCL, London, UK.

References

- Bartlett, J.W.: Covariate adjustment and estimation of mean response in randomised trials. *Pharmaceutical statistics* **17**(5), 648–666 (2018). doi:[10.1002/pst.1880](https://doi.org/10.1002/pst.1880). [1707.04465](https://doi.org/10.1002/pst.1880)
- Daniel, R., Zhang, J., Farewell, D.: Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biometrical Journal* **63**(3), 528–557 (2021). doi:[10.1002/bimj.201900297](https://doi.org/10.1002/bimj.201900297). <https://onlinelibrary.wiley.com/doi/pdf/10.1002/bimj.201900297>
- Davidian, M., Tsiatis, A.A., Leon, S., An, H., Little, R., Brumback, B.A., Brumback, L.C., Molenberghs, G., Schafer, J.L., Kang, J.D.Y.: Semiparametric estimation of treatment effect in a pretest-posttest study with missing data. *Statistical Science* **20**(3), 261–301 (2005). doi:[10.1214/088342305000000151](https://doi.org/10.1214/088342305000000151)
- European Medicines Agency: Guideline on adjustment for baseline covariates in clinical trials (2015). http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2015/03/WC500184923.pdf
- Fay, M.P., Graubard, B.I.: Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics* **57**(4), 1198–1206 (2001). doi:[10.1111/j.0006-341X.2001.01198.x](https://doi.org/10.1111/j.0006-341X.2001.01198.x)
- Food and Drug Administration: Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products Guidance for Industry DRAFT GUIDANCE (2021). <https://www.fda.gov/media/148910/download> Accessed 11/06/2021
- Gruber, S., van der Laan, M.J.: tmle: An R package for targeted maximum likelihood estimation. *Journal of Statistical Software* **51**(13), 1–35 (2012). doi:[10.18637/jss.v051.i13](https://doi.org/10.18637/jss.v051.i13)
- Hauck, W.W., Anderson, S., Marcus, S.M.: Should we adjust for covariates in nonlinear regression analyses of randomized trials? *Controlled Clinical Trials* **19**(3), 249–256 (1998). doi:[10.1016/S0197-2456\(97\)00147-5](https://doi.org/10.1016/S0197-2456(97)00147-5)
- Kahan, B.C., Morris, T.P.: Reporting and analysis of trials using stratified randomisation in leading medical journals: review and reanalysis. *BMJ* **345** (2012). doi:[10.1136/bmj.e5840](https://doi.org/10.1136/bmj.e5840). <https://www.bmj.com/content/345/bmj.e5840.full.pdf>
- Kahan, B.C., Jairath, V., Doré, C.J., Morris, T.P.: The risks and rewards of covariate adjustment in randomized trials: An assessment of 12 outcomes from 8 studies. *Trials* **15**(1), 1–7 (2014). doi:[10.1186/1745-6215-15-139](https://doi.org/10.1186/1745-6215-15-139)
- Kahan, B.C., Rushton, H., Morris, T.P., Daniel, R.M.: A comparison of methods to adjust for continuous covariates in the analysis of randomised trials. *BMC Medical Research Methodology* **16**(1), 1–10 (2016). doi:[10.1186/s12874-016-0141-3](https://doi.org/10.1186/s12874-016-0141-3)
- Luque-Fernandez, M.A., Schomaker, M., Rachet, B., Schnitzer, M.E.: Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Statistics in Medicine* **37**(16), 2530–2546 (2018). doi:[10.1002/sim.7628](https://doi.org/10.1002/sim.7628). <https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.7628>
- Mao, H., Li, L.: PSW: Propensity Score Weighting Methods for Dichotomous Treatments. (2018). R package version 1.1-3. <https://CRAN.R-project.org/package=PSW>
- Nemes, S., Jonasson, J.M., Genell, A., Steineck, G.: Bias in odds ratios by logistic regression modelling and sample size. *BMC Medical Research Methodology* **9**(1), 1–5 (2009). doi:[10.1186/1471-2288-9-56](https://doi.org/10.1186/1471-2288-9-56)
- Petersen, M., Schwab, J., Gruber, S., Blaser, N., Schomaker, M., van der Laan, M.: Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. *Journal of Causal Inference* **2**(2), 147–185 (2014). doi:[10.1515/jci-2013-0007](https://doi.org/10.1515/jci-2013-0007)
- R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2021). R Foundation for Statistical Computing. <https://www.R-project.org/>
- Raad, H., Cornelius, V., Chan, S., Williamson, E., Cro, S.: An evaluation of inverse probability weighting using the propensity score for baseline covariate adjustment in smaller population randomised controlled trials with a continuous outcome. *BMC Medical Research Methodology* **20**(1), 1–12 (2020). doi:[10.1186/s12874-020-00947-7](https://doi.org/10.1186/s12874-020-00947-7)
- Roy, J., Lum, K.J., Daniels, M.J.: A bayesian nonparametric approach to marginal structural models for point treatments and a continuous or survival outcome. *Biostatistics* **18**(1), 32–47 (2016). doi:[10.1093/biostatistics/kxw029](https://doi.org/10.1093/biostatistics/kxw029). <https://academic.oup.com/biostatistics/article-pdf/18/1/32/9607725/kxw029.pdf>
- Schuler, M.S., Rose, S.: Targeted maximum likelihood estimation for causal inference in observational studies. *American Journal of Epidemiology* **185**(1), 65–73 (2017). doi:[10.1093/aje/kww165](https://doi.org/10.1093/aje/kww165)
- Senn, S.: Seven myths of randomisation in clinical trials. *Statistics in Medicine* **32**(9), 1439–1450 (2013). doi:[10.1002/sim.5713](https://doi.org/10.1002/sim.5713)

- Sjolander, A., Dahlqvist, E.: stdReg: Regression Standardization. (2020). R package version 3.4.0. <https://CRAN.R-project.org/package=stdReg>
- Stefanski, L.A., Boos, D.D.: The calculus of M-estimation. *American Statistician* **56**(1), 29–38 (2002). doi:[10.1198/000313002753631330](https://doi.org/10.1198/000313002753631330)
- Tsiatis, A., Davidian, M., Zhang, M., Lu, X.: Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine* **27**(23), 4658–4677 (2008)
- van der Laan, M.J., Rubin, D.: Targeted maximum likelihood learning. *International Journal of Biostatistics* **2**(1) (2006). doi:[10.2202/1557-4679.1043](https://doi.org/10.2202/1557-4679.1043)
- van der Laan, M.J., Rose, S.: Targeted Learning: Causal Inference for Observational and Experimental Data. Springer Series in Statistics. Springer, ??? (2011). <https://books.google.co.uk/books?id=RGnSX5aCagQC>
- Wang, B., Ogburn, E.L., Rosenblum, M.: Analysis of covariance in randomized trials: More precision and valid confidence intervals, without model assumptions. *Biometrics* **75**(4), 1391–1400 (2019). doi:[10.1111/biom.13062](https://doi.org/10.1111/biom.13062)
- Wang, B., Susukida, R., Mojtabai, R., Amin-Esmaili, M., Rosenblum, M.: Model-robust inference for clinical trials that improve precision by stratified randomization and covariate adjustment. *Journal of the American Statistical Association* **0**(ja), 1–31 (2021). doi:[10.1080/01621459.2021.1981338](https://doi.org/10.1080/01621459.2021.1981338). <https://doi.org/10.1080/01621459.2021.1981338>
- White, I.R., Morris, T.P., Williamson, E.: Covariate adjustment in randomised trials: canonical link functions protect against model mis-specification, 1–10 (2021). [2107.07278](https://doi.org/10.1002/sim.9991)
- Williamson, E.J., Forbes, A., White, I.R.: Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in Medicine* **33**(5), 721–737 (2014). doi:[10.1002/sim.5991](https://doi.org/10.1002/sim.5991)
- Williamson, T., Eliasziw, M., Fick, G.H.: Log-binomial models: Exploring failed convergence. *Emerging Themes in Epidemiology* **10**(1), 1–10 (2013). doi:[10.1186/1742-7622-10-14](https://doi.org/10.1186/1742-7622-10-14)
- Yang, L., Tsiatis, A.A.: Efficiency study of estimators for a treatment effect in a pretest-posttest trial. *American Statistician* **55**(4), 314–321 (2001). doi:[10.1198/000313001753272466](https://doi.org/10.1198/000313001753272466)
- Zhou, T., Tong, G., Li, F., Thomas, L., Li, F.: PSweight: Propensity Score Weighting for Causal Inference with Observational Studies and Randomized Trials. (2021). R package version 1.1.4. <https://CRAN.R-project.org/package=PSweight>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [AdditionalFile1.pdf](#)