BMC Medical Research Methodology

RESEARCH Open Access

Statistical analysis of two arm randomized pre-post designs with one post-treatment measurement

Fei Wan

Abstract

Background: Randomized pre-post designs, with outcomes measured at baseline and after treatment, have been commonly used to compare the clinical effectiveness of two competing treatments. There are vast, but often conflicting, amount of information in current literature about the best analytic methods for pre-post designs. It is challenging for applied researchers to make an informed choice.

Methods: We discuss six methods commonly used in literature: one way analysis of variance ("*ANOVA*"), analysis of covariance main effect and interaction models on the post-treatment score ("*ANCOVA*I" and "*ANCOVA*II"), *ANOVA* on the change score between the baseline and post-treatment scores ("*ANOVA-Change*"), repeated measures ("*RM*") and constrained repeated measures ("*cRM*") models on the baseline and post-treatment scores as joint outcomes. We review a number of study endpoints in randomized pre-post designs and identify the mean difference in the post-treatment score as the common treatment effect that all six methods target. We delineate the underlying differences and connections between these competing methods in homogeneous and heterogeneous study populations.

Results: *ANCOVA* and *cRM* outperform other alternative methods because their treatment effect estimators have the smallest variances. *cRM* has comparable performance to *ANCOVA*I in the homogeneous scenario and to *ANCOVA*II in the heterogeneous scenario. In spite of that, *ANCOVA* has several advantages over cRM: i) the baseline score is adjusted as covariate because it is not an outcome by definition; ii) it is very convenient to incorporate other baseline variables and easy to handle complex heteroscedasticity patterns in a linear regression framework.

Conclusions: *ANCOVA* is a simple and the most efficient approach for analyzing pre-post randomized designs.

Keywords: Pre-post design, ANCOVA, ANOVA, Repeated measures, Change score, Treatment effect

Background

Two arm parallel randomized trials have been widely used to compare the clinical effectiveness of competing treatments in improving patients' health outcomes. In these trials, continuous outcomes of interest were routinely measured at baseline (defined as "baseline score") and one post treatment time point (defined as "post-treatment score"). The primary purpose of designing a pre-post randomized study is to answer the scientific question of interest: is treatment A more effective than treatment B? To assess the difference in the treatment effectiveness between two treatments, we need to select a study endpoint and quantify a treatment effect. Common study endpoints include the post treatment score, the change score from baseline to post treatment, a percentage

Correspondence: wan.fei@wustl.edu
Department of Surgery, Division of Public Health Sciences, Washington University School of Medicine, Campus Box 8100, 660 S. Euclid Ave, St. Louis, MO, USA

change from baseline, and rate of change from baseline. The difference between two arms on selected study endpoints is defined as the treatment effect. Few studies have investigated the links between these different metrics of treatment effect in a randomized pre-post trial. These underlying connections are critical in understanding the equivalence among some statistical methods that may appear to be very different at the first sight. We need to be certain about the type of treatment effect each method targets and select the one that yields an unbiased and the most efficient estimator of the treatment effect of our interest.

There are a number of statistical methods commonly used in analyzing pre-post trials. We can analyze the post-treatment score using one way analysis of variance model (***ANOVA***) [1, 2], analysis of covariance model adjusting for the baseline score (***ANCOVA*I**) [2–7], and ANCOVA including a baseline score by treatment interaction (***ANCOVA*II**) [3, 4, 8–10]. We can also analyze the change score using ***ANOVA*** (***ANOVA-Change***) [11]. Alternatively, we can model the baseline and post-treatment scores jointly using repeated measures models (***RM***) and constrained repeated measures models (***cRM***) [10, 12–14]. Despite of the simplicity and wide application of randomized pre-post designs, which method is the best analytic approach has been a debated topic and many methodological studies have been performed to compare different statistical methods for past decades [1–13]. However, it is challenging for applied researchers to evaluate this vast, but often conflicting, amount of information in current literature and make an informed choice.

In this study we aim to review ***ANOVA, ANCOVA*I, *ANCOVA*II, *ANOVA-Change, RM,*** **and** ***cRM*** from a practical standpoint, with the focus on delineating the differences and underlying connections between them. In section Methods, we first provide notations and assumptions for a typical pre-post design, define homogeneous and heterogeneous study populations, and discuss some common study endpoints and the associated metrics of treatment effects. We next analytically assess differences and connections between these competing models in the homogeneous and heterogeneous scenarios by first describing each model using the same set of population mean, variance, and covariance parameters. In section Results, we compare the relative efficiency of these competing methods theoretically using three simulated weight loss trial examples (homogeneous data, heterogeneous data with balanced design, heterogeneous data with unbalanced design). In the last two sections, we discuss the results and give recommendation on the best analytical approach in a randomized pre-post design.

Methods

A hypothetical weight loss trial and metrics of treatment effects

Notations

In a hypothetical two arm parallel weight loss trial comparing the effect of a new drug ("treatment") and a placebo ("control") in reducing participants' body weights, we use Y_{ijt} to denote body weight of the i th subject ($i = 1, 2, 3, \ldots n_j$) in the jth treatment arm ($j = 0, 1$) at the t th time ($t = t_0, t_1$). n_0 and n_1 are the number of subjects in the control and treatment arms.

We denote the mean baseline weights for the treatment and control arms by μ_{1t_0} and μ_{0t_0}, respectively. Random allocation guarantees $\mu_{1t_0} = \mu_{0t_0}$ and we let μ_{t_0} denote the overall mean baseline weight. The mean weights of the treatment and control arms at time t_1 are denoted by μ_{1t_1} and μ_{0t_1}, respectively (Fig. 1). We define homogeneous and heterogeneous study populations as follows:

i) ***The homogeneous scenario:*** every participant has the same pattern of variance and covariance structure for their baseline and post-treatment weights, which is parameterized as below:

$$\sum = \begin{bmatrix} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{bmatrix},$$

where σ_0^2 and σ_1^2 are the variances of the baseline and post-treatment weights, ρ is the correlation coefficient between the baseline and post-treatment weights.

ii) ***The heterogeneous scenario:*** variance and covariance structures of the baseline and post-treatment weights differ between the treatment and control arms. Formally, we have

$$\sum\nolimits_0 = \begin{bmatrix} \sigma_0^2 & \rho_0\sigma_0\sigma_{01} \\ \rho_0\sigma_0\sigma_{01} & \sigma_{01}^2 \end{bmatrix},$$

and

$$\sum\nolimits_1 = \begin{bmatrix} \sigma_0^2 & \rho_1\sigma_0\sigma_{11} \\ \rho_1\sigma_0\sigma_{11} & \sigma_{11}^2 \end{bmatrix},$$

where σ_0^2 is the common variance of the baseline body weight in the control and treatment arms. Randomization guarantees that the variances of the baseline weights in both arms are equal to σ_0^2. σ_{01}^2 and σ_{11}^2 are the variances of the post-treatment weight in the control and treatment arms. ρ_0 and ρ_1 are the correlation coefficients between the baseline and post-treatment weights in the control and treatment arms,

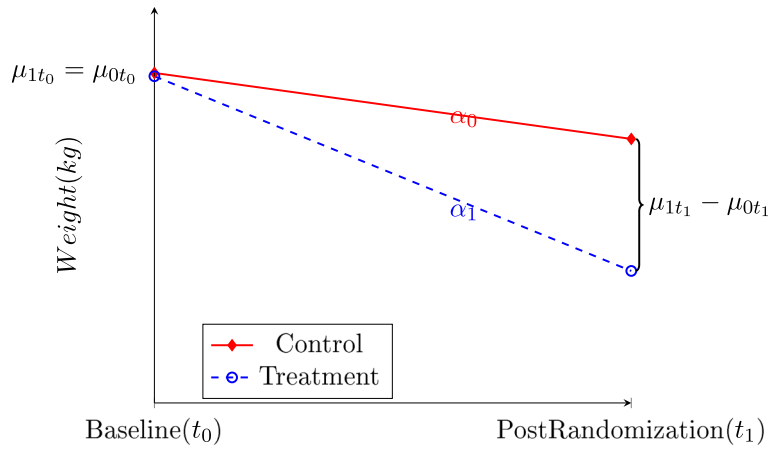

Fig. 1 Hypothetical two arm pre-post weight loss randomized trial

respectively. In practice, participants may respond to the treatment more differently so that variability of the post-treatment weight tends to be larger in the treatment arm than in the control arm and the correlation between pre- and post-treatment weights are usually stronger in the control arm than in the treatment arm. i.e., $\rho_0 > \rho_1$ and $\sigma_{11}^2 > \sigma_{01}^2$.

Metrics of treatment effect

We discuss the following three metrics of treatment effect commonly reported in pre-post trials:

i) The primary endpoint is the post-treatment weight measured at t_1. The difference in the mean post-treatment weights of two arms is defined as a treatment effect, which is parameterized as follows:

$$\tau = \mu_{1t_1} - \mu_{0t_1}$$

For example, if $\tau = -10$, we can interpret the results as "at the end of the trial, the mean weight was 10 pounds lower in the treatment group than in the control group."

ii) The primary endpoint is the change score calculated by subtracting the baseline weight from the post-treatment weight. i.e., $\Delta_{ij} = Y_{ijt_1} - Y_{ijt_0}$. The difference in the mean change scores of two arms is a treatment effect. Formally, we have:

$$\tilde{\tau} = \left(\mu_{1t_1} - \mu_{1t_0}\right) - \left(\mu_{0t_1} - \mu_{0t_0}\right)$$

e.g. if $\tilde{\tau} = -10$, this difference is usually interpreted as "weight **reductions** were 10 pounds greater in the treatment group than in the control group". Since randomization ensures $\mu_{0t_1} = \mu_{0t_0}$, it follows directly $\tilde{\tau} = \tau$. When we code "0" for t_0 and "1" for t_1, the mean change score for each arm can also be interpreted as the mean change rate per unit time for each arm, represented by slopes in Fig. 1. Thus, the difference in slopes, denoted by $\widetilde{\tilde{\tau}} = \alpha_1 - \alpha_0$, is also equivalent to τ. As shown in previous section, ***ANOVA*** and ***ANCOVA*** target τ, ***ANOVA-CHANGE*** targets $\tilde{\tau}$, and ***RM*** targets $\widetilde{\tilde{\tau}}$. However, we can compare these statistical methods targeting seemingly very different types of treatment effects in a meaningful way because of the equivalence between τ, $\tilde{\tau}$, and $\widetilde{\tilde{\tau}}$ in randomized pre-post designs.

iii) The primary endpoint is the percent change from baseline weight, denoted by $\phi_{ij} = \frac{(Y_{ijt_1} - Y_{ijt_0})}{Y_{ijt_0}}$. The mean difference in the percent change between two arms is defined as a treatment effect and parameterized as follows:

$$\tau^* = \overline{\phi}_1 - \overline{\phi}_0,$$

where $\overline{\phi}_1$ and $\overline{\phi}_0$ are the mean percent changes of the treatment and control arms. Although the percent change is popular among clinical researchers, this metric has several drawbacks [1, 15, 16]: i) the percent change is a function of ratio $\frac{Y_{ijt_1}}{Y_{ijt_0}}$. The distribution of the percent change is highly skewed. Analyzing it with normal-theory based statistical methods is not justified and non-parametric statistical methods are generally less powerful; ii) the percent change is not a symmetric measure. For example, the mean weight of adults over 20 in US is 197.8 pound for men and 170.5 pound for women. The mean difference is 27.3 pound between men and women. Men weight 16% (i.e.,100 × ((197.8–170.5)/170.5)) more

than women, whereas women weight 13.8% (i.e., 100 × ((197.8–170.5)/197.8)) less than men. The differences could be different depending on which sex is used as devisor; iii) the percent change is not an additive measure. For example, if a participant's weight increases by 10% in first 6 months and fall by 10% for the next 6 months, the 2 % changes do not cancel out. The participant's weight at the end would be only 99% of the participant's starting weight.

Statistical models

In this section, we focus on six methods that estimate τ. We describe each statistical model using the same set of population mean, variance, and covariance parameters defined in section Methods for homogeneous and heterogeneous scenarios, separately. For each method, we present the closed-form expressions of the point estimator of treatment effect and its variance. It often goes unnoticed in practice that different statistical methods have different types of variances (i.e., conditional vs. unconditional variances) associated with their treatment effect estimators. For example, the OLS model-based variances for ANCOVA are conditional because OLS assumes the baseline weight is fixed. Generally speaking, the baseline weight is random because we rarely enroll participants into randomized trials based on predetermined values of the baseline weight. Thus, the unconditional variance and the corresponding unconditional inference is of greater interest because we want the findings derived from the current sample to be generalizable to the population of interest. We will discuss in details whether the OLS model-based conditional inference (i.e., test statistics and *p*-values from standard statistical softwares) for ANCOVA is still valid for unconditional hypothesis testing and the potential fixes that we can use to draw valid unconditional inference if the usual OLS model-based inference is biased.

When the study population is homogeneous

Method 1:*ANOVA modeling post treatment measure ("ANOVA-Post").* We model the post-treatment body weight Y_{ijt_1} using the binary treatment indicator G_{ij} (1 if in the treatment arm; 0 if in the control arm) as follows:

$$Y_{ijt_1} = \beta_0^{(1)} + \beta_1^{(1)} G_{ij} + e_{ij}^{(1)}, i = 1, 2, ..., n_j; j = 0, 1; \quad (1)$$

$$e_{ij}^{(1)} \sim N\left(0, \sigma_1^2\right),$$

where $\beta_0^{(1)} = \mu_{0t_1}, \beta_1^{(1)} = \mu_{1t_1} - \mu_{0t_1} = \tau$, and $e_{ij}^{(1)}$ is independently and identically distributed (i.i.d) random error. $\beta_1^{(1)}$ represents the treatment effect. Model (1) is homoscedastic with a constant residual variance σ_1^2.

We can fit an ordinary least squares (OLS) regression to estimate the coefficients and standard errors of model (1). The closed-form expressions of the OLS estimator $\hat{\beta}_{1,ols}^{(1)}$ and its "unconditional" variance, denoted by $var(\hat{\beta}_{1,ols}^{(1)})$, are presented in Table 1. $\hat{\beta}_{1,ols}^{(1)}$ is estimated by the sample group mean difference in the post-treatment weight between two arms. $\hat{\beta}_{1,ols}^{(1)}$ is unbiased for τ. The OLS model-based variance of $\hat{\beta}_{1,ols}^{(1)}$ assuming known σ_1^2 is:

$$var_{ols}\left(\hat{\beta}_{1,ols}^{(1)}\right) = \frac{\sigma_1^2}{\sum_{j=0}^{1}\sum_{i=1}^{n_j}\left(G_{ij} - G_{..}\right)^2},$$

where $G_{..} = \frac{\sum_{j=0}^{1}\sum_{i=1}^{n_j} G_{ij}}{n_0+n_1} = \frac{n_1}{n_0+n_1}$. σ_1^2 is estimated by

$$\hat{\sigma}_1^2 = \frac{\sum_{j=0}^{1}\sum_{i=1}^{n_j}\left(y_{ijt_1} - \hat{y}_{ijt_1}^{(1)}\right)^2}{(n_0 + n_1 - 2)},$$

where $\hat{y}_{ijt_1}^{(1)} = \hat{\beta}_{0,ols}^{(1)} + \hat{\beta}_{1,ols}^{(1)} G_{ij}$ is the predicted value from model (1). We let $\widehat{var}_{ols}(\hat{\beta}_{1,ols}^{(1)})$ denote the OLS model-based variance estimator with $\hat{\sigma}_1^2$ substituted for σ_1^2, which is output by standard statistical softwares (Table 1). Since $\sum_{j=0}^{1}\sum_{i=1}^{n_j}(G_{ij} - G_{..})^2 = \frac{n_0 n_1}{n_0+n_1}$, it follows that $var_{ols}(\hat{\beta}_{1,ols}^{(1)}) = var(\hat{\beta}_{1,ols}^{(1)})$. It is well established that $\widehat{var}_{ols}(\hat{\beta}_{1,ols}^{(1)})$ is unbiased for $var_{ols}(\hat{\beta}_{1,ols}^{(1)})$. Thus, $\widehat{var}_{ols}(\hat{\beta}_{1,ols}^{(1)})$ is unbiased for $var(\hat{\beta}_{1,ols}^{(1)})$. The usual OLS model-based inference (i.e., test statistics $t = \frac{\hat{\beta}_{1,ols}^{(1)}}{\sqrt{\widehat{var}_{ols}(\hat{\beta}_{1,ols}^{(1)})}}$ and the associated *p*-value) is valid for testing $H_o: \tau = 0$ unconditionally.

Method 2:*ANCOVA modeling post treatment measure ("ANCOVA1"):* We model the post-treatment weight Y_{ijt_1} using the binary treatment indicator G_{ij} and the baseline weight Y_{ijt_0}.

$$Y_{ijt_1} = \beta_0^{(2)} + \beta_1^{(2)} G_{ij} + \beta_2^{(2)} Y_{ijt_0} + e_{ij}^{(2)}, i = 1, 2, ..., n_j; j = 0, 1; \quad (2)$$

$$e_{ij}^{(2)} \sim N\left(0, \sigma_{e^{(2)}}^2\right) \text{ and } \sigma_{e^{(2)}}^2 = \left(1 - \rho^2\right)\sigma_1^2,$$

, where $\beta_0^{(2)} = \mu_{0t_1} - \rho\frac{\sigma_1}{\sigma_0}\mu_{t_0}$, $\beta_1^{(2)} = \tau$, $\beta_2^{(2)} = \rho\frac{\sigma_1}{\sigma_0}$, and $e_{ij}^{(2)}$ is i.i.d random error. $\beta_1^{(2)}$ measures the treatment effect τ and $\beta_2^{(2)}$ represents the slope of the pre-post association between Y_{ijt_1} and Y_{ijt_0}. Model (2) has a common residual variance $\sigma_{e^{(2)}}^2$ and implicitly assumes that two arms share the common baseline mean μ_{t_0}.

Table 1 Estimators of treatment effect and variance estimators in a homogeneous study population

Model	Estimator of treatment effect (τ)	Type[a]	True variance of treatment effect estimator	OLS model based variance estimator
ANOVA-Post	$\hat{\beta}^{(1)}_{1,ols} = \overline{y}_{.1t_1} - \overline{y}_{.0t_1}$	U	$var(\hat{\beta}^{(1)}_{1,ols}) = \frac{\sigma_1^2}{n_0} + \frac{\sigma_1^2}{n_1}$	$\widehat{var}_{ols}(\hat{\beta}^{(1)}_{1,ols}) = \frac{\hat{\sigma}_1^2}{\sum_{j=0}^{1}\sum_{i=1}^{n_j}(G_{ij}-G_{\cdot,\cdot})^2}$ $\hat{\sigma}_1^2 = \frac{\sum_{j=0}^{1}\sum_{i=1}^{n_j}(y_{ijt_1}-\hat{y}_{ijt_1})^2}{(n_0+n_1-2)}$
ANCOVA-Post I	$\hat{\beta}^{(2)}_{1,ols} = (\overline{y}_{.1t_1} - \overline{y}_{.0t_1}) - \hat{\beta}^{(2)}_{2,ols}(\overline{y}_{.1t_0} - \overline{y}_{.0t_0})$	C	$var(\hat{\beta}^{(2)}_{1,ols}\|Y_{ijt_0}) = (\frac{1}{n_0} + \frac{1}{n_1} + \frac{(\overline{y}_{.1t_0} - \overline{y}_{.0t_0})^2}{\sum_{j=0}^{1}\sum_{i=1}^{n_j}(y_{ijt_0} - \overline{y}_{.jt_0})^2})\sigma^2_{\epsilon^{(2)}}$, $\sigma^2_{\epsilon^{(2)}} = (1-\rho^2)\sigma_1^2$	$\widehat{var}_{ols}(\hat{\beta}^{(2)}_{1,ols}\|Y_{ijt_0}) = (\frac{1}{n_0} + \frac{1}{n_1} + \frac{(\overline{y}_{.1t_0} - \overline{y}_{.0t_0})^2}{\sum_{j=0}^{1}\sum_{i=1}^{n_j}(y_{ijt_0} - \overline{y}_{.jt_0})^2})\hat{\sigma}^2_{e^{(2)}_{ij}}$, $\hat{\sigma}^2_{e^{(2)}_{ij}} = \frac{\sum_{j=0}^{1}\sum_{i=1}^{n_j}(y_{ijt_1}-\hat{y}_{ijt_1})^2}{(n_0+n_1-4)}$
		U	$var(\hat{\beta}^{(2)}_{1,ols}) = (\frac{1}{n_0} + \frac{1}{n_1})(1-\rho^2)\sigma_1^2$	
RM	$\hat{\gamma}^{(3)}_{3,\ gls} = (\overline{y}_{.1t_1} - \overline{y}_{.1t_0}) - (\overline{y}_{.0t_1} - \overline{y}_{.0t_0})$	U	$var(\hat{\gamma}^{(3)}_{3,\ gls}) = (\frac{1}{n_0} + \frac{1}{n_1})(\sigma_1^2 + \sigma_0^2 - 2\rho\sigma_0\sigma_1)$	
cRM	$\hat{\gamma}^{(4)}_{3,\ gls} = (\overline{y}_{.1t_1} - \overline{y}_{.0t_1}) - \frac{\rho\sigma_0\sigma_1}{\sigma_0^2}(\overline{y}_{.1t_0} - \overline{y}_{.0t_0})$	U	$var(\hat{\gamma}^{(4)}_{3,gls}) = (\frac{1}{n_0} + \frac{1}{n_1})(1-\rho^2)\sigma_1^2$	
ANOVA-Change	$\hat{\beta}^{(5)}_{1,ols} = (\overline{y}_{.1t_1} - \overline{y}_{.1t_0}) - (\overline{y}_{.0t_1} - \overline{y}_{.0t_0})$	U	$var(\hat{\beta}^{(5)}_{1,ols}) = (\frac{1}{n_0} + \frac{1}{n_1})(\sigma_1^2 + \sigma_0^2 - 2\rho\sigma_0\sigma_1)$	$\widehat{var}_{ols}(\hat{\beta}^{(5)}_{1,ols}) = \frac{\hat{\sigma}^2_{\epsilon^{(5)}}}{\sum_{j=0}^{1}\sum_{i=1}^{n_j}(G_{ij}-G_{\cdot,\cdot})^2}$, $\hat{\sigma}^2_{\epsilon^{(5)}} = \frac{\sum_{j=0}^{1}\sum_{i=1}^{n_j}(\Delta_{ij}-\hat{\Delta}^{(5)}_{ij})^2}{(n_0+n_1-2)}$

[a]U- unconditional variance; C- conditional variance

The coefficients and standard errors of model (2) are also estimated using an OLS regression. The OLS estimator $\hat{\beta}^{(2)}_{1,ols}$ is derived as the sample mean difference in the post-treatment weight adjusting for the sample mean difference in the baseline weight between two arms. The group mean difference in the baseline weight can be seen as chance imbalance in a randomized trial. $\hat{\beta}^{(2)}_{1,ols}$ is unbiased for τ both conditional on Y_{ijt_0} and unconditionally. The formulas of $\hat{\beta}^{(2)}_{1,ols}$ and its "unconditional" variance $var(\hat{\beta}^{(2)}_{1,ols})$ are listed in Table 1. However, OLS assumes that the baseline weight Y_{ijt_0} is fixed. OLS targets the conditional variance of $\hat{\beta}^{(2)}_{1,ols}$, denoted by $var(\hat{\beta}^{(2)}_{1,ols}|Y_{ijt_0})$, instead of $var(\hat{\beta}^{(2)}_{1,ols})$. The formula of $var(\hat{\beta}^{(2)}_{1,ols}|Y_{ijt_0})$ with a known common residual variance $\sigma^2_{\epsilon^{(2)}}$ is presented in Table 1. Since $\sigma^2_{\epsilon^{(2)}}$ is generally unknown, it is estimated by the following sample residual variance:

$$\hat{\sigma}^2_{e^{(2)}_{ij}} = \frac{\sum_{j=0}^{1}\sum_{i=1}^{n_j}\left(y_{ijt_1}-\hat{y}^{(2)}_{ijt_1}\right)^2}{(n_0+n_1-3)}$$

, where $\hat{y}^{(2)}_{ijt_1} = \hat{\beta}^{(2)}_{0,ols} + \hat{\beta}^{(2)}_{1,ols}G_{ij} + \hat{\beta}^{(2)}_{2,ols}Y_{ijt_0}$, the predicted value from model (2). We let $\widehat{var}_{ols}(\hat{\beta}^{(2)}_{1,ols}|Y_{ijt_0})$ denote the OLS model-based variance estimator with $\hat{\sigma}^2_{\epsilon^{(2)}}$ substituted for $\sigma^2_{\epsilon^{(2)}}$. Note that $\widehat{var}_{ols}(\hat{\beta}^{(2)}_{1,ols}|Y_{ijt_0})$ is reported by standard statistical softwares (e.g. "proc reg" in SAS). Its formula is presented in Table 1.

Since we want to generalize our conclusions to a general population and Y_{ijt_0} can take different values from those collected in the current sample, we may wonder whether significance tests based on the model-based conditional variance assuming Y_{ijt_0} is fixed (e.g., $t = \frac{\hat{\beta}^{(2)}_{1,ols}}{\sqrt{\widehat{var}_{ols}(\hat{\beta}^{(2)}_{1,ols}|Y_{ijt_0})}}$) is comparable to unconditional inference (e.g., $t = \frac{\hat{\beta}^{(2)}_{1,ols}}{\sqrt{var(\hat{\beta}^{(2)}_{1,ols})}}$), in which Y_{ijt_0} is treated as random variable, for testing $H_o: \tau = 0$. To establish this equivalence, we need to show: i) $\widehat{var}_{ols}(\hat{\beta}^{(2)}_{1,ols}|Y_{ijt_0})$ is unbiased for $var(\hat{\beta}^{(2)}_{1,ols}|Y_{ijt_0})$; ii) $var(\hat{\beta}^{(2)}_{1,ols}|Y_{ijt_0})$ is unbiased for $var(\hat{\beta}^{(2)}_{1,ols})$. The first part is well established in a homoscedastic linear model. The second part holds because we can show that $var(\hat{\beta}^{(2)}_{1,ols})$ =E($var(\hat{\beta}^{(2)}_{1,ols}|Y_{ijt_0})$) using the law of total variance formula and the fact that $\hat{\beta}^{(2)}_{1,ols}$ is unbiased for τ. That is, the unconditional variance of $\hat{\beta}^{(2)}_{1,ols}$ is the average of its conditional variance over the distribution of the baseline weight. Therefore, the usual model-based standard errors and associated p-values are valid for unconditional inference [3, 5, 17].

Method 3:*Repeated measures model ("RM"):RM* models the baseline and post-treatment weights (Y_{ijt_0} , Y_{ijt_1}) jointly using the binary treatment indicator G_{ij}, the binary time factor T_{ij}, the time by treatment interaction $G_{ij} \times T_{ij}$ as follows:

$$\begin{aligned} Y_{ijt} &= \gamma^{(3)}_0 + \gamma^{(3)}_1 G_{ij} + \gamma^{(3)}_2 T_{ij} + \gamma^{(3)}_3 G_{ij} \times T_{ij} \\ &\quad + e^{(3)}_{ijt}, i \\ &= 1, 2, ..., n_j; j = 0, 1; t = t_0, t_1, \end{aligned} \tag{3}$$

$$\begin{pmatrix} e^{(3)}_{ijt_0} \\ e^{(3)}_{ijt_1} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sum\right),$$

When $t_0 = 0$ and $t_1 = 1$, $\gamma^{(3)}_0 = \mu_{0t_0}$, $\gamma^{(3)}_1 = \mu_{1t_0} - \mu_{0t_0}$, $\gamma^{(3)}_2 = \mu_{0t_1} - \mu_{0t_0}$, and $\gamma^{(3)}_3 = (\mu_{1t_1} - \mu_{1t_0}) - (\mu_{0t_1} - \mu_{0t_0})$. $\gamma^{(3)}_0$ represents the mean baseline weight of the control arm, $\gamma^{(3)}_1$ represents the difference in the mean baseline weights of the treatment and control arms, $\gamma^{(3)}_2$ represents the mean change from baseline in the control arm, and $\gamma^{(3)}_3$ is generally interpreted as the difference in the mean change from baseline in a unit time interval between the treatment and control arms ("difference in difference"), also known as the difference in slopes. We have $\mu_{1t_0} = \mu_{0t_0}$ from random allocation and it follows that $\gamma^{(3)}_1 = 0$ and $\gamma^{(3)}_3 = \mu_{1t_1} - \mu_{1t_1} = \tau$. Thus, testing $H_o: \gamma^{(3)}_3 = 0$ is equivalent to testing $H_o: \tau = 0$.

The generalized least squares (GLS) model with correlated outcomes is routinely used to estimate the coefficients and standard errors of model (3). The GLS estimator of the treatment effect $\hat{\gamma}^{(3)}_{3,\ gls}$ and its variance $var(\hat{\gamma}^{(3)}_{3,\ gls})$ given known variance and covariance parameters are presented in Table 1. $\hat{\gamma}^{(3)}_{3,\ gls}$ is estimated by the sample mean difference in body weight change between two arms and is unbiased for τ in a large sample. The variance and covariance parameters are generally unknown and need to be estimated using the restricted maximum likelihood (REML). The conventional maximal likelihood estimation (MLE) should be avoided. The REML variance estimator $\widehat{var}_{reml}(\hat{\gamma}^{(3)}_{3,\ gls})$ is derived by plugging the REML estimators of the variance and covariance parameters (i.e., $\sigma^2_0, \sigma^2_1, \rho\sigma_0\sigma_1$) into the formula of $var(\hat{\gamma}^{(3)}_{3,\ gls})$.We use Kenward and Roger method

[18](“ddfm = kenwardroger” in SAS proc. mixed procedure) to adjust for the potential finite sample bias in $\widehat{var}_{reml}(\hat{\gamma}^{(3)}_{3,\ gls})$ because of its failure to incorporate variabilities of the REML estimators of the variance and covariance parameters. This adjustment involves inflating the variance and covariance matrix and computing an adjusted approximation degrees of freedom.

Method 4:*Constrained Repeated measures Model (“cRM”)*: By specifying $\gamma^{(3)}_1$ in the model, ***RM*** model (3) assumes the mean baseline weight is different between two arms. Liang and Zeger [8] proposed the following ***cRM*** model by fixing $\gamma^{(3)}_1 = 0$ to force the treatment and control arms to have the same intercept. Intuitively, ***cRM*** is more efficient than ***RM*** because ***cRM*** estimates one less parameter. Formally, we model the baseline and post-treatment weights (Y_{ijt_0}, Y_{ijt_1}) jointly using the binary factor T_{ij}, a time by treatment interaction $G_{ij} \times T_{ij}$ in the following ***cRM*** model:

$$Y_{ijt} = \gamma^{(4)}_0 + \gamma^{(4)}_2 T_{ij} + \gamma^{(4)}_3 G_{ij} \times T_{ij} + e^{(4)}_{ijt}, i = 1, 2, ..., n_j; j = 0, 1; t = t_0, t_1 \quad (4)$$

$$\begin{pmatrix} e^{(4)}_{ijt_0} \\ e^{(4)}_{ijt_1} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sum \right),$$

where $\gamma^{(4)}_0 = \mu_{t_0}, \gamma^{(4)}_2 = \mu_{0t_1} - \mu_{0t_0}$, and $\gamma^{(4)}_3 = \tau$. Interpretations of $\gamma^{(4)}_0$, $\gamma^{(4)}_2$, and $\gamma^{(4)}_3$ are the same as their counterparts in ***RM***. The formulas of the GLS point estimator $\hat{\gamma}^{(4)}_{3,\ gls}$ and its variance $var(\hat{\gamma}^{(4)}_{3,\ gls})$ are listed in Table 1. $\hat{\gamma}^{(4)}_{3,\ gls}$ is unbiased for τ asymptotically. The empirical or the model-based variance estimate for $var(\hat{\gamma}^{(4)}_{3,\ gls})$ is derived using REML in the same way as a regular ***RM*** model.

Method 5:*ANOVA with change score (“ANOVA-Change”)*: We model change score $\Delta_{ij} = Y_{ijt_1} - Y_{ijt_0}$ using the binary treatment indicator G_{ij} as follows:

$$\Delta_{ij} = \beta^{(5)}_0 + \beta^{(5)}_1 G_{ij} + e^{(5)}_{ij}, i = 1, 2, ..., n_j; j = 0, 1; \quad (5)$$

$$e^{(5)}_{ij} \sim N\left(0, \sigma^2_{\epsilon^{(5)}}\right) \text{ and } \sigma^2_{\epsilon^{(5)}} = \sigma^2_1 + \sigma^2_0 - 2\rho\sigma_0\sigma_1,$$

where $\beta^{(5)}_0 = \mu_{0t_1} - \mu_{0t_0}$, $\beta^{(5)}_1 = (\mu_{1t_1} - \mu_{1t_0}) - (\mu_{0t_1} - \mu_{0t_0})$, and $e^{(3)}_{ij}$ is i.i.d random error. $\beta^{(5)}_0$ measures the mean difference score in the control arm. $\beta^{(5)}_1$ measures the treatment effect $\tilde{\tau}$. Since $\mu_{1t_0} = \mu_{0t_0}$ due to randomization at baseline, $\beta^{(5)}_1$ is reduced to τ. The closed-form expressions of $\hat{\beta}^{(5)}_{1,ols}$ and $var(\hat{\beta}^{(5)}_{1,ols})$ are listed in Table 1. $\hat{\beta}^{(5)}_{1,ols}$ is derived as the sample mean difference in the change score between two arms (“difference in difference”) and is unbiased for τ. The OLS model-based variance of $\hat{\beta}^{(5)}_{1,ols}$ assuming known $\sigma^2_{\epsilon^{(5)}}$ is

$$var_{ols}\left(\hat{\beta}^{(5)}_{1,ols}\right) = \frac{\sigma^2_{\epsilon^{(5)}}}{\sum^1_{j=0}\sum^{n_j}_{i=1}\left(G_{ij} - G_{..}\right)^2},$$

where $G_{..} = \frac{\sum^1_{j=0}\sum^{n_j}_{i=1} G_{ij}}{n_0+n_1} = \frac{n_1}{n_0+n_1}$. $\sigma^2_{\epsilon^{(5)}}$ is estimated by

$$\hat{\sigma}^2_{\epsilon^{(5)}} = \frac{\sum^1_{j=0}\sum^{n_j}_{i=1}\left(\Delta_{ij} - \hat{\Delta}^{(5)}_{ij}\right)^2}{(n_0 + n_1 - 2)},$$

where $\hat{\Delta}^{(5)}_{ij}$ is the fitted value from model (5). We let $\widehat{var}_{ols}(\hat{\beta}^{(5)}_{1,ols})$ denote the OLS model-based variance estimator with $\hat{\sigma}^2_{\epsilon^{(5)}}$ substituted for $\sigma^2_{\epsilon^{(5)}}$ Table 1, which is reported by standard statistical softwares. Since $\sum^1_{j=0}\sum^{n_j}_{i=1}(G_{ij} - G_{..})^2 = \frac{n_0 n_1}{n_0+n_1}$, it follows that $var_{ols}(\hat{\beta}^{(5)}_{1,ols}) = var(\hat{\beta}^{(5)}_{1,ols})$. It is well established that $\widehat{var}_{ols}(\hat{\beta}^{(5)}_{1,ols})$ is unbiased for $var_{ols}(\hat{\beta}^{(5)}_{1,ols})$, and thus for $var(\hat{\beta}^{(5)}_{1,ols})$. The usual OLS model-based inference is valid for unconditional hypothesis testing.

When the study population is heterogeneous

Method 6:*ANCOVA*II: Different variance and covariance structures in the treatment and control arms suggest a baseline measurement by treatment interaction term in ANCOVA [2, 3, 9, 10]. To estimate τ using an interaction model, we first compute the mean centered baseline weight $\tilde{Y}_{ijt_0}$ by subtracting the overall mean baseline weight from individual baseline weights. i.e., $\tilde{Y}_{ijt_0} = Y_{ijt_0} - \mu_{t_0}$. We then model the post-treatment body weight Y_{ijt_1} using the binary treatment indicator G_{ij}, the mean centered baseline weight $\tilde{Y}_{ijt_0}$, and the baseline weight by treatment interaction $G_{ij} \times \tilde{Y}_{ijt_0}$ as follows:

$$Y_{ijt_1} = \beta^{(6)}_0 + \beta^{(6)}_1 G_{ij} + \beta^{(6)}_2 \tilde{Y}_{ijt_0} + \beta^{(6)}_3 G_{ij} \times \tilde{Y}_{ijt_0} + e^{(6)}_{ij}, i = 1, 2, ..., n_j; j = 0, 1; \quad (6)$$

$$e^{(6)}_{i0} \sim N\left(0,\ \sigma^2_{\epsilon^{(6)}_0}\right) \text{ and } \sigma^2_{\epsilon^{(6)}_0} = \left(1 - \rho^2_0\right)\sigma^2_{01}$$

$$e^{(6)}_{i1} \sim N\left(0,\ \sigma^2_{\epsilon^{(6)}_1}\right) \text{ and } \sigma^2_{\epsilon^{(6)}_1} = \left(1 - \rho^2_1\right)\sigma^2_{11}$$

, where $\beta^{(6)}_0 = \mu_{0t_1}$, $\beta^{(6)}_1 = \tau$, $\beta^{(6)}_2 = \rho_0 \frac{\sigma_{0t_0}}{\sigma_0}$, and $\beta^{(6)}_3 = \rho_1 \frac{\sigma_{1t_1}}{\sigma_0} - \rho_0 \frac{\sigma_{0t_0}}{\sigma_0}$. $e^{(6)}_{i0}$ and $e^{(6)}_{i1}$ are i.i.d random errors in the control and treatment arms. $\beta^{(6)}_1$ measures the treatment effect. $\beta^{(6)}_2$ is the regression slope of the baseline body weight in the control arm. $\beta^{(6)}_3$ measures the difference

in the regression slopes of the baseline weight between the treatment and control arms. Model (6) is heteroscedastic because the error terms in the treatment and control arms have different residual variances.

As presented in Table 2, the OLS estimator $\hat{\beta}^{(6)}_{1,ols}$ is the adjusted mean difference in the post-treatment body weights controlling for a weighted mean difference of the baseline body weights between two arms with unequal weighting coefficients for treatment and control arms (i.e., $\hat{\beta}^{(6)}_{2,ols}+\hat{\beta}^{(6)}_{3,ols}$ for the treatment group, and $\hat{\beta}^{(6)}_{2,ols}$ for the control group). $\hat{\beta}^{(6)}_{1,ols}$ is unbiased for τ. The conditional variance of $\hat{\beta}^{(6)}_{1,ols}$, denoted by $var(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$, incorporates two different residual variances $\sigma^2_{\epsilon^{(6)}_0}$ and $\sigma^2_{\epsilon^{(6)}_1}$ (Table 2). Standard statistical softwares such as SAS do not output $var(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$ because OLS incorrectly assumes a common residual variance $\sigma^2_{\epsilon^{(6)}}$, which is the following weighted average of $\sigma^2_{\epsilon^{(6)}_0}$ and $\sigma^2_{\epsilon^{(6)}_1}$:

$$\sigma^2_{\epsilon^{(6)}}=\frac{n_0}{n_0+n_1}\ \sigma^2_{\epsilon^{(6)}_0}+\frac{n_1}{n_0+n_1}\sigma^2_{\epsilon^{(6)}_1}$$

We let $var_{ols}(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$ denote the OLS model-based conditional variance of $\hat{\beta}^{(6)}_{1,ols}$ incorporating $\sigma^2_{\epsilon^{(6)}}$ (Table 2). Since $\sigma^2_{\epsilon^{(6)}}$ is generally unknown, $\sigma^2_{\epsilon^{(6)}}$ is estimated by

$$\hat{\sigma}^2_{\epsilon^{(6)}}=\frac{\sum^1_{j=0}\sum^{n_j}_{i=1}\left(y_{ijt_1}-\hat{y}_{ijt_1}\right)^2}{(n_0+n_1-4)},$$

where $\hat{y}_{ijt_1}$ is the predicted value of y_{ijt_1}. We let $\widehat{var}_{ols}(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$ denote the OLS model-based variance estimator of $\hat{\beta}^{(6)}_{1,ols}$ with $\hat{\sigma}^2_{\epsilon^{(6)}}$ substituted for $\sigma^2_{\epsilon^{(6)}}$. and known constant μ_{t_0} (Table 2). $\widehat{var}_{ols}(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$ is reported by standard statistical softwares (e.g., "proc reg" in SAS). To assess the validity of the model-based standard errors and *p*-values from a regular ***ANCOVAII*** model for unconditional inference, we need to examine: i) whether $\widehat{var}_{ols}(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$ is unbiased for $var(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$; ii) whether $var(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$ is unbiased for $var(\hat{\beta}^{(6)}_{1,ols})$.

First, $\widehat{var}_{ols}(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$ is unbiased for $var_{ols}(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$. However, the unbiasedness of $\widehat{var}_{ols}(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$ as an estimator of $var(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$ depends on the relationship between $var_{ols}(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$ and $var(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$. Asymptotically, we have

$$\begin{aligned}\Delta_{\hat{\beta}^{(6)}_{1,ols}} &= var_{ols}(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})-var(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})\\ &=\left(\sigma^2_{\epsilon^{(6)}_0}-\sigma^2_{\epsilon^{(6)}_1}\right)\left(\frac{1}{n_1}-\frac{1}{n_0}\right)\end{aligned}$$

It can be shown in a balanced design ($n_0 = n_1$),

$$var_{ols}(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})\approx var(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0}).$$

Thus, $\widehat{var}_{ols}(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$ is nearly unbiased for $var(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$ [3]. When the design is unbalanced ($n_0 \neq n_1$),

$$var_{ols}(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})\neq var(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0}).$$

Hence, $\widehat{var}_{ols}(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$ is biased for $var(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$. Due to heteroscedasticity, $\widehat{var}_{ols}(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$ overestimates $var(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$ if the group with a larger residual variance has larger sample size and the group with a smaller residual variance has smaller sample size, and otherwise may underestimate $var(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$ [3, 4].

Second, the common mean baseline weight μ_{t_0} is generally unknown. We need to estimate μ_{t_0} in $\tilde{Y}_{ijt_0}$ using the overall sample mean $\hat{\mu}_{t_0}=\frac{\sum^1_{j=0}\sum^{n_j}_{i=1}Y_{ijt_0}}{n_0+n_1}$ but ANCOVA treats $\hat{\mu}_{t_0}$ as fixed and fails to capture this additional variability in the conditional variances. As shown below, it turns out that $var(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$ underestimates $var(\hat{\beta}^{(6)}_{1,ols})$ by a factor of $\beta^{[6]2}_{3,ols}\ var(\hat{\mu}_{t_0})$ [3]:

$$var\left(\hat{\beta}^{(6)}_{1,ols}\right)=E\left(var\left(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0}\right)\right)+\beta^{(6)2}_{3,ols}\ var(\hat{\mu}_{t_0}).$$

Thus, the OLS model-based conditional inference is biased for unconditional hypothesis testing because of heteroscedasticity and neglecting of sampling variability in $\hat{\mu}_{t_0}$. To fix these two problems, we can use the following adjusted heteroscedasticity-consistent (HC) variance estimator to replace $\widehat{var}_{ols}(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$ for valid unconditional inference:

$$\widehat{var}_{aHC}(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})=\widehat{var}_{HC}(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})+\hat{\beta}^{(6)2}_{3,ols}\frac{\hat{\sigma}^2_0}{n_0+n_1},$$

where $\widehat{var}_{HC}(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$ is a HC variance estimator for $var(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$ [19] and can be output from standard softwares. HC variance estimators are consistent (i.e., unbiased in large sample). Among all available HC variance estimators, HC2 was shown to have the best

Table 2 Estimators of treatment effect and variance estimators in a heterogeneous study population

Model	Estimator of treatment effect (τ)	Type	True variance of treatment effect estimator	Variance estimator from OLS model
ANCOVA-Post II	$\hat{\beta}^{(6)}_{1,ols} = (\overline{y}_{.1t_1} - (\hat{\beta}^{(6)}_{2,ols} + \hat{\beta}^{(6)}_{3,ols})\overline{\tilde{y}}_{.1t_0}) - (\overline{y}_{.0t_0} - \hat{\beta}^{(6)}_{2,ols}\overline{\tilde{y}}_{.0t_0})$	C	$var(\hat{\beta}^{(6)}_{1,ols}\|\tilde{Y}_{ijt_0}) = (\frac{1}{n_0} + \frac{\overline{\tilde{y}}^2_{.\alpha_0}}{\sum_{i=1}^{n_0}(\tilde{y}_{i\alpha_0} - \overline{\tilde{y}}_{.\alpha_0})^2})\,\sigma^2_{\epsilon_0^{(6)}} + (\frac{1}{n_1} + \frac{\overline{\tilde{y}}^2_{.1t_0}}{\sum_{i=1}^{n_0}(\tilde{y}_{i1t_0} - \overline{\tilde{y}}_{.1t_0})^2})\,\sigma^2_{\epsilon_1^{(6)}}$ $\sigma^2_{\epsilon_0^{(6)}} = (1-\rho_0^2)\sigma^2_{01}, \quad \sigma^2_{\epsilon_1^{(6)}} = (1-\rho_1^2)\sigma^2_{11}$	$\widehat{var}_{ols}(\hat{\beta}^{(6)}_{1,ols}\|\tilde{Y}_{ijt_0}) = (\frac{1}{n_0} + \frac{1}{n_1} + \frac{\overline{\tilde{y}}^2_{.\alpha_0}}{\sum_{i=1}^{n_0}(\tilde{y}_{i\alpha_0} - \overline{\tilde{y}}_{.\alpha_0})^2} + \frac{\overline{\tilde{y}}^2_{.1t_0}}{\sum_{i=1}^{n_0}(\tilde{y}_{i1t_0} - \overline{\tilde{y}}_{.1t_0})^2})\hat{\sigma}^2_{\epsilon^{(6)}}$ $\hat{\sigma}^2_{\epsilon^{(6)}} = \frac{\sum_{j=0}^{1}\sum_{i=1}^{n_j}(y_{ijt_1} - \hat{y}_{ijt_1})^2}{(n_0+n_1-5)}$
		U	$var(\hat{\beta}^{(6)}_{1,ols}) = \frac{1}{n_0}(1-\rho_0^2)\sigma^2_{01} + \frac{1}{n_1}(1-\rho_1^2)\sigma^2_{11} + (\rho_1\frac{\sigma_{11}}{\sigma_0} - \rho_0\frac{\sigma_{01}}{\sigma_0})^2\frac{\sigma_0^2}{n_0+n_1}$	
ANCOVA-Post I	$\hat{\beta}^{(7)}_{1,ols} = (\overline{y}_{.1t_1} - \overline{y}_{.0t_1}) - \hat{\beta}^{(7)}_{2,ols}(\overline{y}_{.1t_0} - \overline{y}_{.0t_0})$	C	$var(\hat{\beta}^{(7)}_{1,ols}\|Y_{ijt_0}) = (\frac{1}{n_0} + \frac{\sum_{i=1}^{n_0}(y_{i1t_0} - \overline{y}_{.0t_0})^2(\overline{y}_{.1t_0} - \overline{y}_{.0t_0})}{\sum_{j=0}^{1}\sum_{i=1}^{n_j}(y_{ijt_0} - \overline{y}_{.jt_0})^2})\,\sigma^2_{\epsilon_0^{(7)}} + (\frac{1}{n_1} + \frac{\sum_{i=1}^{n_1}(y_{i1t_0} - \overline{y}_{.1t_0})^2(\overline{y}_{.1t_0} - \overline{y}_{.0t_0})}{\sum_{j=0}^{1}\sum_{i=1}^{n_j}(y_{i1t_0} - \overline{\tilde{y}}_{.1t_0})^2})\,\sigma^2_{\epsilon_1^{(7)}}$ $\sigma^2_{\epsilon_0^{(7)}} = (1-\rho_0^2)\sigma^2_{01}, \quad \sigma^2_{\epsilon_1^{(7)}} = (1-\rho_1^2)\sigma^2_{11}$	$\widehat{var}_{ols}(\hat{\beta}^{(7)}_{1,ols}\|Y_{ijt_0}) = (\frac{1}{n_0} + \frac{1}{n_1} + \frac{\sum_{i=1}^{n_0}(y_{i1t_0} - \overline{y}_{.0t_0})^2(\overline{y}_{.1t_0} - \overline{y}_{.0t_0})}{\sum_{j=0}^{1}\sum_{i=1}^{n_j}(y_{ijt_0} - \overline{y}_{.jt_0})^2} + \frac{\sum_{i=1}^{n_1}(y_{i1t_0} - \overline{y}_{.1t_0})^2(\overline{y}_{.1t_0} - \overline{y}_{.0t_0})}{\sum_{j=0}^{1}\sum_{i=1}^{n_j}(y_{i1t_0} - \overline{\tilde{y}}_{.1t_0})^2})\hat{\sigma}^2_{\epsilon^{(7)}}$ $\hat{\sigma}^2_{\epsilon^{(7)}} = \frac{\sum_{j=0}^{1}\sum_{i=1}^{n_j}(y_{ijt_1} - \hat{y}_{ijt_1})^2}{(n_0+n_1-4)}$
		U	$var(\hat{\beta}^{(7)}_{1,ols}) = \frac{1}{n_0}[(1-\rho_0^2)\sigma^2_{01} + ((\rho_1\frac{\sigma_{11}}{\sigma_0} - \rho_0\frac{\sigma_{01}}{\sigma_0})p_1)^2\sigma^2_0 + \frac{1}{n_1}[(1-\rho_1^2)\,\sigma^2_{11} + ((\rho_1\frac{\sigma_{11}}{\sigma_0} - \rho_0\frac{\sigma_{01}}{\sigma_0})p_0)^2\sigma^2_0]$	
cRM	$\hat{\gamma}^{(4)}_{3,\ gls} = (\overline{y}_{.1t_1} - \overline{y}_{.0t_1}) - (\frac{\rho_0\sigma_0\sigma_{01}}{\sigma_0^2}(\overline{y}_{.1t_0} - \overline{y}_{..t_0}) - \frac{\rho_1\sigma_0\sigma_{11}}{\sigma_0^2}(\overline{y}_{.1t_0} - \overline{y}_{..t_0})$	U	$var(\hat{\gamma}^{(4)}_{3,\ gls}) = \frac{1}{n_0}[(1-\rho_0^2)\sigma^2_{01} + ((\rho_1\frac{\sigma_{11}}{\sigma_0} - \rho_0\frac{\sigma_{01}}{\sigma_0})p_1)^2\sigma^2_0 + \frac{1}{n_1}[(1-\rho_1^2)\,\sigma^2_{11} + ((\rho_1\frac{\sigma_{11}}{\sigma_0} - \rho_0\frac{\sigma_{01}}{\sigma_0})p_0)^2\sigma^2_0]$	

performance in finite samples [3, 4] (e.g. "HCCMETHOD = 2" in proc. reg or "EMPIRICAL" in proc. mixed, SAS). $\hat{\beta}^{[6]}_{3,ols}$ is the OLS estimator of $\beta^{(6)}_3$, and $\hat{\sigma}^2_0$ is the overall sample variance of the baseline body weight. It follows directly that $\widehat{var}_{aHC}(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})$ is asymptotically unbiased for $var(\hat{\beta}^{(6)}_{1,ols})$ and we can construct a valid test $t = \frac{\hat{\beta}^{(6)}_{1,ols}}{\sqrt{\widehat{var}_{aHC}(\hat{\beta}^{(6)}_{1,ols}|\tilde{Y}_{ijt_0})}}$ for testing H_o: $\tau = 0$ unconditionally.

Method 7*ANCOVA*I: We model the post-treatment weight Y_{ijt_1} using the binary treatment G and the baseline weight Y_{ijt_0}:

$$Y_{ijt_1} = \beta^{(7)}_0 + \beta^{(7)}_1 G_{ij} + \beta^{(7)}_2 Y_{ijt_0} + e^{(7)}_{ij} \quad (7)$$

$$e^{(7)}_{i0} \sim N\left(0,\ \sigma^2_{\epsilon^{(7)}_0}\right) \text{ and } \sigma^2_{\epsilon^{(7)}_0} = (1-\rho^2_0)\sigma^2_{01} + \left(\beta^{(6)}_3 p_1\right)^2 \sigma^2_0$$

$$e^{(7)}_{i1} \sim N\left(0,\ \sigma^2_{\epsilon^{(7)}_1}\right) \text{ and } \sigma^2_{\epsilon^{(7)}_1} = (1-\rho^2_1)\sigma^2_{11} + \left(\beta^{(6)}_3 p_0\right)^2 \sigma^2_0$$

, where $\beta^{(7)}_0 = \beta^{(6)}_0 - \beta^{(6)}_3 p_0 \mu_0$, and $\beta^{(7)}_1 = \tau$. $e^{(7)}_{i0}$ and $e^{(7)}_{i1}$ are random errors in the control and treatment arms. Since $e^{(7)}_{i0}$ and $e^{(7)}_{i1}$ have different variances in general, model (7) is heteroscedastic and the severity of heteroscedasticity is determined by the correlation coefficient, the variances of the post-treatment weights in two arms, and whether the design is balanced.

As shown in Table 2, the OLS estimator $\hat{\beta}^{(7)}_{1,ols}$ is an adjusted mean difference in the post-treatment weights controlling for a weighted mean difference of the baseline weights between two arms with equal weighting coefficient for the treatment and control arms (i.e., $\hat{\beta}^{(7)}_{2,ols}$ for both arms). $\hat{\beta}^{(7)}_{1,ols}$ is unbiased for τ. The true conditional variance $var(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0})$ incorporates two different residual variances. Similar to ***ANCOVA*II**, the OLS model-based inference for ***ANCOVA*I** also mistakenly assumes a constant residual variance $\sigma^2_{\epsilon^{(7)}}$, which is a weighted average of $\sigma^2_{\epsilon^{(7)}_0}$ and $\sigma^2_{\epsilon^{(7)}_1}$, as follows:

$$\sigma^2_{\epsilon^{(7)}} = \frac{n_0}{n_0+n_1}\ \sigma^2_{\epsilon^{(7)}_0} + \frac{n_1}{n_0+n_1}\sigma^2_{\epsilon^{(7)}_1}.$$

Since $\sigma^2_{\epsilon^{(7)}}$ is unknown, it is estimated by

$$\hat{\sigma}^2_{\epsilon^{(7)}} = \frac{\sum^1_{j=0}\sum^{n_j}_{i=1}\left(y_{ijt_1} - \hat{y}_{ijt_1}\right)^2}{n_0+n_1-3},$$

where $\hat{y}_{ijt_1}$ is the predicted value of y_{ijt_1} from model (7). The closed form expressions of the OLS model-based conditional variance $var_{ols}(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0})$ incorporating $\sigma^2_{\epsilon^{(7)}}$ and the OLS model-based variance estimator $\widehat{var}_{ols}(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0})$ with $\hat{\sigma}^2_{\epsilon^{(7)}}$ substituted for $\sigma^2_{\epsilon^{(7)}}$ are given in Table 2. Recall that standard statistical softwares report $\widehat{var}_{ols}(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0})$. To show the model-based standard errors and p-values are valid for unconditional inference, we need to examine: i) whether $\widehat{var}_{ols}(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0})$ is unbiased for $var(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0})$; ii) whether $var(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0})$ is unbiased for $var(\hat{\beta}^{(7)}_{1,ols})$.

First, $\widehat{var}_{ols}(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0})$ is unbiased for $var_{ols}(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0})$ but the unbiasedness of $\widehat{var}_{ols}(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0})$ as an estimator of $var(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0})$ depends on the relationship between $var_{ols}(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0})$ and $var(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0})$. Asymptotically, we have

$$\Delta_{\hat{\beta}^{(7)}_{1,ols}} = var_{ols}(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0}) - var\left(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0}\right) = \left(\sigma^2_{\epsilon^{(7)}_0} - \sigma^2_{\epsilon^{(7)}_1}\right)\left(\frac{1}{n_1} - \frac{1}{n_0}\right)$$

When sample sizes are equal between two arms, we have

$$var_{ols}(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0}) \approx var(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0}).$$

Thus, $\widehat{var}_{ols}(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0})$ is nearly unbiased for $var(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0})$ in a balanced design [3]. When sample sizes are not equal between two arms,

$$var_{ols}(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0}) \neq var(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0}),$$

it follows directly that $\widehat{var}_{ols}(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0})$ is biased for $var(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0})$ due to heteroscedasticity. $\widehat{var}_{ols}(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0})$ may over-estimate $var(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0})$ when the group with a larger residual variance has larger sample size and the group with a smaller residual variance has smaller sample size, and otherwise may underestimate $var(\hat{\beta}^{(7)}_{1,ols}|\tilde{Y}_{ijt_0})$ [3, 4] . ***ANCOVA*I** is robust against heteroscedasticity in a balanced design, but not in an unbalanced design.

Second, different from ***ANCOVA*II**, $var(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0})$ is unbiased for $var(\hat{\beta}^{(7)}_{1,ols})$ because $var(\hat{\beta}^{(7)}_{1,ols}) = E(\ var(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0}))$.

Thus, the model-based standard errors and *p*-values are valid for unconditional inference in a balanced design but are biased in an unbalanced design only due to heteroscedasticity. This bias can be easily corrected by replacing $\widehat{var}_{ols}(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0})$ with an HC variance estimator $\widehat{var}_{HC}(\hat{\beta}^{(7)}_{1,ols}|Y_{ijt_0})$ [4, 19] and corrected ***ANCOVA*I** will provide valid unconditional inference.

***Constrained Repeated Measures heterogeneous variance model ("cRM")*:** We model the baseline and post-treatment weights (Y_{ijt_0}, Y_{ijt_1}) jointly using the binary time point T_{ij}, time by treatment interaction $G_{ij} \times T_{ij}$:

$$Y_{ijt} = \gamma_0^{(8)} + \gamma_1^{(8)} T_{ij} + \gamma_2^{(8)} G_{ij} \times T_{ij} + e_{ijt}^{(8)} \quad j = 0, 1; i = 1, 2, \ldots n_j. \tag{8}$$

$$\begin{pmatrix} e^{(8)}_{i0t_0} \\ e^{(8)}_{i0t_1} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sum\nolimits_0 \right) \text{ in the control arm,}$$

$$\begin{pmatrix} e^{(8)}_{i1t_0} \\ e^{(8)}_{i1t_1} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sum\nolimits_1 \right) \text{ in the treatment arm,}$$

where $\gamma_0^{(8)} = \mu_{t_0}, \gamma_2^{(8)} = \mu_{0t_1} - \mu_{0t_0}$, and $\gamma_2^{(8)} = \tau$. Noting that subjects in the treatment and control arms have different variance-covariance structures for the association between the pre- and post-treatment weights, we fit a ***cRM*** heterogeneous variance GLS model with group specific variance-covariance structure ("repeated/group= " in SAS proc. mixed procedure specifies distinct variance-covariance structure for each treatment arm). The formulas of $\hat{\gamma}^{(8)}_{2,gls}$ and $var(\hat{\gamma}^{(8)}_{2,gls})$ are listed in Table 2. The GLS estimator $\hat{\gamma}^{(8)}_{2,gls}$ is asymptotically unbiased for $\gamma_2^{(8)}$. REML is used to derive the empirical or model-based variance estimator $\widehat{var}_{reml}(\hat{\gamma}^{(8)}_{2,\ gls})$.

Results

All treatment effect estimators, except the ANOVA estimator, are expressed as the mean difference in post-treatment measurements adjusting for the chance imbalance in baseline measurement between two arms in certain ways. Nonetheless, all estimators are unbiased for τ. To compare these competing methods, we evaluate the efficiency of point estimators of treatment effect by comparing their "unconditional" variances. Since the hypothesis testing of no treatment effect is based on dividing the point estimator by its standard error (i.e., variance divided by sample size) and rejecting the null hypothesis when this ratio exceeds a given threshold, the method that produces unbiased point estimate with the smallest unconditional variance is preferred because standard error in the dominator of statistical test determines the statistical power.

When study population is homogeneous

ANCOVA*I** is a more efficient alternative to ***ANOVA because $var(\hat{\beta}^{(2)}_{1,ols}) \leq var(\hat{\beta}^{(1)}_{1,ols})$ (Table 1). This advantage of ANCOVA over ANOVA can also be observed from the fact that the residual error variance of ***ANCOVA*I** is less than the residual error variance of ***ANOVA*** (i.e.,$(1-\rho^2)\sigma_1^2 \leq \sigma_1^2$). When the correlation coefficient ρ becomes larger, the ***ANCOVA*I** estimator has smaller variance. Since Y_{ijt_1} and Y_{ijt_0} are highly correlated in general, the inclusion of Y_{ijt_0} in ***ANCOVA*I** explains away some variability in Y_{ijt_1} and thus reduces the residual variance and yields a more efficient estimator of treatment effect than ***ANOVA***.

ANOVA-Change and ***RM*** have exactly same point estimators of τ and thus have the same variances (Table 1). To compare ***ANOVA-Change*** or ***RM*** with ***ANOVA***, we can derive the difference between the unconditional variances of their treatment effect estimators as follows:

$$\Delta_1 = \sigma_0(1-2\rho\sigma_1).$$

When $\rho < \frac{1}{2\sigma_1}$, $\Delta_1 > 0$ and ***ANOVA*** outperforms ***ANOVA-Change*** and ***RM*** because the ***ANOVA*** estimator has smaller variance**.** When $\rho > \frac{1}{2\sigma_1}$, $\Delta_1 < 0$ and ***ANOVA*** underperforms the other two methods.

It can be shown that the difference between the unconditional variances of the ***ANCOVA*I** or ***cRM*** estimators and those of **the*ANOVA-Change*** or ***RM*** estimators are always nonnegative:

$$\begin{aligned} \Delta_2 &= \left(\sigma_1^2 + \sigma_0^2 - 2\rho\sigma_0\sigma_1\right) - \left(1-\rho^2\right)\sigma_1^2 \\ &= (\sigma_0 - \rho\sigma_1)^2 \geq 0 \end{aligned}$$

Thus, ***ANOVA-Change*** or ***RM*** is less efficient than either ***ANCOVA*I** or ***cRM*** because their estimators have larger variances. Intuitively ***ANCOVA*I** or ***cRM*** assumes that mean baseline weights in two arms are equal in a randomized study but ***ANOVA-Change*** or ***RM*** assumes that there is a baseline difference and needs to estimate an extra parameter.

As shown in Table 1, the ***ANCOVA*I** and ***cRM*** estimators of τ are equivalent because $\beta^{(2)}_{1,ols} = \frac{\rho\sigma_0\sigma_1}{\sigma_0^2}$. However, ***ANCOVA*I** plugs in the OLS estimators $\hat{\beta}^{(2)}_{1,ols}$, whereas ***cRM*** plugs in the REML estimators of the variance and covariance parameters. The numerical difference between $\hat{\beta}^{(2)}_{1,ols}$ and $\hat{\gamma}^{(4)}_{3,gls}$ becomes negligible as sample size

increases. Because of this equivalence between $\hat{\beta}^{(2)}_{1,ols}$ and $\hat{\gamma}^{(4)}_{3,gls}$, $var(\hat{\beta}^{(2)}_{1,ols})$ and $var(\hat{\gamma}^{(4)}_{3,\ gls})$ are equal [3]. As discussed previously, ***ANCOVA*I** is a conditional model assuming fixed baseline covariates. Even though the model-based variance estimates are conditional, they are unbiased for the unconditional variance and thus the usual model-based conditional inference is still valid for unconditional hypothesis testing. ***ANCOVA*I** performs comparably to ***cRM*** [3, 17]**.**

When study population is heterogeneous

A heterogeneous study population justifies the inclusion of a treatment by baseline weight interaction term. Thus, ***ANCOVA*II** is the correctly specified model, whereas ***ANCOVA*I** is a mis-specified model. In this case, the "conditional" treatment effect is not constant across different values of baseline weight. The "marginal" treatment effect τ is simply the average of the conditional treatment effect over the distribution of the baseline weight and measures an overall treatment effect. As shown previously, both ANCOVA models can be used to estimate τ even though ***ANCOVA*I** is mis-specified. Then, what is the advantage of using a more complex interaction model over a main effect model? It turns out the ***ANCOVA*II** estimator $\hat{\beta}^{(6)}_{1,ols}$ is more efficient than the ***ANCOVA*I** estimator $\hat{\beta}^{(7)}_{1,ols}$ because $var(\hat{\beta}^{(6)}_{1,ols}) \leq var(\hat{\beta}^{(7)}_{1,ols})$ [5]. Only in a balanced design $var(\hat{\beta}^{(6)}_{1,ols}) = var(\hat{\beta}^{(7)}_{1,ols})$ and the two ANCOVA models perform comparably. Note that the OLS model-based variance estimates for ***ANCOVA*I** and **II** are both biased for the corresponding unconditional variances, but the HC-variance estimators provide simple fixes.

The ***ANCOVA*II** and ***cRM*** estimators of τ are equivalent because $\beta^{(6)}_2 + \beta^{(6)}_3 = \frac{\rho_0\sigma_0\sigma_{01}}{\sigma^2_0}$ and $\beta^{(6)}_2 = \frac{\rho_1\sigma_0\sigma_{11}}{\sigma^2_0}$ (Table 2). Two methods only differ in the way two estimators are estimated. ***ANCOVA*II** plugs in the OLS estimators $\hat{\beta}^{(6)}_{2,ols}$ and $\hat{\beta}^{(6)}_{3,ols}$, whereas ***cRM*** plugs in the REML estimators of the variance and covariance parameters. The numerical difference between the ***ANCOVA*II** and ***cRM*** estimators becomes smaller as sample size increases. As discussed previously, standard statistical softwares such as SAS does not output unconditional variance for ***ANCOVA*II** directly but the usual OLS model-based standard errors and *p*-values are biased for unconditional inference in heterogeneous scenario. The adjusted HC-variance estimator fixes this bias. Corrected ***ANCOVA*II** provides valid unconditional inference and performs comparably to ***cRM*.** Another alternative approach to estimate variances of the ***ANCOVA*I and II** estimators is to use bootstrap method [20].

Data example

No human data was used in this study. Instead we simulated three weight loss trial data sets based on a published study for three scenarios: homogeneous data, heterogeneous data with balanced and unbalanced designs as follows [21]:

1) The baseline weights for the control and treatment arms were generated from normal distribution with mean 88 kg and standard deviation 14 kg. Weights at 6 month after treatment for the control arm have mean 86 kg and standard deviation 15 kg. This gives a ~ 2.3% change from baseline. The mean and standard deviation of body weight at the sixth month in the treatment arm are 83 kg and 15 kg, respectively; This corresponds to a 5.7% change from baseline.
2) In the homogeneous data, the correlation coefficient between the pre- and post-treatment weights is 0.9. One hundred eighty subjects were assigned to the treatment and control arms equally. In the heterogeneous data, the correlation coefficient between the pre- and post-treatment weights in the control arm is 0.9 and 0.7 in the treatment arm. Sample sizes are ($n_0 = 90$, $n_1 = 90$) for the balanced design and ($n_0 = 60$, $n_1 = 120$) for the unbalanced design. We analyzed the data examples using the methods outlined in section Methods. The statistical results were reported in Table 3 (SAS programs are provided in the Additional file 1).

In the first data example, ***ANOVA*** produced the largest standard error and the largest *p*-value. ***ANOVA-Change*** and ***RM*** both outperformed ***ANOVA*** with much smaller standard errors and *p*-values. ***ANCOVA*I** and ***cRM*** outperformed ***ANOVA-Change*** and ***RM*** with smaller standard errors and *p*-values. Although ***ANCOVA*I** and ***cRM*** are equivalent when sample size is large, there are still minor numerical differences between the two in finite sample.

For the second data example with a balanced design, Fig. 2a shows that there is a strong baseline weight by treatment interaction. Both ***ANCOVA*I** and **II** have heteroscedastic errors by treatment arm (Fig. 2b and c). As shown in Table 2, the OLS model-based standard error of ***ANCOVA*I** is very similar to its HC and bootstrap standard errors. Thus, heteroscedasticity does not bias the model-based standard error of ***ANCOVA*I**. Although ***ANCOVA*II** is robust against heteroscedasticity in the balanced design, the OLS model-based standard error of ***ANCOVA*II** (*s.e* = 1.333) is still not correct because OLS fails to consider the variability of estimating the overall mean baseline weight. The adjusted HC standard error for ***ANCOVA*II** is 1.402, which is closer to the model-

Table 3 Statistical analysis of the three simulated data examples

Scenario	Method	Estimate	Standard error	*p-value*
Homogeneous	***ANOVA***	−3.089	2.106	0.144
	***ANCOVA*I**	−2.422	0.955	0.0121
	ANOVA-Change	−2.354	0.971	0.0163
	RM	−2.354	0.971	0.0163
	cRM	−2.434	0.944	0.0108
Heterogeneous ($n_0 = 90$, $n_1 = 90$)	***ANCOVA*I**	−3.203	1.403[a]	0.0235
			1.397[b]	0.0231
			1.400[d]	n/a
	***ANCOVA*II**	−3.165	1.333[a]	0.0187
			1.402[c]	0.0252
			1.397[d]	n/a
	cRM	−3.203	1.405	0.0241
Heterogeneous ($n_0 = 60$, $n_1 = 120$)	***ANCOVA*I**	−3.416	1.415[a]	0.0167
			1.279[b]	0.0083
			1.281[d]	n/a
	***ANCOVA*II**	−3.399	1.376[a]	0.0145
			1.258[c]	0.0076
			1.260[d]	n/a
	cRM	−3.396	1.262	0.0078

[a]OLS regression model-based standard error
[b]HC standard error for ANCOVA I (main effect) model
[c]Modified HC standard error for ANCOVA II (interaction) model
[d]Bootstrapping standard error ($n = 5000$)

based and HC standard errors of ***ANCOVA*I**. The bootstrapping standard errors for ***ANCOVA*I** and **II** are close to their HC or adjusted HC standard errors, which suggests the HC and adjusted HC variances perform well in estimating the unconditional variances. The ***cRM*** estimate and its standard error are close to those from ***ANCOVA*I** and **II**.

For the third example with an unbalanced design, Fig. 2d also reveals a baseline weight by treatment interaction. Both ANCOVA models have heteroscedastic errors by treatment arm (Fig. 2e and f). The model-based standard errors of ***ANCOVA*I** and **II** are not valid. The model-based standard errors were larger than the HC standard errors and thus overestimated the true conditional variances. Compared with ***ANCOVA*I**, ***ANCOVA*II** has a smaller HC standard error (also smaller *p*-value) and thus is slightly more efficient. The adjusted HC standard error for ***ANCOVA*II** is very close to the model-based standard error for ***cRM***. The bootstrapping standard errors for ***ANCOVA*I** and **II** are very close to their HC or adjusted HC standard errors.

Discussion

In this study we compare the efficiency of six unbiased methods analyzing pre-post designs. We found ANCOVA and ***cRM*** are the equally most efficient methods compared with other alternatives in homogeneous and heterogeneous scenarios. In this study, we focus on the scenario in which randomization is properly performed and these competing methods all target the same causal quantity. In the scenarios where the treatment is not properly randomized or not randomized at all (e.g., in an observational study), the baseline score will not be balanced by design. In this case these competing methods may target different causal quantities. Debate over using change-score analysis (or RM) verse ANCOVA in the non-randomized setting, generally known as the lord's paradox, is a well-known example [22, 23].

The majority of previous studies has only examined homogeneous study population. In this setting, ***ANOVA*** is one of the least efficient approaches for analyzing pre-post designs because it does not utilize any baseline information. ***ANOVA-Change*** and ***RM*** incorporate the baseline score as part of outcome, whereas ***ANCOVA*I** includes the baseline score as a covariate. ***ANCOVA*I** outperforms ***ANOVA-Change*** and ***RM*** because ***ANCOVA*I** utilizes the assumption that the baseline scores are balanced between two arms in a randomized study. Thus, change score is a less efficient way to utilize the

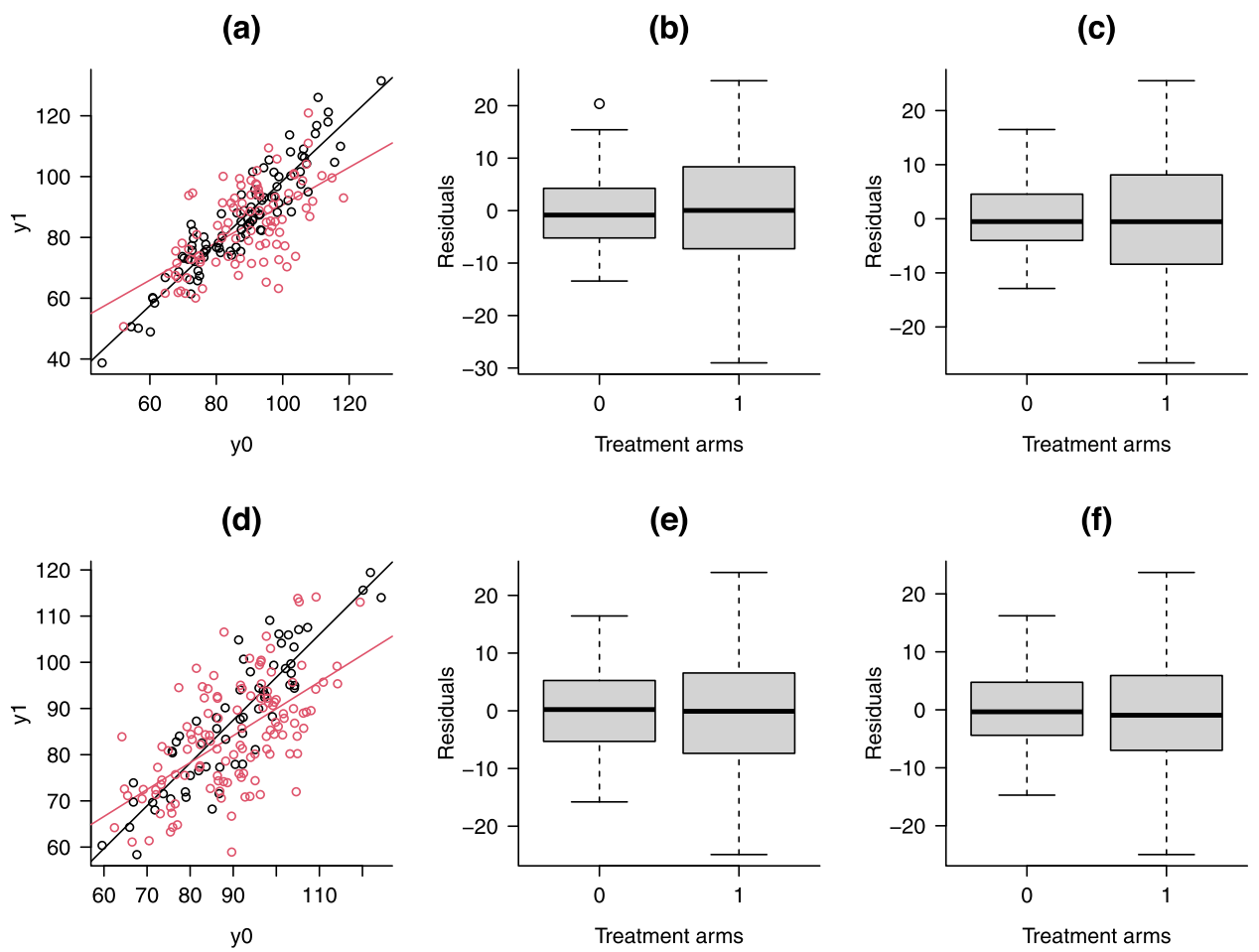

Fig. 2 Diagnosis plots of ANCOVA main and interaction models in heterogeneous scenario. **a** Scatter plot of baseline and follow-up weights in balanced design. Black and red solid dots are data points in the treatment and control arms. Black and red solid lines are the regression slopes of baseline weight against follow-up weight in the treatment and control arms. **b** Boxplot of residuals from the treatment and control arms from ***ANCOVA*I** model in balanced design; **c** Boxplot of residuals from the treatment and control arms from ***ANCOVA*II** model in balanced design; **d** Scatter plot of baseline and follow-up weights in unbalanced design. Black and red solid dots are data points in the treatment and control arms. Black and red solid lines are the regression slopes of baseline weight against follow-up weight in the treatment and control arms. **e** Boxplot of residuals from the treatment and control arms from ***ANCOVA*I** model in an unbalanced design; **f** Boxplot of residuals from the treatment and control arms from ***ANCOVA*II** model in an unbalanced design

baseline score than including the baseline score as a covariate. Since we seldom can control the values of the baseline score in randomized trials, the OLS assumption that the baseline score is fixed casts doubt on the validity of ANCOVA for hypothesis testing [6, 12]. Crager proved ***ANCOVA*I** is valid for unconditional inference in homogeneous scenario [6]. This conclusion can be simply attributed to that the conditional variance of the ***ANCOVA*I** estimator is an unbiased estimate for its unconditional variance [3].

A few studies investigated further a heterogeneous scenario [3, 4, 10, 12, 24]. Although the heterogeneity justifies the inclusion of the baseline measurement by treatment interaction term, ***ANCOVA*I** and **II** are both unbiased. Yang and Tsiatis showed that ***ANCOVA*II** has a smaller unconditional variance estimator than that of ***ANCOVA*I** unless in a balanced design [9]. However, the OLS model-based variances of the ***ANCOVA*I** and **II** estimators, reported by standard statistical softwares, are conditional variances, not unconditional variances. The OLS model-based standard errors and associated *p*-values for ***ANCOVA*II** are generally questionable for unconditional inference, and the model-based inference for ***ANCOVA*I** is biased only when the design is unbalanced [3, 4, 10, 24]. With the corrected HC variance estimators, both models provide valid unconditional inference. Choosing between ***ANCOVA*I** and **II** then becomes an evaluation of a trade-off between simplicity and some gains in efficiency.

In homogenous setting, ***cRM*** was suggested as a superior choice to ***ANCOVA*I** because the unconditional variance of the ***cRM*** estimator is smaller than the conditional variance of the ***ANCOVA*I** estimator [25]. Kenward et al. pointed out that such direct comparison between the conditional and unconditional variances is not meaningful. Since both estimators are equivalent, it can be shown that ***cRM*** coupled with REML and Kenward-roger adjustment performs almost identically to ***ANCOVA*I** in finite samples [17]. In heterogeneous scenario, ***cRM*** is comparable to ***ANCOVA*II** [3]. In presence of missing data, applied researchers often prefer ***cRM*** over ANCOVA because it can utilize all observed data but ANCOVA uses only complete cases**.** However**,** imputation methods which utilize the strong pre-post correlation, such as weighting and regression imputation, can improve the statistical power for ANCOVA without biasing estimates, making it comparable to ***cRM*** [17].

Furthermore, ANCOVA has several advantages over ***cRM***: first, outcome should only be the variable that can be influenced by treatment. Baseline measurement is certainly not an outcome by this definition. It is conceptually more appropriate to include the baseline score as covariate, not model it as outcome [5]; Second, it is very convenient to include other baseline variables in a regression model for more efficient estimates of treatment effect. Third, it is easy to adjust for other patterns of heteroscedastic errors in an OLS regression. For example, we may expect larger variability in the post-treatment weights associated with larger baseline weights. ***cRM*** cannot handle this more complex type of heteroscedasticity easily. HC-variance estimators for ANCOVA are simple fixes and readily implemented in statistical softwares.

Conclusion

Comparing with other alternative methods, ANCOVA is a simple and the most efficient approach analyzing a pre-post randomized design. When there exists a baseline score by treatment interaction, we need to assess the heteroscedasticity of ANCOVA particularly when the design is not balanced. The HC-variances should be used for valid inference when heteroscedasticity is present. Adding an interaction term in ANCOVA can gain some efficiency but not including this term does not bias results.

Abbreviations

ANOVA: Analysis of variance model; ***ANCOVA*I**: Analysis of covariance model adjusting for the baseline measurement; ***ANCOVA*II**: Analysis of covariance model adjusting for the baseline measurement by treatment interaction; *RM*: Constrained repeated measure model; ***cRM***: Constrained repeated measure model; HC: Heteroscadascity-consistent

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-021-01323-9.

Additional file 1.

Acknowledgements

Not applicable

Author's contributions

FW developed the idea for the paper, performed analysis, and drafted the manuscript. The author(s) read and approved the final manuscript.

Funding

Not applicable

Availability of data and materials

SAS code is provided as the Additional file 1. There is no real data used. All data generated or analyzed during this study are included in this published article [and its supplementary information files].

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The author declare that he has no competing risk

Received: 10 January 2021 Accepted: 19 May 2021
Published online: 24 July 2021

References

1. Vickers AJ. Analysis of variance is easily misapplied in the analysis of randomized trials: a critique and discussion of alternative statistical approaches. Psychosom Med. 2005;67(4):652–5. https://doi.org/10.1097/01.psy.0000172624.52957.a8.
2. O'Connell NS, Dai L, Jiang Y, Speiser JL, Ward R, Wei W, et al. Methods for analysis of pre-post data in clinical research: a comparison of five common methods. J Biom Biostat. 2017;8(1):1–8. https://doi.org/10.4172/2155-6180.1000334.
3. Wan F. Analyzing pre-post randomized studies with one post-randomization score using repeated measures and ANCOVA models. Stat Methods Med Res. 2019;28(10-11):2952–74. https://doi.org/10.1177/0962280218789972.
4. Wan F. Analyzing pre-post designs using the analysis of covariance models with and without the interaction term in a heterogeneous study population. Stat Methods Med Res. 2020;29(1):189–204. https://doi.org/10.1177/0962280219827971.
5. Senn S. Change from baseline and analysis of covariance revisited. Stat Med. 2006;25(24):4334–44. https://doi.org/10.1002/sim.2682.
6. Crager MR. Analysis of covariance in parallel-group clinical trials with pretreatment baseline. Biometrics. 1987;43(4):895–901. https://doi.org/10.2307/2531543.
7. Frison L, Pocock SJ. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. Stat Med. 1992;11(13):1685–704. https://doi.org/10.1002/sim.4780111304.
8. Brogan DR, Kutner MH. Comparative analyses of pretest-posttest research designs. Am Stat. 1980;34:229–32.
9. Yang L, Tsiatis AA. Efficiency study of estimators for a treatment effect in a pretest-posttest trial. Am Stat. 2001;55(4):314–21. https://doi.org/10.1198/000313001753272466.
10. Winkens B, van Breukelen GJ, Schouten HJ, Berger MP. Randomized clinical trials with a pre- and a post-treatment measurement: repeated measures versus ANCOVA models. Contemp Clin Trials. 2007;28(6):713–9. https://doi.org/10.1016/j.cct.2007.04.002.
11. Dimitrov DM, Rumrill J, Phillip D. Pretest-posttest designs and measurement of change. Work. 2003;20:159–65.

12. Chen X. The adjustment of random baseline measurements in treatment effect estimation. J Stat Plan Inference. 2006;136(12):4161–75. https://doi.org/10.1016/j.jspi.2005.08.046.
13. Jennings E. Models for pretest-posttest data: repeated measures ANOVA revisited. J Educ Behav Stat. 1988;13(3):273–80. https://doi.org/10.3102/10769986013003273.
14. Liang K, Zeger S. Longitudinal data analysis of continuous and discrete responses for pre-post designs. Sankhya. 2000;62:134–48.
15. Donald AB, Gregory DA. Symmetrized percent change for treatment comparisons. Am Stat. 2006;60:27–31.
16. Cole TJ, Altman DG. Statistics notes: what is a percentage difference? BMJ. 2017;358:j3663. https://doi.org/10.1136/bmj.j3663.
17. Kenward MG, White IR, Carpenter JR. Re: should baseline be a covariate or dependent variable in analyses of change from baseline in clinical trials? (Liu GF et al., Stat Med 2009; 28: 2509–30). Stat Med. 2010;29(13):1455–6. https://doi.org/10.1002/sim.3868.
18. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics. 1997;53(3):983–97. https://doi.org/10.2307/2533558.
19. Long J, Ervin L. Using heteroscedasticity: consistent standard errors in the linear regression model. Am Stat. 2000;54:217–24.
20. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall; 1993. https://doi.org/10.1007/978-1-4899-4541-9.
21. Boozer CN, Daly PA, Homel P, et al. Herbal ephedra/caffeine for weight loss: a 6-month randomized safety and efficacy trial. Int J Obes Relat Metab Disord. 2002;6:593–604.
22. Lord FM. A paradox in the interpretation of group comparisons. Psychol Bull. 1967;68(5):304–5. https://doi.org/10.1037/h0025105.
23. Egbewale BE, Lewis M, Sim J. Bias, precision and statistical power of analysis of covariance in the analysis of randomized trials with baseline imbalance: a simulation study. BMC Med Res Methodol. 2014;14(1):49. https://doi.org/10.1186/1471-2288-14-49.
24. Senn S. Various varying variances: the challenge of nuisance parameters to the practicing biostatistician. Stat Methods Med Res. 2015;24(4):403–19. https://doi.org/10.1177/0962280214520728.
25. Liu GF, Lu K, Mogg R, Mallick M, Mehrotra DV. Should baseline be a covariate or dependent variable in analyses of change from baseline in clinical trials? Stat Med. 2009;28(20):2509–30. https://doi.org/10.1002/sim.3639.

Publisher's Note