METHODOLOGY

Open Access

# Covariate-constrained randomization in cluster randomized 2 × 2 factorial trials: application to a diabetes prevention study

Juned Siddique[1*], Zhehui Li[1] and Matthew J. O'Brien[2]

## Abstract

**Background** Cluster randomized trials (CRTs) are randomized trials where randomization takes place at an administrative level (e.g., hospitals, clinics, or schools) rather than at the individual level. When the number of available clusters is small, researchers may not be able to rely on simple randomization to achieve balance on cluster-level covariates across treatment conditions. If these cluster-level covariates are predictive of the outcome, covariate imbalance may distort treatment effects, threaten internal validity, lead to a loss of power, and increase the variability of treatment effects. Covariate-constrained randomization (CR) is a randomization strategy designed to reduce the risk of imbalance in cluster-level covariates when performing a CRT. Existing methods for CR have been developed and evaluated for two- and multi-arm CRTs but not for factorial CRTs.

**Methods** Motivated by the BEGIN study—a CRT for weight loss among patients with pre-diabetes—we develop methods for performing CR in 2 × 2 factorial cluster randomized trials with a continuous outcome and continuous cluster-level covariates. We apply our methods to the BEGIN study and use simulation to assess the performance of CR versus simple randomization for estimating treatment effects by varying the number of clusters, the degree to which clusters are associated with the outcome, the distribution of cluster level covariates, the size of the constrained randomization space, and analysis strategies.

**Results** Compared to simple randomization of clusters, CR in the factorial setting is effective at achieving balance across cluster-level covariates between treatment conditions and provides more precise inferences. When cluster-level covariates are included in the analyses model, CR also results in greater power to detect treatment effects, but power is low compared to unadjusted analyses when the number of clusters is small.

**Conclusions** CR should be used instead of simple randomization when performing factorial CRTs to avoid highly imbalanced designs and to obtain more precise inferences. Except when there are a small number of clusters, cluster-level covariates should be included in the analysis model to increase power and maintain coverage and type 1 error rates at their nominal levels.

**Keywords** CRT, Balance, Confounding

## Background

Cluster randomized trials (CRTs) are randomized controlled trials where randomization takes place at an administrative level (e.g., hospitals, clinics, or schools) rather than at the individual level. CRTs are an attractive research design when there are concerns of treatment contamination among participants, when it is logistically

*Correspondence:
Juned Siddique
siddique@northwestern.edu
[1] Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, 680 N Lake Shore Drive, Suite 1400, Chicago, IL, USA
[2] Department of Medicine, Northwestern University Feinberg School of Medicine, 750 N Lake Shore Drive, 10th floor, Chicago, IL, USA

easier to conduct the trial by randomizing at the cluster level, and when the intervention of interest is delivered at the cluster level [1].

A major practical limitation when conducting CRTs is the ability to enroll a large number of clusters. When the number of available clusters is small, researchers may not be able to rely on simple randomization to achieve balance on cluster-level covariates across treatment conditions [2]. If these cluster-level covariates are predictive of the outcome, covariate imbalance across treatment conditions may distort treatment effects, threaten internal validity, lead to a loss of power, increase the variability of the treatment effect, and usually requires statistical adjustment in the analysis stage [3]. For example, in a two-arm CRT where clinics are randomized to treatment conditions and where the size of a clinic is related to the outcome of interest, researchers would want equal numbers of small and large clinics in the treatment and control conditions, respectively.

Factorial experiments are an efficient approach to determine which of several possible components of a proposed intervention have effects of practical significance [4]. When implementing factorial experiments at the cluster level, the challenges involved in balancing cluster-level covariates across arms is magnified because there are more than two treatment conditions. For example, in a 2 × 2 factorial CRT, clusters will be randomized to one of 4 treatment conditions.

One approach to address imbalance in prognostic cluster-level covariates across treatment conditions is to include these covariates in the analysis model which can help ensure an unbiased estimate of the treatment effect. The drawback to including cluster-level covariates in the analysis model is the subsequent loss of degrees of freedom that are available to estimate treatment effects. This resulting loss of power can be substantial when there are a small number of clusters [5]. An alternative to model-based covariate adjustment is to control for potential confounders at the *design* stage, by balancing the distribution of select measured characteristics across treatment arms. This can help ensure more precise treatment effects as well as confidence that observed treatment effects are not due to imbalance in prognostic covariates while at the same time avoiding the resulting loss of power due to covariate adjustment.

Individually randomized trials often rely on stratification to achieve balance on prognostic factors across treatment conditions. In CRTs with a small number of clusters, stratifying on more than one variable can be challenging because of an insufficient number of clusters to distribute among strata. This phenomenon is only exacerbated in factorial trials where there are at least 4 treatment conditions. For example, with two binary stratification variables there will be total of four strata. To conduct a 2 × 2 factorial CRT would require at least 4 clusters per stratum (16 clusters total) to avoid unequal allocation of treatments within strata [3]. Furthermore, stratifying on a continuous factor requires converting it to a categorical variable, a process that can result in a loss of information.

Covariate-constrained randomization (CR) is an alternative procedure for achieving balance across treatment conditions on a set of predetermined cluster-level covariates. Unlike individual level trials where participants are recruited sequentially, the participating units in a CRT are generally assembled at the start of the study so that cluster-level covariate values such as geographic location, clinic size, and the income level of patients are available at the design stage.

The first step in CR is to identify those cluster-level covariates that are predictive of the outcome on which one wishes to achieve balance. Using the terminology of Li et al. [6], we refer to these covariates as “potential confounders" because they are cluster-level prognostic factors that, when imbalanced, could distort estimates of treatment effects.

The second step in CR is—for every possible randomization scheme (or a random subset of schemes when the number of clusters is large)—to calculate a balance score that measures the difference in the distribution of cluster-level covariates across treatment conditions [3, 7]. Next, a subset of schemes is chosen that meet some pre-specified balance criteria, such the 10% of schemes with the best balance scores. Finally, an allocation is randomly selected among those schemes that meet the pre-specified criteria and is used to randomize clusters. CR tends to produce better balance on average across treatment conditions as compared to simple randomization in which a randomization scheme is selected from all possible schemes with equal probability assigned to each scheme. Compared with stratification, CR may be preferred due to its capacity to accommodate multiple covariates, both categorical and continuous [8].

There are numerous variations of CR that use different balance metrics and different analysis strategies. In the two-arm setting, Raab and Butcher [7] and Li et al. [6] consider weighted and unweighted pairwise balance scores based on the difference in covariate means between arms. In the multi-arm setting, Zhou et al. [9] extend the pairwise balance score method, while Watson et al. [10] present a balance metric based on the sum of cluster-level mean differences. Ciolino et al. [11] calculate a Kruskal-Wallis test for each covariate across arms and assesses balance based on the $p$-values of these tests where a minimum $p$-value greater than 0.30 was found to appropriately identify acceptable balance.

The tradeoffs involved in selecting the size of the constrained allocation space is an important consideration when using CR. As noted by Moulton [12], a highly constrained design—while ensuring good balance across treatment conditions—may open the investigator to accusations of manipulation in favor of their hypothesis. Furthermore, there can be a departure of the nominal type 1 error rate when correlated clusters have a high or low probability of being included in the same arm [6, 12, 13].

In a two-arm setting, Li et al. [6] found that type 1 error rates were conservative when using CR if not all the covariates used for randomization were included in the analysis. Watson et al. [10], and Zhou et al. [9] found the same result in simulations of multi-arm trials. For these reasons, Li et al. [6], Watson et al. [10], and Zhou et al. [9] all recommend adjustment for potential confounders in the analyses stage to maintain type 1 error and provide adequate power. The most common approach for the analysis of CRTs is mixed-effects regression modeling with random cluster-level effects to account for within-cluster correlation. In a longitudinal CRT, mixed-effects models are sufficiently flexible to account for variability at both the cluster and participant levels.

Existing work on CR has focused on two- or three-arm CRTs. The performance of CR in a factorial setting—where the minimum number of randomization conditions is 4—has not been explored. At the randomization stage, a factorial design and a multi-arm parallel design are similar. For example, a 2 × 2 factorial trial and a four-arm trial both aim to enroll equal numbers of clusters to one of four conditions. For this reason, CR methods for a 2 × 2 factorial design and a 4-arm parallel design are the same, as they both seek to achieve balance across conditions on prognostic cluster-level covariates. However, the analysis of data from factorial designs and multi-arm parallel designs is different. In a four-arm trial, each of the 4 arms consists of a single intervention and analysis involves comparing mean outcomes in each of the arms to each other or each of the three arms to a control arm.

In a 2 × 2 factorial trial, main effects are estimated by combining the mean outcomes from two conditions and comparing them to the mean outcomes from the other two conditions. For example, in Table 1 below describing the design of our motivating example, the effect of the in-person intervention is obtained by estimating the mean outcome in conditions "a" and "c" and subtracting it from the mean outcome in conditions "b" and "d." Similarly, the effect of text messages is estimated by taking the mean outcome in conditions "b" and "c" and subtracting it from the mean outcome in conditions "a" and "d." In this way, factorial designs are able to test multiple intervention components efficiently, by recycling [14] clusters when estimating intervention effects and their interactions. This is especially important for CRTs where recruiting clusters can be challenging. Assessing whether CR operates differently in the factorial setting is an area that requires further investigation.

**Table 1** 2x2 factorial design of the BEGIN Study

| | Intervention | |
|---|---|---|
| **Condition** | **In-person** | **Text messages** |
| a | On | Off |
| b | Off | On |
| c | On | On |
| d | Off | Off |

### Motivating example

Our methods are motivated by the Behavioral Nudges for Diabetes Prevention (BEGIN) study [15], a 2 × 2 factorial CRT studying two pragmatic behavioral interventions that prompt patients to adopt evidence-based treatment for prediabetes in primary care, thereby promoting modest weight loss. Preventing type 2 diabetes (T2D) has become a top public health priority given the high prevalence of prediabetes and the availability of evidence-based treatments to prevent T2D [16, 17]. With 682 million office visits made by US adults annually, primary care is a critical venue for promoting weight loss and T2D prevention [18].

BEGIN takes place at the Erie Family Health Center, a Federally-funded primary care clinic network in Chicago serving 85,000 vulnerable patients, 83% of whom live in poverty and 79% of whom are Hispanic/Latino. Given their reach and unique access to high-risk populations, community health centers are an ideal venue for studying primary care-based interventions that promote prediabetes treatment uptake and modest weight loss.

The two BEGIN primary care interventions are (1) in-person behavioral nudges via a pre-diabetes decision aid delivered by existing health educators; and (2) automated behavioral nudges via motivational letters and text messages. These two interventions are being tested in 8 Erie Family Health Center clinics using a 2 × 2 factorial design. Two clinics are randomly assigned to each of the the four conditions in Table 1. These four conditions are:

a. In-person intervention alone
b. Text message intervention alone
c. Both in-person and text message interventions
d. Neither intervention

Because randomization occurs at the clinic level, there is a risk of imbalance in clinic-level characteristics across treatment conditions. Table 2 presents data on three

**Table 2** Clinic volume, percent female, and mean BMI of visits by patients who met the BEGIN eligibility criteria in 2019–2020 for each of the 8 clinics in the BEGIN trial

| Clinic number | Clinic volume | Percent female | Mean BMI |
|---|---|---|---|
| C1 | 29,933 | 73.57 | 31.19 |
| C2 | 26,613 | 88.54 | 31.20 |
| C3 | 23,940 | 77.59 | 31.53 |
| C4 | 18,869 | 77.52 | 30.55 |
| C5 | 14,660 | 84.65 | 30.32 |
| C6 | 24,119 | 81.71 | 31.11 |
| C7 | 34,637 | 74.39 | 30.58 |
| C8 | 3429 | 71.19 | 31.33 |

clinic-level covariates from the 8 clinics in the BEGIN study on which the BEGIN investigators sought to achieve balance. The data in Table 2 are based on clinic visits in 2019–2020 (prior to the start of the BEGIN study) among patients who met the eligibility criteria of the BEGIN study. These three potential confounders are (1) clinic volume, as measured by the number of office visits; (2) percent of office visits by female patients; and (3) mean BMI of visits. It is worth noting that mean BMI is similar across the 8 clinics, but total volume varies considerably.

In this manuscript, motivated by the BEGIN study, we extend and evaluate CR methods for multi-arm trials [9–11]—with a continuous outcome and continuous cluster-level covariates—to the 2 × 2 factorial CRT setting. The outline for the rest of this paper is as follows. In the Methods section, we present methods for CR in the setting of a 2 × 2 factorial CRT and describe a simulation study to assess the performance of our methods as compared to simple randomization of clusters. In the Results section, we present the results of our simulation study and apply our methods to the BEGIN study. The Discussion section provides discussion and areas of future work. We conclude in the Conclusions section.

## Methods

As mentioned above, once a set of potential cluster-level confounders are identified, the next step in performing CR is to calculate a balance metric to measure the difference in the distribution of these cluster-level covariates across treatment conditions for all possible randomization schemes. In this section, we describe a balance metric for factorial trials that extends the balance metrics of Li et al. [6], Raab and Butcher [7], and Watson et al. [10].

Let $J$ be the number of clusters and $T$ be the number of treatment conditions so that $n_T = \frac{J}{T}$ clusters are randomized to each treatment condition. Let $x_{jk}$ be the value of the $k$th covariate ($k = 1, \ldots, K$) in cluster $j$ ($j = 1, \ldots, J$), and $\bar{x}_{tk} = \frac{1}{n_T} \sum_{j \in t} x_{jk}$ the mean value of the $k$th covariate in clusters assigned to condition $t$, ($t = 1, \ldots, T$). Finally $\bar{x}_k = \frac{1}{J} \sum_{j=1}^{J} x_{jk}$ is the overall mean of covariate $k$ across all clusters. Our balance metric is:

$$B = \sum_{k=1}^{K} d_k \sum_{t=1}^{T} (\bar{x}_{tk} - \bar{x}_k)^2 \tag{1}$$

where $d_k$ is a predetermined scaling factor for the $k$th covariate. Following Raab and Butcher [7] and Li et al. [6], we set $d_k$ as the inverse of the variance of the $k$th covariate across all clusters. That is

$$d_k = \frac{1}{s_k^2} = \frac{J-1}{\sum_{j=1}^{J} (x_{jk} - \bar{x}_k)^2}. \tag{2}$$

The metric in (1) and (2) describe the balance score introduced by Watson et al. [10] for use in multi-arm trials. A limitation to this metric is that balance is purely defined by covariate values and does not take into account clinical importance. For example, in the BEGIN study, if clinic volume is considered to be a stronger predictor of weight loss than percent of female visits, we may want to give clinic volume greater weight in the balance metric so that smaller balance scores using the weighted metric will reflect better balance on clinic volume at the expense of less balance on clinic percent female. To incorporate weights into the balance metric in (1), we use the approach of Yu et al. [8] to produce the weighted balance metric:

$$B_w = \sum_{k=1}^{K} w_k d_k \sum_{t=1}^{T} (\bar{x}_{tk} - \bar{x}_k)^2 \tag{3}$$

where $w_k$ is a user-defined weight for the $k$th covariate. If $w_k = 1$ for all covariates, then (3) reduces to the balance metric in (1). When researchers consider certain variables to be more predictive of the outcome than others or for which there is greater variability across clusters, a user-defined weight $w_k > 1$ could be assigned to those variables when calculating balance scores [6].

To perform CR, the balance metric $B$ (or $B_w$) is generated for all possible randomization schemes of the $J$ clusters. The final allocation is chosen from a subset of allocations that meet a pre-specified balance criteria. Here, we select a cutoff value $q$ which is the $q$th percentile of the balance scores. Yu et al. [8] note that the cutoff value $q$ should be small and away from 1.0 (simple randomization) so that only the more balanced randomization schemes are retained in the constrained space. For example, Yu et al. [8] set $q = 0.1$ so that only the schemes in the top 10% of balance scores are included in the constrained allocation space.

When the number of clusters is small, it is feasible to calculate the balance score for all possible allocations where the number of allocations is $\frac{J!}{[(J/T)!]^T}$. For example, when $J = 8$ and $T = 4$, there are only 2520 possible ways to randomize clusters. Since treatment assignments can be labeled $4! = 24$ different ways, these 2520 possible allocations correspond to only $2520/24 = 105$ unique balance scores. Thus, when $J = 8$, it is computationally feasible to select the final allocation from the top $(q \times 100)\%$ of the allocations corresponding to the 105 unique balance scores. For example, when $q = 0.1$, we draw from the top $10 \times 24 = 240$ balanced allocations.

However, for CRTs with more clusters, for example, when $J = 12$ and $T = 4$, there are 369,600 possible ways to randomize the clusters and enumerating all possible allocations becomes computationally expensive. Following Li et al. [6], when $J > 8$, we randomly sample a subset of 20,000 allocations from all possible allocations, remove duplicate allocations, then select our final allocation from the top $(q \times 100)\%$ of allocations in terms of balance scores.

## Simulation study

We use simulation to assess our method of CR in the setting of a $2 \times 2$ factorial cluster randomized trial and how it compares to simple randomization in terms of estimating treatment effects. Following Li et al. [6] we simulate data using the following approach. Let $x_{j1}, x_{j2}, x_{j3}$ be three correlated cluster level covariates for cluster $j$, $(j = 1, \ldots, J)$; that are normally distributed with mean 1 and variance $\sigma_x^2$ on which we wish to achieve balance. The correlations between cluster-level covariates are based on the BEGIN data in Table 2 and are: $\text{corr}(x_{j1}, x_{j2}) = 0.13$, $\text{corr}(x_{j1}, x_{j3}) = -.04$, and $\text{corr}(x_{j2}, x_{j3}) = -0.19$. Let $y_{ij}$ be the outcome of interest for subject $i$, $(i = 1, \ldots, n_j)$; in cluster $j$. We set $n_j = 100$ throughout. Let $\text{Trt}_{1j}$ and $\text{Trt}_{2j}$ indicate—using dummy coding—whether cluster $j$ is assigned to treatments 1 and/or 2, respectively, where treatment is based on the factorial design in Table 1. We generate $y_{ij}$ from the following linear mixed-effects model:

$$y_{ij} = \beta_1 x_{j1} + \beta_2 x_{j2} + \beta_3 x_{j3} + \gamma_1 \text{Trt}_{1j} + \gamma_2 \text{Trt}_{2j} + \text{b}_\text{j} + \varepsilon_{\text{ij}} \quad (4)$$

The parameters $\beta_1$, $\beta_2$, and $\beta_3$ are regression coefficients on the cluster-level covariates that are predictive of the outcome (when $\beta \neq 0$). For simplicity, we let $\beta_1 = \beta_2 = \beta_3$. The coefficients $\gamma_1$ and $\gamma_2$ correspond to the effects of the two interventions. We set $\gamma_1 = 5$ and $\gamma_2 = 0$. The parameter $b_j$ is a cluster-level random effect where $b_j \sim N\left(0, \sigma_b^2\right)$ and $\varepsilon_{ij}$ is an error term where $\varepsilon_{ij} \sim N\left(0, \sigma_\varepsilon^2\right)$. We assume $\sigma_\varepsilon^2 = 36$ and an intra-cluster correlation (ICC) of $\rho = 0.05$ so that $\sigma_b^2 = \rho\sigma_\varepsilon^2/(1-\rho)$.

When controlling for cluster-level covariates in the analysis model, the analysis model is identical to (4) and the variance of the outcome is the same across all simulation scenarios and is equal to $\text{Var}(\text{y}_{\text{ij}}|\mathbf{x}_\text{j}) = \sigma_\varepsilon^2 + \sigma_\text{b}^2$.

When the analysis model does not control for cluster-level covariates, the covariates $x_{j1}, x_{j2}, x_{j3}$ are excluded from the model and the variance of the outcome varies across simulation scenarios and is reflected in an inflated between-cluster variance. That is,

$$\begin{aligned} \text{Var}(\text{y}_{\text{ij}}) &= E\left\{\text{Var}(\text{y}_{\text{ij}}|\mathbf{x}_\text{j})\right\} + \text{Var}\left\{\text{E}(\text{y}_{\text{ij}}|\mathbf{x}_\text{j})\right\} \\ &= \sigma_\varepsilon^2 + \sigma_b^2 + 3\beta^2\sigma_x^2, \end{aligned} \quad (5)$$

where the term $3\beta^2\sigma_x^2$ is the increase in variance due to not conditioning on covariates. Since $\sigma_\varepsilon^2$ and $\sigma_b^2$ are fixed in our simulations, the variance of the outcome will be the same when the *product* of $\beta^2$ and $\sigma_x^2$ are the same.

We sought to investigate the following factors in our simulation study and examine how their effects differ when using CR as compared to simple randomization: number of clusters, the size of the constrained randomization space, the variability of cluster-level covariates, the magnitude of cluster-level effects on the outcome, and whether or not cluster-level covariates are controlled for in the analysis model. Table 3 shows the factors that vary in the simulation. With five factors with two to four levels each, we evaluated a total of $2 \times 4 \times 3 \times 3 \times 2 = 144$ scenarios. Simulation is based on the following steps:

1. Simulate $K = 3$ correlated cluster level covariates of size $J$ from a multivariate normal distribution with mean 1, variance $\sigma_x^2$, and correlations $\text{corr}(x_{j1}, x_{j2}) = 0.13$, $\text{corr}(x_{j1}, x_{j3}) = -.04$, and $\text{corr}(x_{j2}, x_{j3}) = -0.19$.
2. Use either CR (see code in Appendix 1 for implementing CR in R) or simple randomization to randomize the J clusters to one of the 4 conditions in Table 1.
3. Draw $\varepsilon_{ij} \sim N\left(0, \sigma_\varepsilon^2\right)$, $i = 1, \ldots, n_j$; $j = 1, \ldots, J$. Here, we fix $\sigma_\varepsilon{}^2 = 36$.
4. Draw $b_j \sim N\left(0, \sigma_b^2\right)$, $j = 1, \ldots, J$ where $\sigma_b^2 = \rho\sigma_\varepsilon^2/(1-\rho)$, and $\rho = 0.05$ is the ICC.
5. Generate $n_j = 100$ values of $y_{ij}$ using (4).
6. Analyze the data using a linear mixed-effects model with a random intercept for cluster and indicator variables for the two treatment conditions. Based on the

**Table 3** Factors that vary in the simulation study

| Factor | Values |
|---|---|
| Number of clusters $J$ | $J = 8, 12$ |
| Size of constrained randomization space | Top 10%, 20%, 50%, 100% (SR) of balance scores |
| SD of cluster-level covariates $\sigma_x$ | 0.5, 1, 2 |
| Cluster-level covariate effects $\beta$ | None ($\beta = 0$), medium ($\beta = 0.5$), large ($\beta = 1$) |
| Analysis model | Control/do not control for covariates |

*Note: SR* simple randomization, *SD* standard deviation simulation scenario, the analysis model either controls for or does not control for cluster-level covariates.

Steps 1–6 were performed 10,000 times to generate 10,000 parameter estimates for each of the 144 simulation scenarios. We focus our attention on the performance of the treatment effects $\gamma$. Specifically, using $\gamma_1$ we assess the percent bias, variance, mean squared error (MSE), coverage and width of the 95% confidence interval, and the power to reject the null hypothesis. Using $\gamma_2$, we assess type 1 error under a nominal type 1 error rate of 0.05. For each scenario, we also report the mean, minimum, and maximum of the balance metric in (1) across the 10,000 simulations.

### *Interaction effects*

The primary reason for using a factorial design in the BEGIN trial was efficiency, as it requires a smaller sample size than a three-arm trial. However, an additional advantage of a factorial design is the ability to estimate whether treatments interact. To assess our method of CR when estimating an interaction effect, we repeated our simulations but now replacing the data generating model in (4) with the following model:

$$y_{ij} = \beta_1 x_{j1} + \beta_2 x_{j2} + \beta_3 x_{j3} + \gamma_1 \mathrm{Trt}_{1j} + \gamma_2 \mathrm{Trt}_{2j} + \gamma_3 (\mathrm{Trt}_{1j} * \mathrm{Trt}_{2j}) + \mathrm{b_j} + \varepsilon_{\mathrm{ij}}. \tag{6}$$

Equation (6) differs from (4) in two respects. First, an interaction term has been included in the model for the interaction between the two interventions. Second, the variables $\mathrm{Trt}_{1j}$ and $\mathrm{Trt}_{1j}$ are entered using effect coding (1, − 1) so that the main effects and interaction term are orthogonal to each other [19]. Since the low and high levels of these effect codes are further apart than the dummy codes in (4), we set $\gamma_1 = 2.5$ so that the magnitude of the main effect of treatment 1 is the same as before. We again set $\gamma_2 = 0$ to assess type 1 error in the presence of an interaction effect. Finally, we set $\gamma_3 = 2.5$ so that the magnitude of the interaction effect was the same as that of the main effect.

We again performed Steps 1–6 10,000 times to generate 10,000 parameter estimates for each of the 144 simulation scenarios in Table 3. We report the performance characteristics for $\gamma_1$ (the main effect of treatment 1) and $\gamma_3$ (the interaction of treatments 1 and 2). Using $\gamma_2$ (the main effect of treatment 2), we assess type 1 error under a nominal type 1 error rate of 0.05.

## Results

### Simulation results

Tables 4 and 5 summarize the results of our simulation study for 8 and 12 clinics, respectively, using both CR and simple randomization under various degrees of cluster-level variability ($\sigma_x$), cluster-level covariate effects ($\beta$), and allocation space sizes. The results in Tables 4 and 5 are from simulations where cluster-level covariates are not controlled for in the analysis model and only main effects are estimated.

Looking at Table 4, comparing CR to simple randomization, the percent bias is essentially 0 for both CR and simple randomization. As the magnitude of cluster-level covariate effects increases (as measured by $\beta$) variance and MSE increase, with both performance criteria better under CR. A similar trend is seen with increasing values of cluster-level variability (as measured by $\sigma_x$), where variance and MSE increase as $\sigma_x$ increases and both performance criteria are worse under simple randomization. Coverage and type 1 error tend to be conservative under CR while these values are at their nominal levels under simple randomization.

Power in Table 4 is similar for both CR and simple randomization. However, in those settings where the magnitude of potential confounding is high and cluster-level variability is also high, power is low for both CR and simple randomization. For example, when $\sigma_x = 2$ and $\beta = 1$, power is 26% under CR (top 10% of balance scores) and 35% under simple randomization.

As the allocation space increases from 10% to 50%, the simulation results are more like those under simple randomization. Variance and MSE increase while coverage and type 1 error are almost identical when using the top 10% or 20% of balance scores and slightly less conservative when using the top 50% of balance scores. As expected, covariate balance (last column in Table 4) is better and less variable as the allocation space becomes more constrained. Because the balance metric in (1) standardizes each covariate by the inverse of its variance, values of $\sigma_x$ do not have an effect on the balance metric and mean balance and its range across the 10,000 simulations only depends on the size of the allocation space.

The results in Table 5 based on 12 clusters are similar to those based on 8 clinics, with better variance and MSE under CR and similar power as compared to simple randomization. Again, coverage and type 1 error are conservative under CR while these criteria are at their nominal level under simple randomization. However, with 12 clusters, power is much greater than in the setting with 8 clusters such that power is only inadequate in the scenario with the highest potential confounding ($\beta = 1$) and the highest between-cluster variability ($\sigma_x = 2$).

As noted in the Simulation Study section, the performance criteria in Tables 4 and 5 are the same when the

**Table 4** Simulation results for the effect of treatment with 8 clusters, based on a main effects only analysis model that does not control for cluster-level covariates

| Covariate SD | Degree Confounding | %Bias | Var | MSE | Cov | Power | 95% CI Width | Type 1 Error | Mean balance (min, max) |
|---|---|---|---|---|---|---|---|---|---|
| *Covariate-constrained randomization: top 10% of balance scores* | | | | | | | | | |
| $\sigma_x = 0.5$ | $\beta = 0.0$ | 0.07 | 1.15 | 1.15 | 0.95 | 0.96 | 5.18 | 0.05 | 2.13 (0.35, 3.58) |
| | $\beta = 0.5$ | 0.00 | 1.19 | 1.19 | 0.95 | 0.95 | 5.42 | 0.05 | 2.13 (0.35, 3.58) |
| | $\beta = 1.0$ | − 0.06 | 1.32 | 1.32 | 0.96 | 0.90 | 6.08 | 0.04 | 2.13 (0.35, 3.58) |
| $\sigma_x = 1$ | $\beta = 0.0$ | 0.08 | 1.15 | 1.15 | 0.95 | 0.96 | 5.18 | 0.05 | 2.13 (0.35, 3.58) |
| | $\beta = 0.5$ | − 0.06 | 1.32 | 1.32 | 0.96 | 0.90 | 6.08 | 0.04 | 2.13 (0.35, 3.58) |
| | $\beta = 1.0$ | − 0.19 | 1.84 | 1.84 | 0.98 | 0.69 | 8.21 | 0.03 | 2.13 (0.35, 3.58) |
| $\sigma_x = 2$ | $\beta = 0.0$ | 0.07 | 1.15 | 1.15 | 0.95 | 0.96 | 5.18 | 0.05 | 2.13 (0.35, 3.58) |
| | $\beta = 0.5$ | − 0.19 | 1.84 | 1.84 | 0.98 | 0.69 | 8.21 | 0.03 | 2.13 (0.35, 3.58) |
| | $\beta = 1.0$ | − 0.45 | 3.92 | 3.92 | 0.99 | 0.26 | 13.77 | 0.01 | 2.13 (0.35, 3.58) |
| *Covariate-constrained randomization: top 20% of balance scores* | | | | | | | | | |
| $\sigma_x = 0.5$ | $\beta = 0.0$ | − 0.27 | 1.13 | 1.13 | 0.95 | 0.96 | 5.19 | 0.05 | 2.56 (0.35, 3.79) |
| | $\beta = 0.5$ | − 0.29 | 1.18 | 1.18 | 0.96 | 0.95 | 5.42 | 0.05 | 2.56 (0.35, 3.79) |
| | $\beta = 1.0$ | − 0.31 | 1.33 | 1.33 | 0.96 | 0.90 | 6.07 | 0.04 | 2.56 (0.35, 3.79) |
| $\sigma_x = 1$ | $\beta = 0.0$ | − 0.27 | 1.13 | 1.13 | 0.95 | 0.96 | 5.19 | 0.05 | 2.56 (0.35, 3.79) |
| | $\beta = 0.5$ | − 0.31 | 1.33 | 1.33 | 0.96 | 0.90 | 6.07 | 0.04 | 2.56 (0.35, 3.79) |
| | $\beta = 1.0$ | − 0.36 | 1.95 | 1.95 | 0.98 | 0.69 | 8.15 | 0.03 | 2.56 (0.35, 3.79) |
| $\sigma_x = 2$ | $\beta = 0.0$ | − 0.27 | 1.13 | 1.13 | 0.95 | 0.96 | 5.19 | 0.05 | 2.56 (0.35, 3.79) |
| | $\beta = 0.5$ | − 0.34 | 1.94 | 1.95 | 0.98 | 0.69 | 8.15 | 0.03 | 2.56 (0.35, 3.79) |
| | $\beta = 1.0$ | − 0.42 | 4.39 | 4.39 | 0.98 | 0.27 | 13.59 | 0.02 | 2.56 (0.35, 3.79) |
| *Covariate-constrained randomization: top 50% of balance scores* | | | | | | | | | |
| $\sigma_x = 0.5$ | $\beta = 0.0$ | − 0.06 | 1.14 | 1.14 | 0.95 | 0.96 | 5.21 | 0.05 | 3.34 (0.50, 4.80) |
| | $\beta = 0.5$ | − 0.06 | 1.21 | 1.21 | 0.95 | 0.95 | 5.43 | 0.05 | 3.34 (0.50, 4.80) |
| | $\beta = 1.0$ | − 0.06 | 1.41 | 1.41 | 0.95 | 0.90 | 6.03 | 0.04 | 3.34 (0.50, 4.80) |
| $\sigma_x = 1$ | $\beta = 0.0$ | − 0.06 | 1.14 | 1.14 | 0.95 | 0.96 | 5.21 | 0.05 | 3.34 (0.50, 4.80) |
| | $\beta = 0.5$ | − 0.07 | 1.41 | 1.41 | 0.95 | 0.90 | 6.03 | 0.04 | 3.34 (0.50, 4.80) |
| | $\beta = 1.0$ | − 0.07 | 2.23 | 2.23 | 0.97 | 0.70 | 8.00 | 0.04 | 3.34 (0.50, 4.80) |
| $\sigma_x = 2$ | $\beta = 0.0$ | − 0.08 | 1.14 | 1.14 | 0.95 | 0.96 | 5.21 | 0.05 | 3.34 (0.50, 4.80) |
| | $\beta = 0.5$ | − 0.08 | 2.22 | 2.22 | 0.97 | 0.70 | 8.00 | 0.04 | 3.34 (0.50, 4.80) |
| | $\beta = 1.0$ | − 0.09 | 5.49 | 5.49 | 0.97 | 0.31 | 13.20 | 0.03 | 3.34 (0.50, 4.80) |
| *Simple randomization* | | | | | | | | | |
| $\sigma_x = 0.5$ | $\beta = 0.0$ | − 0.06 | 1.17 | 1.17 | 0.95 | 0.96 | 5.17 | 0.05 | 4.50 (0.49, 9.42) |
| | $\beta = 0.5$ | − 0.12 | 1.26 | 1.26 | 0.95 | 0.94 | 5.37 | 0.05 | 4.50 (0.49, 9.42) |
| | $\beta = 1.0$ | − 0.18 | 1.53 | 1.53 | 0.95 | 0.90 | 5.92 | 0.05 | 4.50 (0.49, 9.42) |
| $\sigma_x = 1$ | $\beta = 0.0$ | − 0.06 | 1.17 | 1.17 | 0.95 | 0.96 | 5.17 | 0.05 | 4.50 (0.49, 9.42) |
| | $\beta = 0.5$ | − 0.18 | 1.53 | 1.53 | 0.95 | 0.90 | 5.92 | 0.05 | 4.50 (0.49, 9.42) |
| | $\beta = 1.0$ | − 0.30 | 2.59 | 2.59 | 0.95 | 0.71 | 7.75 | 0.05 | 4.50 (0.49, 9.42) |
| $\sigma_x = 2$ | $\beta = 0.0$ | − 0.06 | 1.17 | 1.17 | 0.95 | 0.96 | 5.17 | 0.05 | 4.50 (0.49, 9.42) |
| | $\beta = 0.5$ | − 0.30 | 2.59 | 2.59 | 0.95 | 0.71 | 7.75 | 0.05 | 4.50 (0.49, 9.42) |
| | $\beta = 1.0$ | − 0.54 | 6.79 | 6.79 | 0.95 | 0.35 | 12.66 | 0.05 | 4.50 (0.49, 9.42) |

*Note: SD* standard deviation, *%Bias* percent bias, *Var* variance, *MSE* mean squared error, *Cov* coverage of the 95% confidence interval, *CI* confidence interval, *Type 1 Error* proportion of type 1 errors under a nominal type 1 error rate of 0.05, *Balance* covariate balance

*product* of $\sigma_x$ and $\beta$ are the same due to the fact that the marginal variance of $y_{ij}$ is a function of the product of $\sigma_x$ and $\beta$ (Eq. 5). For example, the performance criteria when $\sigma_x = 0.5$ and $\beta = 1.0$ are the same as when $\sigma_x = 1.0$ and $\beta = 0.5$.

Appendix 2 Tables 7 and 8 summarize the simulation results for 8 and 12 clusters, respectively, now based on a main effects only analysis model that controls for cluster-level covariates. Here, the analysis model is identical to the data generating model so that the

**Table 5** Simulation results for the effect of treatment with 12 clusters, based on a main effects only analysis model that does not control for cluster-level covariates

| **Covariate** | **Degree** | **%Bias** | **Var** | **MSE** | **Cov** | **Power** | **95% CI** | **Type 1** | **Mean balance** |
|---|---|---|---|---|---|---|---|---|---|
| **SD** | **Confounding** | | | | | | **Width** | **Error** | **(min, max)** |
| *Covariate-constrained randomization: top 10% of balance scores* | | | | | | | | | |
| $\sigma_X = 0.5$ | $\beta = 0.0$ | – 0.16 | 0.76 | 0.76 | 0.95 | 1.00 | 3.81 | 0.05 | 1.26 (0.11, 1.92) |
| | $\beta = 0.5$ | – 0.17 | 0.79 | 0.79 | 0.95 | 1.00 | 3.97 | 0.04 | 1.26 (0.11, 1.92) |
| | $\beta = 1.0$ | – 0.18 | 0.86 | 0.86 | 0.96 | 0.99 | 4.42 | 0.04 | 1.26 (0.11, 1.92) |
| $\sigma_X = 1$ | $\beta = 0.0$ | – 0.16 | 0.76 | 0.76 | 0.95 | 1.00 | 3.81 | 0.05 | 1.26 (0.11, 1.92) |
| | $\beta = 0.5$ | – 0.19 | 0.86 | 0.86 | 0.96 | 0.99 | 4.42 | 0.04 | 1.26 (0.11, 1.92) |
| | $\beta = 1.0$ | – 0.21 | 1.17 | 1.18 | 0.98 | 0.94 | 5.90 | 0.02 | 1.26 (0.11, 1.92) |
| $\sigma_X = 2$ | $\beta = 0.0$ | – 0.16 | 0.76 | 0.76 | 0.95 | 1.00 | 3.81 | 0.05 | 1.26 (0.11, 1.92) |
| | $\beta = 0.5$ | – 0.22 | 1.18 | 1.18 | 0.98 | 0.94 | 5.90 | 0.02 | 1.26 (0.11, 1.92) |
| | $\beta = 1.0$ | – 0.27 | 2.41 | 2.41 | 0.99 | 0.52 | 9.80 | 0.01 | 1.26 (0.11, 1.92) |
| *Covariate-constrained randomization: top 20% of balance scores* | | | | | | | | | |
| $\sigma_X = 0.5$ | $\beta = 0.0$ | 0.05 | 0.75 | 0.75 | 0.95 | 1.00 | 3.81 | 0.05 | 1.54 (0.19, 2.24) |
| | $\beta = 0.5$ | 0.06 | 0.78 | 0.78 | 0.95 | 1.00 | 3.96 | 0.05 | 1.54 (0.19, 2.24) |
| | $\beta = 1.0$ | 0.08 | 0.87 | 0.87 | 0.96 | 0.99 | 4.41 | 0.04 | 1.54 (0.19, 2.24) |
| $\sigma_X = 1$ | $\beta = 0.0$ | 0.05 | 0.75 | 0.75 | 0.95 | 1.00 | 3.81 | 0.05 | 1.54 (0.19, 2.24) |
| | $\beta = 0.5$ | 0.08 | 0.87 | 0.87 | 0.96 | 0.99 | 4.41 | 0.04 | 1.54 (0.19, 2.24) |
| | $\beta = 1.0$ | 0.11 | 1.23 | 1.23 | 0.98 | 0.94 | 5.86 | 0.03 | 1.54 (0.19, 2.24) |
| $\sigma_X = 2$ | $\beta = 0.0$ | 0.05 | 0.75 | 0.75 | 0.95 | 1.00 | 3.81 | 0.05 | 1.54 (0.19, 2.24) |
| | $\beta = 0.5$ | 0.10 | 1.23 | 1.23 | 0.98 | 0.94 | 5.86 | 0.03 | 1.54 (0.19, 2.24) |
| | $\beta = 1.0$ | 0.16 | 2.72 | 2.72 | 0.99 | 0.54 | 9.71 | 0.01 | 1.54 (0.19, 2.24) |
| *Covariate-constrained randomization: top 50% of balance scores* | | | | | | | | | |
| $\sigma_X = 0.5$ | $\beta = 0.0$ | 0.01 | 0.76 | 0.76 | 0.95 | 1.00 | 3.80 | 0.05 | 2.09 (0.23, 3.06) |
| | $\beta = 0.5$ | 0.01 | 0.80 | 0.80 | 0.95 | 1.00 | 3.95 | 0.05 | 2.09 (0.23, 3.06) |
| | $\beta = 1.0$ | 0.00 | 0.92 | 0.92 | 0.96 | 0.99 | 4.39 | 0.04 | 2.09 (0.23, 3.06) |
| $\sigma_X = 1$ | $\beta = 0.0$ | 0.01 | 0.76 | 0.76 | 0.95 | 1.00 | 3.80 | 0.05 | 2.09 (0.23, 3.06) |
| | $\beta = 0.5$ | 0.00 | 0.92 | 0.92 | 0.96 | 0.99 | 4.39 | 0.04 | 2.09 (0.23, 3.06) |
| | $\beta = 1.0$ | – 0.01 | 1.42 | 1.42 | 0.97 | 0.93 | 5.80 | 0.03 | 2.09 (0.23, 3.06) |
| $\sigma_X = 2$ | $\beta = 0.0$ | 0.01 | 0.76 | 0.76 | 0.95 | 1.00 | 3.80 | 0.05 | 2.09 (0.23, 3.06) |
| | $\beta = 0.5$ | – 0.01 | 1.42 | 1.42 | 0.97 | 0.94 | 5.80 | 0.03 | 2.09 (0.23, 3.06) |
| | $\beta = 1.0$ | – 0.03 | 3.41 | 3.41 | 0.98 | 0.54 | 9.55 | 0.03 | 2.09 (0.23, 3.06) |
| *Simple Randomization* | | | | | | | | | |
| $\sigma_X = 0.5$ | $\beta = 0.0$ | – 0.02 | 0.75 | 0.75 | 0.95 | 1.00 | 3.80 | 0.05 | 2.99 (0.30, 8.60) |
| | $\beta = 0.5$ | – 0.06 | 0.81 | 0.81 | 0.95 | 1.00 | 3.95 | 0.05 | 2.99 (0.30, 8.60) |
| | $\beta = 1.0$ | – 0.10 | 0.98 | 0.98 | 0.95 | 0.99 | 4.36 | 0.05 | 2.99 (0.30, 8.60) |
| $\sigma_X = 1$ | $\beta = 0.0$ | – 0.02 | 0.75 | 0.75 | 0.95 | 1.00 | 3.80 | 0.05 | 2.99 (0.30, 8.60) |
| | $\beta = 0.5$ | – 0.10 | 0.98 | 0.98 | 0.95 | 0.99 | 4.36 | 0.05 | 2.99 (0.30, 8.60) |
| | $\beta = 1.0$ | – 0.17 | 1.66 | 1.66 | 0.95 | 0.93 | 5.71 | 0.05 | 2.99 (0.30, 8.60) |
| $\sigma_X = 2$ | $\beta = 0.0$ | – 0.02 | 0.75 | 0.75 | 0.95 | 1.00 | 3.80 | 0.05 | 2.99 (0.30, 8.60) |
| | $\beta = 0.5$ | – 0.17 | 1.66 | 1.66 | 0.95 | 0.93 | 5.71 | 0.05 | 2.99 (0.30, 8.60) |
| | $\beta = 1.0$ | – 0.32 | 4.42 | 4.42 | 0.95 | 0.55 | 9.33 | 0.05 | 2.99 (0.30, 8.60) |

*Note: SD* standard deviation, *%Bias* percent bias, *Var* variance, *MSE* mean squared error, *Cov* coverage of the 95% confidence interval, *CI* confidence interval, *Type 1 Error* proportion of type 1 errors under a nominal type 1 error rate of 0.05, *Balance* covariate balance

results for CR are the same across all scenarios within a fixed allocation space and the results for simple randomization are the same across all scenarios. Overall, even when controlling for covariates in the analyses, there is a benefit to using CR as compared to simple randomization in terms of lower MSE, greater power, and narrower confidence interval width. And unlike in the unadjusted analyses, coverage and type 1 error are not conservative and are close to their nominal levels when using CR with 8 clusters and at their nominal levels with 12 clusters.

Comparing simulations with 8 clusters where the analyses does not control for covariates (Table 4) to simulations with 8 clusters where the analysis model does control for covariates (Appendix 2 Table 7) we see that controlling for covariates has an especially adverse effect on power such that power is only 54% under CR (top 10% of balance scores) and 43% under simple randomization. The only scenario where controlling for covariates produces better results than not controlling for covariates is the extreme scenario with the highest potential confounding and the highest between-cluster variability. Here, variance, MSE, power, and CI width are all better when controlling for covariates.

With 12 clusters (Appendix 2 Table 8) there appears to be a clear advantage to controlling for cluster-level covariates in the analyses. The effect on power as compared to not controlling for covariates (Table 5) is modest, and in those scenarios with a high degree of potential confounding, controlling for covariates results in a marked increase in power. For example, under CR (top 10% of balance scores) in the scenario with the highest potential confounding and the highest between-cluster variability, power goes from 0.52 when not controlling for covariates to 0.99 when controlling for covariates. And as mentioned earlier, coverage and type 1 error are at their nominal levels when controlling for covariates.

Results from simulations that included an interaction term are reported in Appendix 3 Tables 9 through 12. Performance criteria (bias, coverage, power) is same for the main effect and the interaction. However, power in these simulations is lower than power in simulations that only include main effects of treatment, owing to the additional cluster-level coefficient in the analysis model for the interaction term. This reduction in power was especially pronounced in simulations based on 8 clusters. Furthermore, when controlling for cluster-level covariates with 8 clusters (Appendix 3 Table 11), coverage was well below the nominal level and type 1 error was elevated compared to the main effects only analysis (Appendix 2 Table 7).

### Application to the BEGIN study

We applied our methods for CR in factorial trials to the BEGIN study, using the cluster-level covariate information in Table 2. With 8 clusters and 4 treatment conditions there are $\frac{8!}{2^4} = 2520$ possible schemes. Using the balance metric in (3), we calculated the balance score for each of these possible 2520 allocation schemes. Based on a belief by the BEGIN investigators that clinic volume was an important predictor of weight loss, and the fact that mean BMI was similar across all clinics, clinic volume was given a weight of 2 in (3), while percent female and mean BMI were given weights of 1. Figure 1 displays a histogram of the balance scores for all 2520 possible schemes. The vertical red line in Fig. 1 indicates the cutoff corresponding to the top 10% balance scores among the 2520 scores.

As mentioned above, for a given set of clinic matches, the treatment assignments can be labeled $4! = 24$ different ways, so that our 2520 possible allocations correspond to only $2520/24 = 105$ unique balance scores. The allocations corresponding to the top 10 unique balance scores are listed in Table 6.

Note that in seven of the ten allocations in Table 6, clinics C7 and C8 are matched together. Clinics C7 and C8 have the largest and smallest clinic volumes, respectively. Assigning them to the same treatment condition helps ensure balance across treatment conditions. Conversely, clinics C1 and C2 are only matched together in two of the ten allocations. Clinics C1 and C2 are the second and third largest clinics. Putting them in *different* treatment conditions also helps ensure balance.

## Discussion

In this paper, we presented a method for performing CR in factorial cluster randomized trials. We performed a simulation study to assess the effectiveness of our method as compared to simple randomization in terms of estimating treatment effects in the setting of a 2 × 2 factorial trial. In all scenarios, bias of the treatment effect was essentially 0. However, by balancing prognostic covariates across treatment arms, CR resulted in more precise estimates of the treatment effect as measured by MSE, a finding also noted by Kalish and Begg [20]. And by constraining the allocation space, CR eliminates the possibility of a highly imbalanced allocation which may significantly undermine the power of a trial as well as threaten its internal validity [10].

When covariates were not controlled for in the analysis model, we found that both CR and simple randomization produced similar rates of power but coverage and Type 1 error rates were conservative under CR, a finding that was also found in Li et al. [6], Watson et al. [10] and Zhou et al. [9]. When covariates were controlled for in the analysis, simulations again showed a clear benefit of CR versus simple randomization across all performance criteria in addition to coverage and type 1 error close to or at their nominal levels. Still, the question of whether or not one should control for covariates in the analysis model is not clear-cut. The rationale to control for cluster-level covariates even when performing CR is that including these covariates helps adjust for any residual imbalances not controlled for during randomization

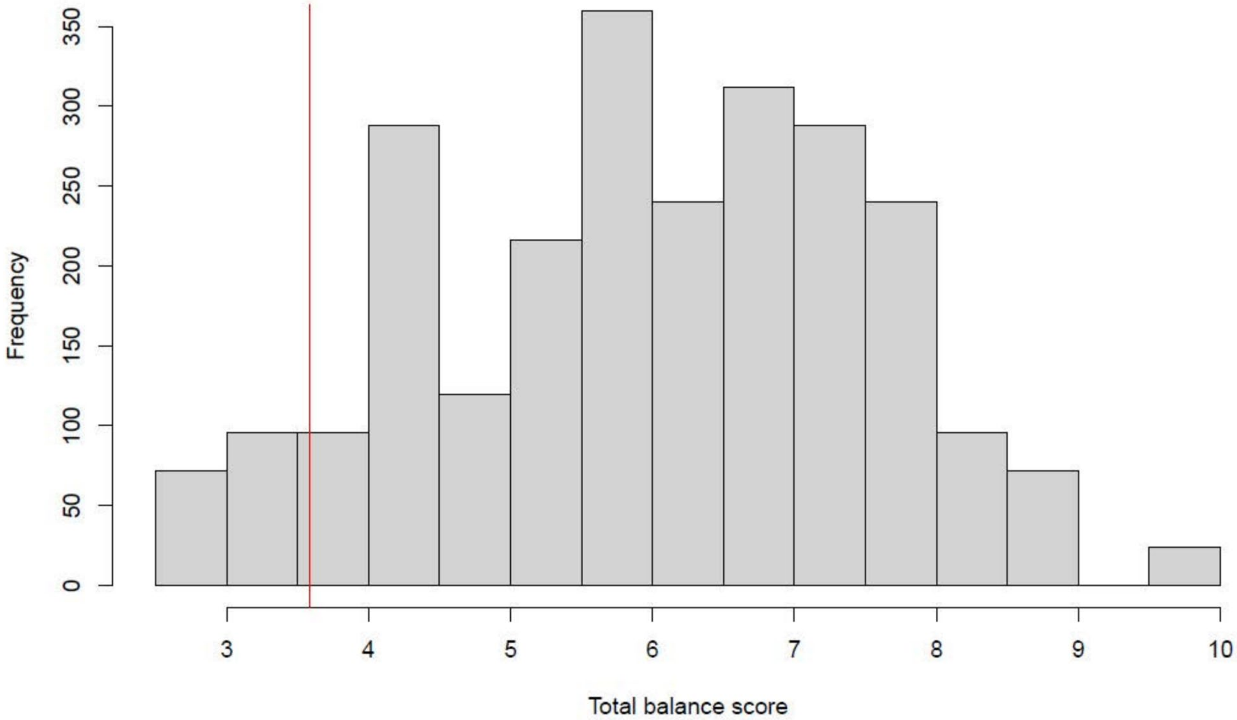

**Fig. 1** Histogram of total balance scores for the 2520 possible allocation schemes for the Behavioral Nudges for Diabetes Prevention (BEGIN) cluster randomized trial with 8 clusters and 4 randomization conditions. The vertical red line indicates the cutoff corresponding to the top 10% of balance scores among the 2520 possible scores

and can also reduce residual variance. The trade-off is a reduction in the number of degrees of freedom for estimating treatment effects. For example, when there are 8 clusters and covariates are not included in the analysis model, there are $8 - 3 = 5$ degrees of freedom available to estimate treatment effects. Including 3 cluster-level covariates in the analysis model reduces this to only 2 degrees of freedom.

In our simulations with 8 clusters, the loss of power when controlling for covariates was so substantial that controlling for covariates is not recommended due to the decrease in degrees of freedom for estimating treatment effects. This loss of power was exacerbated

**Table 6** Clinic pairings associated with the top 10 unique balance scores sorted by total balance score, using data from the Behavioral Nudges for Diabetes Prevention (BEGIN) cluster randomized trial in Table 2

| | Clinic | | | | | | | | Balance score | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Allocation | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | Total | Female | Volume | BMI |
| 1 | a | b | c | b | a | c | d | d | 2.79 | 1.56 | 0.29 | 0.94 |
| 2 | a | b | c | b | c | a | d | d | 2.85 | 1.73 | 0.89 | 0.23 |
| 3 | a | a | b | c | b | c | d | d | 2.92 | 1.35 | 1.18 | 0.39 |
| 4 | a | b | c | c | a | d | d | b | 3.10 | 0.09 | 2.19 | 0.82 |
| 5 | a | b | c | c | a | b | d | d | 3.11 | 2.22 | 0.44 | 0.45 |
| 6 | a | b | b | c | a | c | d | d | 3.17 | 1.57 | 0.42 | 1.18 |
| 7 | a | b | c | d | a | d | c | b | 3.29 | 0.28 | 2.16 | 0.85 |
| 8 | a | b | c | a | c | d | d | b | 3.57 | 0.50 | 2.46 | 0.61 |
| 9 | a | b | c | a | c | b | d | d | 3.58 | 2.63 | 0.70 | 0.25 |
| 10 | a | a | b | b | c | c | d | d | 3.58 | 1.77 | 1.17 | 0.65 |

when an interaction term was included in the model. With 12 clusters, the loss of power when controlling for covariates in the analysis model was minimal, and in some scenarios produced better power than not controlling for covariates. The loss of power when controlling for cluster-level covariates with a small number of clusters highlights another benefit of CR—it allows to user to control for cluster-level covariates during the design stage in order to avoid highly imbalanced designs and obtain more precise inferences—without the resulting decrease in degrees of freedom that would occur if covariates were controlled for in the analysis model.

Power for testing the main effect of treatment was reduced when including an interaction term in the model due to the additional degree of freedom required to estimate this coefficient. While this reduction in power was modest with 12 clusters, it was substantial with 8 clusters. Furthermore, when controlling for cluster-level covariates with 8 clusters (Appendix 3 Table 11), coverage was low and type 1 error was elevated compared to the main effects only analysis (Appendix 2 Table 7). When designing a factorial CRT, investigators need to think carefully about whether estimating interaction terms are of scientific interest and, if so, the study needs to be powered accordingly. Unlike in our simulations where the magnitude of the main effect and interaction terms were the same, in many studies, the interaction term is expected to be smaller in magnitude than the main effect.

As we decreased the size of the constrained allocation space, variance and MSE decreased, with very minor tradeoffs in coverage and type 1 error. When controlling for covariates in the analysis, the size of the allocation space did not affect coverage and type 1 error. Overall, these results suggest that using the top 10% of balance scores is sufficient for achieving balance, reducing MSE, and avoiding highly imbalanced designs.

Although we did not systematically vary the correlations between cluster-level covariates in our simulation study, simulations using independent cluster-level covariates provided results almost identical to those shown here, suggesting that the correlation structure among the cluster-level covariates has little effect on our methods. Ciolino et al. [11] also found in their simulation studies that the magnitude of the correlation between cluster-level covariates had negligible effects on their CR method.

When cluster-level covariates have small variance, as was the case in our simulations when $\sigma_x = 0.5$, there is little benefit to controlling for covariates in the analysis model and a substantial loss of power. This can be seen by comparing Table 4 (top 10% of balance scores) and Appendix 2 Table 7 when $\sigma_x = 0.5$ and $\beta = 1$. Here, power is 90% when not controlling for covariates but only 54% when covariates are included in the analysis model. Only when $\sigma_x = 2$ and the degree of confounding is high is power better when controlling for covariates.

This finding is relevant to the BEGIN study, where there are only 8 clusters and the variability in the cluster-level covariates is small. In our simulation studies, where the mean of the covariates was 1, the coefficient of variation in the cluster-level covariates ranged from 0.7 when $\sigma_x = 0.5$, to 1.4, when $\sigma_x = 2$. In Table 2, the clinic volume coefficient of variation is 0.44. But the coefficient of variation for percent female is 0.08 and the coefficient of variation for mean BMI is only 0.01. These values suggest that the analysis model for the BEGIN study should not control for clinic-level covariates unless the distribution of clinic-level covariates in the actual trial data is much different from the values in Table 2.

In BEGIN, two of the cluster-level covariates in Table 2, percent female and mean BMI, can be collected during the course of the study at the individual level and adjusted for in the analysis model as individual-level covariates. However, variables at the clinic level can have a different effect than a similar covariate at the individual level. For example, a clinic where the mean BMI is high may have providers who are more likely to bring up weight loss with their patients so that prediabetic patients at this clinic are *more* likely to lose weight over time compared to prediabetic patients at other clinics. BMI could have an opposite effect at the individual level. For example, patients with high BMI may be *less* likely to lose weight over time than patients with lower BMI. Controlling for BMI at *both* the clinic level and the individual level may be important to reduce confounding, but one is not necessarily a substitute for the other.

Our simulation results under simple random sampling are averages over all possible allocations and unconstrained randomization retains the possibility to select a highly unbalanced design. As suggested by a reviewer, we performed additional simulations where the allocation space was based on the *bottom* 10% of balance scores. When not controlling for covariates, results (not shown) based on the bottom 10% of balance scores had worse bias, variance, MSE, and coverage than results using simple random sampling. Furthermore, the type 1 error rate exceeded the nominal level.

When controlling for covariates, results based on the bottom 10% of balance scores had worse bias, variance, MSE, and power than results based on simple random sampling. But coverage and type 1 error rates were preserved at their nominal levels.

When randomizing 12 clinics using CR, we sampled from 20,000 of the 369,300 possible allocations. To evaluate whether sampling from 20,000 allocations is sufficient for approximating the entire randomization space, we repeated our simulations but sampled from 50,000 of the possible allocations. These results (not shown) were almost identical to the results in Table 5 and Appendix 2 Table 8 suggesting that—at least for 12 clinics and 4 treatment conditions—sampling from more than 20,000 of the total possible allocations does not affect our results.

An alternative to the model-based methods used in this manuscript are randomization-based methods. Work by Zhou et al. [9] in the multi-arm setting has shown that when baseline covariates are balanced through CR, covariate adjustment at the analysis stage is necessary for *model-based* tests to maintain nominal type 1 error rates but randomization-based tests do not require this adjustment, a finding also reported by Li et al. [6] in two-group designs. Zhou et al. [9] also found that power is better under CR versus simple randomization in unadjusted analyses, although care must be taken when selecting the size of the constrained randomization space. Randomization tests are also more robust to violations of distributional assumptions. For these reasons, developing and evaluating methods for randomization-based inference in factorial CRTs is a promising area of future research.

There are several limitations to our study. We evaluated our methods using a 2 × 2 factorial trial and it is not clear whether our methods would work equally well with a $2^k$ or other larger factorial trial with additional treatment conditions. We evaluated our balance metric using simulated continuous covariates. Future work will evaluate how well our methods perform when binary or categorical group-level covariates are used to constrain the randomization set and the outcome is continuous or binary.

## Conclusions

Our findings provide evidence for the use of CR instead of simple randomization when performing factorial CRTs to avoid highly imbalanced designs and to obtain more precise inferences. Except when there are a small number of clusters per treatment condition, cluster-level covariates should be included in the analysis model to increase power and produce coverage and type 1 error rates at their nominal levels. When there are a small number of clusters, we recommend cluster-level covariates should not be included in the analysis model due to the loss of power even though coverage and type 1 error rates will be conservative in the unadjusted analyses.

## Appendix

### Appendix 1 R code for covariate-constrained randomization in the BEGIN 2 × 2 factorial trial

The R code below is to implement covariate constrained randomization in the BEGIN 2 × 2 factorial trial. The code below is for three potential cluster-level confounders and 4 randomization conditions.

```
CR_fn <- function(data, x1.w=1, x2.w=1, x3.w=1, nsample=20000, con.space=0.1){
  # function to perform covariate-constrained randomization
  # Args:
  #   data: dataframe of 3 cluster-level covariates (x1, x2, x3)
  #   x1.w, x2.w, x3.w: user-defined weights
  #   nsample: number of sampled allocations to be used when there are
  #            more than 8 clusters
  #   con.space: Size of constrained allocation space. Draw from the top
  #              (con.space * 100)% of balance scores
  # Returns:
  #  The data dataset with treatment assignment and total balance appended

  # required packages
  library(doBy)
  library(arrangements)

  # If 8 clinics then enumerate all possible permutations
  if(nrow(data) == 8) nsample <- NULL else nsample <- nsample

  data <- as.data.frame(data)
  J <- nrow(data)

  # calculate inverse of variance
  x1.d <- 1/var(data$x1)
  x2.d <- 1/var(data$x2)
  x3.d <- 1/var(data$x3)

  # calculate the balance scores
  # generate permutations for J clinics and 4 conditions
  perms <- permutations(4, freq=c(J/4, J/4, J/4, J/4), nsample=nsample)

  # remove duplicate allocations
  perms <- unique(perms)

  B <- rep(0, nrow(perms)*4) # vector of balance scores
  dim(B) <- c(nrow(perms), 4)

  for(i in 1:nrow(perms)){
    data$tx <- perms[i,] # calculate by permutation
    means <- summaryBy(cbind(x1, x2, x3) ~ tx, data=data)

    # balance score is the sum of squared deviations of the covariate
    # means multiplied by their weight which is the inverse
    # of the cluster variances
    x1.bal <- (nrow(means)-1) * var(means$x1.mean) * x1.w * x1.d
    x2.bal <- (nrow(means)-1) * var(means$x2.mean) * x2.w * x2.d
    x3.bal <- (nrow(means)-1) * var(means$x3.mean) * x3.w * x3.d

    # add up covariate-specific balance scores
    total.bal <- sum(x1.bal, x2.bal, x3.bal)

    #return all balance scores
    B[i,1:4] <- c(total.bal, x1.bal, x2.bal, x3.bal)
  }

  # merge balance scores with permutations data set
  permdata <- data.frame(perms)
  permdata <- cbind(permdata, B)
  names(permdata) <- c("1":J, "total.bal", "x1.bal", "x2.bal", "x3.bal")

  # sort by total balance score
  permsort <- permdata[order(permdata$total.bal),]

  # Randomly select one allocation from top (con.space * 100)%
  # of balance scores
  n <- ifelse(J==8, round(((factorial(J)/(factorial(J/4)^4))/factorial(4))
              * con.space) * factorial(4), round(con.space*nrow(perms)))

  ralloc <- sample(1:n, 1)
  CR.trt <- permsort[ralloc,]

  # assign treatment 1 (Decision Support) to those clinics who were
  # randomized to conditions 1 or 3
  trt1 <- as.numeric(CR.trt[,1:J] == 1 | CR.trt[,1:J] == 3)

  # assign treatment 2 (Text message) to those clinics who were
  # randomized to conditions 2 or 3
  trt2 <- as.numeric(CR.trt[,1:J] == 2 | CR.trt[,1:J] == 3)

  # add total balance score in the output

  total.bal <- CR.trt[,J+1]

  CR <- cbind(data, trt1, trt2, total.bal)
  return(CR)
}
```

## Appendix 2 Simulation results controlling for cluster-level covariates in a main effects only analysis

**Table 7** Simulation results for the effect of treatment with 8 clusters, based on a main effects only analysis model that controls for cluster-level covariates

| Covariate SD | Degree Confounding | %Bias | Var | MSE | Cov | Power | 95% CI Width | Type 1 Error | Mean balance (min, max) |
|---|---|---|---|---|---|---|---|---|---|
| *Covariate-constrained randomization: top 10% of balance scores* | | | | | | | | | |
| $\sigma_X = 0.5$ | $\beta = 0.0$ | 0.07 | 2.19 | 2.19 | 0.93 | 0.54 | 10.20 | 0.07 | 2.13 (0.35, 3.58) |
| | $\beta = 0.5$ | 0.07 | 2.19 | 2.19 | 0.93 | 0.54 | 10.20 | 0.07 | 2.13 (0.35, 3.58) |
| | $\beta = 1.0$ | 0.07 | 2.19 | 2.19 | 0.93 | 0.54 | 10.20 | 0.07 | 2.13 (0.35, 3.58) |
| $\sigma_X = 1$ | $\beta = 0.0$ | 0.07 | 2.19 | 2.19 | 0.93 | 0.54 | 10.20 | 0.07 | 2.13 (0.35, 3.58) |
| | $\beta = 0.5$ | 0.07 | 2.19 | 2.19 | 0.93 | 0.54 | 10.20 | 0.07 | 2.13 (0.35, 3.58) |
| | $\beta = 1.0$ | 0.07 | 2.19 | 2.19 | 0.93 | 0.54 | 10.20 | 0.07 | 2.13 (0.35, 3.58) |
| $\sigma_X = 2$ | $\beta = 0.0$ | 0.08 | 2.19 | 2.19 | 0.93 | 0.54 | 10.20 | 0.07 | 2.13 (0.35, 3.58) |
| | $\beta = 0.5$ | 0.08 | 2.19 | 2.19 | 0.93 | 0.54 | 10.20 | 0.07 | 2.13 (0.35, 3.58) |
| | $\beta = 1.0$ | 0.08 | 2.19 | 2.19 | 0.93 | 0.54 | 10.20 | 0.07 | 2.13 (0.35, 3.58) |
| *Covariate-constrained randomization: top 20% of balance scores* | | | | | | | | | |
| $\sigma_X = 0.5$ | $\beta = 0.0$ | – 0.20 | 2.34 | 2.34 | 0.93 | 0.52 | 10.77 | 0.07 | 2.56 (0.35, 3.79) |
| | $\beta = 0.5$ | – 0.20 | 2.34 | 2.34 | 0.93 | 0.52 | 10.77 | 0.07 | 2.56 (0.35, 3.79) |
| | $\beta = 1.0$ | – 0.20 | 2.34 | 2.34 | 0.93 | 0.52 | 10.77 | 0.07 | 2.56 (0.35, 3.79) |
| $\sigma_X = 1$ | $\beta = 0.0$ | – 0.21 | 2.34 | 2.34 | 0.93 | 0.52 | 10.77 | 0.07 | 2.56 (0.35, 3.79) |
| | $\beta = 0.5$ | – 0.21 | 2.34 | 2.34 | 0.93 | 0.52 | 10.77 | 0.07 | 2.56 (0.35, 3.79) |
| | $\beta = 1.0$ | – 0.21 | 2.34 | 2.34 | 0.93 | 0.52 | 10.77 | 0.07 | 2.56 (0.35, 3.79) |
| $\sigma_X = 2$ | $\beta = 0.0$ | – 0.20 | 2.34 | 2.34 | 0.93 | 0.52 | 10.77 | 0.07 | 2.56 (0.35, 3.79) |
| | $\beta = 0.5$ | – 0.20 | 2.34 | 2.34 | 0.93 | 0.52 | 10.77 | 0.07 | 2.56 (0.35, 3.79) |
| | $\beta = 1.0$ | – 0.20 | 2.34 | 2.34 | 0.93 | 0.52 | 10.77 | 0.07 | 2.56 (0.35, 3.79) |
| *Covariate-constrained randomization: top 50% of balance scores* | | | | | | | | | |
| $\sigma_X = 0.5$ | $\beta = 0.0$ | 0.11 | 3.27 | 3.27 | 0.93 | 0.48 | 12.00 | 0.07 | 3.34 (0.50, 4.80) |
| | $\beta = 0.5$ | 0.11 | 3.27 | 3.27 | 0.93 | 0.48 | 12.00 | 0.07 | 3.34 (0.50, 4.80) |
| | $\beta = 1.0$ | 0.11 | 3.27 | 3.27 | 0.93 | 0.48 | 12.00 | 0.07 | 3.34 (0.50, 4.80) |
| $\sigma_X = 1$ | $\beta = 0.0$ | 0.09 | 3.27 | 3.27 | 0.93 | 0.48 | 12.00 | 0.07 | 3.34 (0.50, 4.80) |
| | $\beta = 0.5$ | 0.09 | 3.27 | 3.27 | 0.93 | 0.48 | 12.00 | 0.07 | 3.34 (0.50, 4.80) |
| | $\beta = 1.0$ | 0.09 | 3.27 | 3.27 | 0.93 | 0.48 | 12.00 | 0.07 | 3.34 (0.50, 4.80) |
| $\sigma_X = 2$ | $\beta = 0.0$ | 0.08 | 3.27 | 3.27 | 0.93 | 0.47 | 12.00 | 0.07 | 3.34 (0.50, 4.80) |
| | $\beta = 0.5$ | 0.08 | 3.27 | 3.27 | 0.93 | 0.47 | 12.00 | 0.07 | 3.34 (0.50, 4.80) |
| | $\beta = 1.0$ | 0.08 | 3.27 | 3.27 | 0.93 | 0.47 | 12.00 | 0.07 | 3.34 (0.50, 4.80) |
| *Simple randomization* | | | | | | | | | |
| $\sigma_X = 0.5$ | $\beta = 0.0$ | 0.10 | 3.95 | 3.95 | 0.93 | 0.43 | 13.35 | 0.07 | 4.50 (0.49, 9.42) |
| | $\beta = 0.5$ | 0.10 | 3.95 | 3.95 | 0.93 | 0.43 | 13.35 | 0.07 | 4.50 (0.49, 9.42) |
| | $\beta = 1.0$ | 0.10 | 3.95 | 3.95 | 0.93 | 0.43 | 13.35 | 0.07 | 4.50 (0.49, 9.42) |
| $\sigma_X = 1$ | $\beta = 0.0$ | 0.10 | 3.95 | 3.95 | 0.93 | 0.43 | 13.35 | 0.07 | 4.50 (0.49, 9.42) |
| | $\beta = 0.5$ | 0.10 | 3.95 | 3.95 | 0.93 | 0.43 | 13.35 | 0.07 | 4.50 (0.49, 9.42) |
| | $\beta = 1.0$ | 0.10 | 3.95 | 3.95 | 0.93 | 0.43 | 13.35 | 0.07 | 4.50 (0.49, 9.42) |
| $\sigma_X = 2$ | $\beta = 0.0$ | 0.10 | 3.95 | 3.95 | 0.93 | 0.43 | 13.35 | 0.07 | 4.50 (0.49, 9.42) |
| | $\beta = 0.5$ | 0.10 | 3.95 | 3.95 | 0.93 | 0.43 | 13.35 | 0.07 | 4.50 (0.49, 9.42) |
| | $\beta = 1.0$ | 0.10 | 3.95 | 3.95 | 0.93 | 0.43 | 13.35 | 0.07 | 4.50 (0.49, 9.42) |

*Note: SD* standard deviation, *%Bias* percent bias, *Var* variance, *MSE* mean squared error, *Cov* coverage of the 95% confidence interval, *CI* confidence interval, *Type 1 Error* proportion of type 1 errors under a nominal type 1 error rate of 0.05, *Balance* covariate balance

**Table 8** Simulation results for the effect of treatment with 12 clusters, based on a main effects only analysis model that controls for cluster-level covariates

| Covariate SD | Degree Confounding | %Bias | Var | MSE | Cov | Power | 95% CI Width | Type 1 Error | Mean balance (min, max) |
|---|---|---|---|---|---|---|---|---|---|
| *Covariate-constrained randomization: rop 10% of balance scores* | | | | | | | | | |
| $\sigma_x = 0.5$ | $\beta = 0.0$ | – 0.18 | 0.90 | 0.90 | 0.95 | 0.99 | 4.42 | 0.05 | 1.26 (0.11, 1.92) |
| | $\beta = 0.5$ | – 0.18 | 0.90 | 0.90 | 0.95 | 0.99 | 4.42 | 0.05 | 1.26 (0.11, 1.92) |
| | $\beta = 1.0$ | – 0.18 | 0.90 | 0.90 | 0.95 | 0.99 | 4.42 | 0.05 | 1.26 (0.11, 1.92) |
| $\sigma_x = 1$ | $\beta = 0.0$ | – 0.18 | 0.90 | 0.90 | 0.95 | 0.99 | 4.42 | 0.05 | 1.26 (0.11, 1.92) |
| | $\beta = 0.5$ | – 0.18 | 0.90 | 0.90 | 0.95 | 0.99 | 4.42 | 0.05 | 1.26 (0.11, 1.92) |
| | $\beta = 1.0$ | – 0.18 | 0.90 | 0.90 | 0.95 | 0.99 | 4.42 | 0.05 | 1.26 (0.11, 1.92) |
| $\sigma_x = 2$ | $\beta = 0.0$ | – 0.18 | 0.90 | 0.90 | 0.95 | 0.99 | 4.42 | 0.05 | 1.26 (0.11, 1.92) |
| | $\beta = 0.5$ | – 0.18 | 0.90 | 0.90 | 0.95 | 0.99 | 4.42 | 0.05 | 1.26 (0.11, 1.92) |
| | $\beta = 1.0$ | – 0.18 | 0.90 | 0.90 | 0.95 | 0.99 | 4.42 | 0.05 | 1.26 (0.11, 1.92) |
| *Covariate-constrained randomization: top 20% of balance scores* | | | | | | | | | |
| $\sigma_x = 0.5$ | $\beta = 0.0$ | – 0.06 | 0.93 | 0.93 | 0.95 | 0.99 | 4.51 | 0.05 | 1.54 (0.19, 2.24) |
| | $\beta = 0.5$ | – 0.06 | 0.93 | 0.93 | 0.95 | 0.99 | 4.51 | 0.05 | 1.54 (0.19, 2.24) |
| | $\beta = 1.0$ | – 0.06 | 0.93 | 0.93 | 0.95 | 0.99 | 4.51 | 0.05 | 1.54 (0.19, 2.24) |
| $\sigma_x = 1$ | $\beta = 0.0$ | – 0.06 | 0.93 | 0.93 | 0.95 | 0.99 | 4.51 | 0.05 | 1.54 (0.19, 2.24) |
| | $\beta = 0.5$ | – 0.06 | 0.93 | 0.93 | 0.95 | 0.99 | 4.51 | 0.05 | 1.54 (0.19, 2.24) |
| | $\beta = 1.0$ | – 0.06 | 0.93 | 0.93 | 0.95 | 0.99 | 4.51 | 0.05 | 1.54 (0.19, 2.24) |
| $\sigma_x = 2$ | $\beta = 0.0$ | – 0.06 | 0.93 | 0.93 | 0.95 | 0.99 | 4.51 | 0.05 | 1.54 (0.19, 2.24) |
| | $\beta = 0.5$ | – 0.06 | 0.93 | 0.93 | 0.95 | 0.99 | 4.51 | 0.05 | 1.54 (0.19, 2.24) |
| | $\beta = 1.0$ | – 0.06 | 0.93 | 0.93 | 0.95 | 0.99 | 4.51 | 0.05 | 1.54 (0.19, 2.24) |
| *Covariate-constrained randomization: top 50% of balance scores* | | | | | | | | | |
| $\sigma_x = 0.5$ | $\beta = 0.0$ | – 0.06 | 1.02 | 1.02 | 0.95 | 0.98 | 4.69 | 0.05 | 2.09 (0.23, 3.06) |
| | $\beta = 0.5$ | – 0.06 | 1.02 | 1.02 | 0.95 | 0.98 | 4.69 | 0.05 | 2.09 (0.23, 3.06) |
| | $\beta = 1.0$ | – 0.06 | 1.02 | 1.02 | 0.95 | 0.98 | 4.69 | 0.05 | 2.09 (0.23, 3.06) |
| $\sigma_x = 1$ | $\beta = 0.0$ | – 0.06 | 1.02 | 1.02 | 0.95 | 0.98 | 4.69 | 0.05 | 2.09 (0.23, 3.06) |
| | $\beta = 0.5$ | – 0.06 | 1.02 | 1.02 | 0.95 | 0.98 | 4.69 | 0.05 | 2.09 (0.23, 3.06) |
| | $\beta = 1.0$ | – 0.06 | 1.02 | 1.02 | 0.95 | 0.98 | 4.69 | 0.05 | 2.09 (0.23, 3.06) |
| $\sigma_x = 2$ | $\beta = 0.0$ | – 0.06 | 1.02 | 1.02 | 0.95 | 0.98 | 4.69 | 0.05 | 2.09 (0.23, 3.06) |
| | $\beta = 0.5$ | – 0.06 | 1.02 | 1.02 | 0.95 | 0.98 | 4.69 | 0.05 | 2.09 (0.23, 3.06) |
| | $\beta = 1.0$ | – 0.06 | 1.02 | 1.02 | 0.95 | 0.98 | 4.69 | 0.05 | 2.09 (0.23, 3.06) |
| *Simple randomization* | | | | | | | | | |
| $\sigma_x = 0.5$ | $\beta = 0.0$ | – 0.04 | 1.19 | 1.19 | 0.95 | 0.95 | 5.03 | 0.05 | 2.99 (0.30, 8.60) |
| | $\beta = 0.5$ | – 0.04 | 1.19 | 1.19 | 0.95 | 0.95 | 5.03 | 0.05 | 2.99 (0.30, 8.60) |
| | $\beta = 1.0$ | – 0.04 | 1.19 | 1.19 | 0.95 | 0.95 | 5.03 | 0.05 | 2.99 (0.30, 8.60) |
| $\sigma_x = 1$ | $\beta = 0.0$ | – 0.04 | 1.19 | 1.19 | 0.95 | 0.95 | 5.03 | 0.05 | 2.99 (0.30, 8.60) |
| | $\beta = 0.5$ | – 0.04 | 1.19 | 1.19 | 0.95 | 0.95 | 5.03 | 0.05 | 2.99 (0.30, 8.60) |
| | $\beta = 1.0$ | – 0.04 | 1.19 | 1.19 | 0.95 | 0.95 | 5.03 | 0.05 | 2.99 (0.30, 8.60) |
| $\sigma_x = 2$ | $\beta = 0.0$ | – 0.04 | 1.19 | 1.19 | 0.95 | 0.95 | 5.03 | 0.05 | 2.99 (0.30, 8.60) |
| | $\beta = 0.5$ | – 0.04 | 1.19 | 1.19 | 0.95 | 0.95 | 5.03 | 0.05 | 2.99 (0.30, 8.60) |
| | $\beta = 1.0$ | – 0.04 | 1.19 | 1.19 | 0.95 | 0.95 | 5.03 | 0.05 | 2.99 (0.30, 8.60) |

*Note: SD* standard deviation, *%Bias* percent bias, *Var* variance, *MSE* mean squared error, *Cov* coverage of the 95% confidence interval, *CI* confidence interval, *Type 1 Error* proportion of type 1 errors under a nominal type 1 error rate of 0.05, *Balance* covariate balance

## Appendix 3 Simulation results based on data generating and analysis models that include an interaction term

**Table 9** Simulation results for the main effect of treatments and their interaction with 8 clusters, based on an analysis model that does not control for cluster-level covariates

| | | Main effects | | | | Interaction | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Covariate** | **Degree** | **%Bias** | **Cov** | **Power** | **Type 1** | **%Bias** | **Cov** | **Power** | **Mean balance** |
| **SD** | **Confounding** | | | | **Error** | | | | **(min, max)** |
| *Covariate-constrained randomization: top 10% of balance scores* | | | | | | | | | |
| $\sigma_x = 0.5$ | $\beta = 0.0$ | 0.07 | 0.95 | 0.93 | 0.06 | 0.00 | 0.95 | 0.93 | 2.13 (0.35, 3.58) |
| | $\beta = 0.5$ | 0.00 | 0.95 | 0.91 | 0.05 | 0.01 | 0.95 | 0.91 | 2.13 (0.35, 3.58) |
| | $\beta = 1.0$ | − 0.06 | 0.96 | 0.85 | 0.04 | 0.02 | 0.96 | 0.84 | 2.13 (0.35, 3.58) |
| $\sigma_x = 1$ | $\beta = 0.0$ | 0.08 | 0.95 | 0.93 | 0.06 | 0.00 | 0.95 | 0.93 | 2.13 (0.35, 3.58) |
| | $\beta = 0.5$ | − 0.06 | 0.96 | 0.85 | 0.04 | 0.02 | 0.96 | 0.84 | 2.13 (0.35, 3.58) |
| | $\beta = 1.0$ | − 0.19 | 0.98 | 0.58 | 0.02 | 0.04 | 0.98 | 0.59 | 2.13 (0.35, 3.58) |
| $\sigma_x = 2$ | $\beta = 0.0$ | 0.07 | 0.95 | 0.93 | 0.06 | 0.00 | 0.95 | 0.93 | 2.13 (0.35, 3.58) |
| | $\beta = 0.5$ | − 0.19 | 0.98 | 0.58 | 0.02 | 0.04 | 0.98 | 0.59 | 2.13 (0.35, 3.58) |
| | $\beta = 1.0$ | − 0.45 | 0.99 | 0.19 | 0.01 | 0.09 | 0.99 | 0.20 | 2.13 (0.35, 3.58) |
| *Covariate-constrained randomization: top 20% of balance scores* | | | | | | | | | |
| $\sigma_x = 0.5$ | $\beta = 0.0$ | − 0.27 | 0.95 | 0.93 | 0.05 | 0.22 | 0.95 | 0.93 | 2.56 (0.35, 3.79) |
| | $\beta = 0.5$ | − 0.29 | 0.95 | 0.91 | 0.05 | 0.22 | 0.95 | 0.91 | 2.56 (0.35, 3.79) |
| | $\beta = 1.0$ | − 0.31 | 0.96 | 0.85 | 0.04 | 0.22 | 0.96 | 0.84 | 2.56 (0.35, 3.79) |
| $\sigma_x = 1$ | $\beta = 0.0$ | − 0.27 | 0.95 | 0.93 | 0.05 | 0.22 | 0.95 | 0.93 | 2.56 (0.35, 3.79) |
| | $\beta = 0.5$ | − 0.31 | 0.96 | 0.85 | 0.04 | 0.23 | 0.96 | 0.84 | 2.56 (0.35, 3.79) |
| | $\beta = 1.0$ | − 0.36 | 0.98 | 0.60 | 0.02 | 0.24 | 0.97 | 0.61 | 2.56 (0.35, 3.79) |
| $\sigma_x = 2$ | $\beta = 0.0$ | − 0.27 | 0.95 | 0.93 | 0.05 | 0.22 | 0.95 | 0.93 | 2.56 (0.35, 3.79) |
| | $\beta = 0.5$ | − 0.34 | 0.98 | 0.60 | 0.02 | 0.23 | 0.97 | 0.61 | 2.56 (0.35, 3.79) |
| | $\beta = 1.0$ | − 0.42 | 0.99 | 0.21 | 0.01 | 0.24 | 0.98 | 0.22 | 2.56 (0.35, 3.79) |
| *Covariate-constrained randomization: top 50% of balance scores* | | | | | | | | | |
| $\sigma_x = 0.5$ | $\beta = 0.0$ | − 0.06 | 0.95 | 0.93 | 0.05 | − 0.22 | 0.95 | 0.93 | 3.34 (0.50, 4.80) |
| | $\beta = 0.5$ | − 0.06 | 0.95 | 0.91 | 0.05 | − 0.31 | 0.95 | 0.91 | 3.34 (0.50, 4.80) |
| | $\beta = 1.0$ | − 0.06 | 0.96 | 0.85 | 0.04 | − 0.40 | 0.96 | 0.85 | 3.34 (0.50, 4.80) |
| $\sigma_x = 1$ | $\beta = 0.0$ | − 0.06 | 0.95 | 0.93 | 0.05 | − 0.22 | 0.95 | 0.93 | 3.34 (0.50, 4.80) |
| | $\beta = 0.5$ | − 0.07 | 0.96 | 0.85 | 0.04 | − 0.41 | 0.96 | 0.85 | 3.34 (0.50, 4.80) |
| | $\beta = 1.0$ | − 0.07 | 0.97 | 0.62 | 0.03 | − 0.59 | 0.97 | 0.62 | 3.34 (0.50, 4.80) |
| $\sigma_x = 2$ | $\beta = 0.0$ | − 0.08 | 0.95 | 0.93 | 0.05 | − 0.21 | 0.95 | 0.93 | 3.34 (0.50, 4.80) |
| | $\beta = 0.5$ | − 0.08 | 0.97 | 0.62 | 0.03 | − 0.58 | 0.97 | 0.62 | 3.34 (0.50, 4.80) |
| | $\beta = 1.0$ | − 0.09 | 0.97 | 0.26 | 0.02 | − 0.95 | 0.98 | 0.26 | 3.34 (0.50, 4.80) |
| *Simple randomization* | | | | | | | | | |
| $\sigma_x = 0.5$ | $\beta = 0.0$ | − 0.06 | 0.94 | 0.93 | 0.05 | − 0.14 | 0.95 | 0.93 | 4.50 (0.49, 9.42) |
| | $\beta = 0.5$ | − 0.12 | 0.95 | 0.91 | 0.05 | − 0.07 | 0.95 | 0.92 | 4.50 (0.49, 9.42) |
| | $\beta = 1.0$ | − 0.18 | 0.95 | 0.86 | 0.05 | 0.01 | 0.95 | 0.87 | 4.50 (0.49, 9.42) |
| $\sigma_x = 1$ | $\beta = 0.0$ | − 0.06 | 0.94 | 0.93 | 0.05 | − 0.14 | 0.95 | 0.93 | 4.50 (0.49, 9.42) |
| | $\beta = 0.5$ | − 0.18 | 0.95 | 0.86 | 0.05 | 0.01 | 0.95 | 0.87 | 4.50 (0.49, 9.42) |
| | $\beta = 1.0$ | − 0.30 | 0.95 | 0.67 | 0.05 | 0.16 | 0.95 | 0.66 | 4.50 (0.49, 9.42) |
| $\sigma_x = 2$ | $\beta = 0.0$ | − 0.06 | 0.94 | 0.93 | 0.05 | − 0.14 | 0.95 | 0.93 | 4.50 (0.49, 9.42) |
| | $\beta = 0.5$ | − 0.30 | 0.95 | 0.67 | 0.05 | 0.16 | 0.95 | 0.66 | 4.50 (0.49, 9.42) |
| | $\beta = 1.0$ | − 0.54 | 0.95 | 0.32 | 0.05 | 0.46 | 0.95 | 0.32 | 4.50 (0.49, 9.42) |

*Note: SD* standard deviation, *%Bias* percent bias, *Cov* coverage of the 95% confidence interval, *Type 1 Error* proportion of type 1 errors under a nominal type 1 error rate of 0.05, *Balance* covariate balance

**Table 10** Simulation results for the main effect of treatments and their interaction with 12 clusters, based on an analysis model that does not control for cluster-level covariates

| | | Main effects | | | | Interaction | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Covariate** | **Degree** | **%Bias** | **Cov** | **Power** | **Type 1** | **%Bias** | **Cov** | **Power** | **Mean balance** |
| **SD** | **Confounding** | | | | **Error** | | | | **(min, max)** |
| *Covariate-constrained randomization: top 10% of balance scores* | | | | | | | | | |
| $\sigma_x = 0.5$ | $\beta = 0.0$ | − 0.16 | 0.95 | 1.00 | 0.05 | 0.31 | 0.95 | 1.00 | 1.26 (0.11, 1.92) |
| | $\beta = 0.5$ | − 0.17 | 0.95 | 1.00 | 0.05 | 0.31 | 0.95 | 1.00 | 1.26 (0.11, 1.92) |
| | $\beta = 1.0$ | − 0.18 | 0.96 | 0.99 | 0.03 | 0.31 | 0.96 | 0.99 | 1.26 (0.11, 1.92) |
| $\sigma_x = 1$ | $\beta = 0.0$ | − 0.16 | 0.95 | 1.00 | 0.05 | 0.31 | 0.95 | 1.00 | 1.26 (0.11, 1.92) |
| | $\beta = 0.5$ | − 0.18 | 0.96 | 0.99 | 0.03 | 0.31 | 0.96 | 0.99 | 1.26 (0.11, 1.92) |
| | $\beta = 1.0$ | − 0.21 | 0.98 | 0.92 | 0.01 | 0.30 | 0.98 | 0.93 | 1.26 (0.11, 1.92) |
| $\sigma_x = 2$ | $\beta = 0.0$ | − 0.16 | 0.95 | 1.00 | 0.05 | 0.31 | 0.95 | 1.00 | 1.26 (0.11, 1.92) |
| | $\beta = 0.5$ | − 0.22 | 0.98 | 0.92 | 0.01 | 0.31 | 0.98 | 0.93 | 1.26 (0.11, 1.92) |
| | $\beta = 1.0$ | − 0.27 | 0.99 | 0.47 | 0.01 | 0.31 | 0.99 | 0.48 | 1.26 (0.11, 1.92) |
| *Covariate-constrained randomization: top 20% of balance scores* | | | | | | | | | |
| $\sigma_x = 0.5$ | $\beta = 0.0$ | 0.05 | 0.95 | 1.00 | 0.05 | 0.29 | 0.95 | 1.00 | 1.54 (0.19, 2.24) |
| | $\beta = 0.5$ | 0.06 | 0.95 | 1.00 | 0.05 | 0.25 | 0.96 | 1.00 | 1.54 (0.19, 2.24) |
| | $\beta = 1.0$ | 0.08 | 0.96 | 0.99 | 0.04 | 0.21 | 0.96 | 0.99 | 1.54 (0.19, 2.24) |
| $\sigma_x = 1$ | $\beta = 0.0$ | 0.05 | 0.95 | 1.00 | 0.05 | 0.29 | 0.95 | 1.00 | 1.54 (0.19, 2.24) |
| | $\beta = 0.5$ | 0.08 | 0.96 | 0.99 | 0.04 | 0.21 | 0.96 | 0.99 | 1.54 (0.19, 2.24) |
| | $\beta = 1.0$ | 0.11 | 0.98 | 0.92 | 0.02 | 0.13 | 0.98 | 0.93 | 1.54 (0.19, 2.24) |
| $\sigma_x = 2$ | $\beta = 0.0$ | 0.05 | 0.95 | 1.00 | 0.05 | 0.29 | 0.95 | 1.00 | 1.54 (0.19, 2.24) |
| | $\beta = 0.5$ | 0.10 | 0.98 | 0.92 | 0.02 | 0.14 | 0.98 | 0.93 | 1.54 (0.19, 2.24) |
| | $\beta = 1.0$ | 0.16 | 0.99 | 0.49 | 0.01 | -0.02 | 0.99 | 0.49 | 1.54 (0.19, 2.24) |
| *Covariate-constrained randomization: top 50% of balance scores* | | | | | | | | | |
| $\sigma_x = 0.5$ | $\beta = 0.0$ | 0.01 | 0.95 | 1.00 | 0.05 | − 0.22 | 0.95 | 1.00 | 2.09 (0.23, 3.06) |
| | $\beta = 0.5$ | 0.01 | 0.95 | 1.00 | 0.05 | − 0.19 | 0.95 | 1.00 | 2.09 (0.23, 3.06) |
| | $\beta = 1.0$ | 0.00 | 0.96 | 0.99 | 0.04 | − 0.16 | 0.96 | 0.99 | 2.09 (0.23, 3.06) |
| $\sigma_x = 1$ | $\beta = 0.0$ | 0.01 | 0.95 | 1.00 | 0.05 | − 0.23 | 0.95 | 1.00 | 2.09 (0.23, 3.06) |
| | $\beta = 0.5$ | 0.00 | 0.96 | 0.99 | 0.04 | − 0.16 | 0.96 | 0.99 | 2.09 (0.23, 3.06) |
| | $\beta = 1.0$ | − 0.01 | 0.97 | 0.92 | 0.03 | − 0.09 | 0.97 | 0.93 | 2.09 (0.23, 3.06) |
| $\sigma_x = 2$ | $\beta = 0.0$ | 0.01 | 0.95 | 1.00 | 0.05 | − 0.23 | 0.95 | 1.00 | 2.09 (0.23, 3.06) |
| | $\beta = 0.5$ | − 0.01 | 0.95 | 1.00 | 0.05 | − 0.23 | 0.95 | 1.00 | 2.09 (0.23, 3.06) |
| | $\beta = 1.0$ | − 0.03 | 0.98 | 0.52 | 0.02 | 0.05 | 0.98 | 0.52 | 2.09 (0.23, 3.06) |
| *Simple Randomization* | | | | | | | | | |
| $\sigma_x = 0.5$ | $\beta = 0.0$ | − 0.02 | 0.95 | 1.00 | 0.05 | 0.17 | 0.95 | 1.00 | 2.99 (0.30, 8.60) |
| | $\beta = 0.5$ | − 0.06 | 0.95 | 1.00 | 0.05 | 0.12 | 0.95 | 1.00 | 2.99 (0.30, 8.60) |
| | $\beta = 1.0$ | − 0.10 | 0.95 | 0.99 | 0.05 | 0.06 | 0.95 | 0.99 | 2.99 (0.30, 8.60) |
| $\sigma_x = 1$ | $\beta = 0.0$ | − 0.02 | 0.95 | 1.00 | 0.05 | 0.17 | 0.95 | 1.00 | 2.99 (0.30, 8.60) |
| | $\beta = 0.5$ | − 0.10 | 0.95 | 0.99 | 0.05 | 0.06 | 0.95 | 0.99 | 2.99 (0.30, 8.60) |
| | $\beta = 1.0$ | − 0.17 | 0.95 | 0.92 | 0.05 | − 0.04 | 0.95 | 0.92 | 2.99 (0.30, 8.60) |
| $\sigma_x = 2$ | $\beta = 0.0$ | − 0.02 | 0.95 | 1.00 | 0.05 | 0.17 | 0.95 | 1.00 | 2.99 (0.30, 8.60) |
| | $\beta = 0.5$ | − 0.17 | 0.95 | 0.92 | 0.05 | − 0.04 | 0.95 | 0.92 | 2.99 (0.30, 8.60) |
| | $\beta = 1.0$ | − 0.32 | 0.95 | 0.54 | 0.05 | − 0.26 | 0.95 | 0.54 | 2.99 (0.30, 8.60) |

*Note: SD* standard deviation, *%Bias* percent bias, *Cov* coverage of the 95% confidence interval, *Type 1 Error* proportion of type 1 errors under a nominal type 1 error rate of 0.05, *Balance* covariate balance

**Table 11** Simulation results for the main effect of treatments and their interaction with 8 clusters, based on an analysis model that controls for cluster-level covariates

| | | **Main effects** | | | | **Interaction** | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Covariate** | **Degree** | **%Bias** | **Cov** | **Power** | **Type 1** | **%Bias** | **Cov** | **Power** | **Mean balance** |
| **SD** | **Confounding** | | | | **Error** | | | | **(min, max)** |
| *Covariate-constrained randomization: top 10% of balance scores* | | | | | | | | | |
| $\sigma_x = 0.5$ | $\beta = 0.0$ | 0.36 | 0.86 | 0.31 | 0.13 | 0.18 | 0.86 | 0.31 | 2.13 (0.35, 3.58) |
| | $\beta = 0.5$ | 0.36 | 0.86 | 0.31 | 0.13 | 0.18 | 0.86 | 0.31 | 2.13 (0.35, 3.58) |
| | $\beta = 1.0$ | 0.36 | 0.86 | 0.31 | 0.13 | 0.18 | 0.86 | 0.31 | 2.13 (0.35, 3.58) |
| $\sigma_x = 1$ | $\beta = 0.0$ | 0.36 | 0.86 | 0.31 | 0.13 | 0.17 | 0.86 | 0.31 | 2.13 (0.35, 3.58) |
| | $\beta = 0.5$ | 0.36 | 0.86 | 0.31 | 0.13 | 0.17 | 0.86 | 0.31 | 2.13 (0.35, 3.58) |
| | $\beta = 1.0$ | 0.36 | 0.86 | 0.31 | 0.13 | 0.17 | 0.86 | 0.31 | 2.13 (0.35, 3.58) |
| $\sigma_x = 2$ | $\beta = 0.0$ | 0.37 | 0.86 | 0.31 | 0.13 | 0.18 | 0.86 | 0.31 | 2.13 (0.35, 3.58) |
| | $\beta = 0.5$ | 0.37 | 0.86 | 0.31 | 0.13 | 0.18 | 0.86 | 0.31 | 2.13 (0.35, 3.58) |
| | $\beta = 1.0$ | 0.37 | 0.86 | 0.31 | 0.13 | 0.18 | 0.86 | 0.31 | 2.13 (0.35, 3.58) |
| *Covariate-constrained randomization: top 20% of balance scores* | | | | | | | | | |
| $\sigma_x = 0.5$ | $\beta = 0.0$ | − 0.61 | 0.86 | 0.30 | 0.13 | 0.15 | 0.86 | 0.30 | 2.56 (0.35, 3.79) |
| | $\beta = 0.5$ | − 0.61 | 0.86 | 0.30 | 0.13 | 0.15 | 0.86 | 0.30 | 2.56 (0.35, 3.79) |
| | $\beta = 1.0$ | − 0.61 | 0.86 | 0.30 | 0.13 | 0.15 | 0.86 | 0.30 | 2.56 (0.35, 3.79) |
| $\sigma_x = 1$ | $\beta = 0.0$ | − 0.61 | 0.86 | 0.30 | 0.13 | 0.15 | 0.86 | 0.30 | 2.56 (0.35, 3.79) |
| | $\beta = 0.5$ | − 0.61 | 0.86 | 0.30 | 0.13 | 0.15 | 0.86 | 0.30 | 2.56 (0.35, 3.79) |
| | $\beta = 1.0$ | − 0.61 | 0.86 | 0.30 | 0.13 | 0.15 | 0.86 | 0.30 | 2.56 (0.35, 3.79) |
| $\sigma_x = 2$ | $\beta = 0.0$ | − 0.61 | 0.86 | 0.30 | 0.13 | 0.15 | 0.86 | 0.30 | 2.56 (0.35, 3.79) |
| | $\beta = 0.5$ | − 0.61 | 0.86 | 0.30 | 0.13 | 0.15 | 0.86 | 0.30 | 2.56 (0.35, 3.79) |
| | $\beta = 1.0$ | − 0.61 | 0.86 | 0.30 | 0.13 | 0.15 | 0.86 | 0.30 | 2.56 (0.35, 3.79) |
| *Covariate-constrained randomization: top 50% of balance scores* | | | | | | | | | |
| $\sigma_x = 0.5$ | $\beta = 0.0$ | − 0.41 | 0.86 | 0.30 | 0.14 | − 0.98 | 0.86 | 0.29 | 3.34 (0.50, 4.80) |
| | $\beta = 0.5$ | − 0.41 | 0.86 | 0.30 | 0.14 | − 0.98 | 0.86 | 0.29 | 3.34 (0.50, 4.80) |
| | $\beta = 1.0$ | − 0.41 | 0.86 | 0.30 | 0.14 | − 0.98 | 0.86 | 0.29 | 3.34 (0.50, 4.80) |
| $\sigma_x = 1$ | $\beta = 0.0$ | − 0.42 | 0.86 | 0.30 | 0.14 | − 0.97 | 0.86 | 0.29 | 3.34 (0.50, 4.80) |
| | $\beta = 0.5$ | − 0.42 | 0.86 | 0.30 | 0.14 | − 0.97 | 0.86 | 0.29 | 3.34 (0.50, 4.80) |
| | $\beta = 1.0$ | − 0.42 | 0.86 | 0.30 | 0.14 | − 0.97 | 0.86 | 0.29 | 3.34 (0.50, 4.80) |
| $\sigma_x = 2$ | $\beta = 0.0$ | − 0.43 | 0.86 | 0.30 | 0.14 | − 0.96 | 0.86 | 0.29 | 3.34 (0.50, 4.80) |
| | $\beta = 0.5$ | − 0.43 | 0.86 | 0.30 | 0.14 | − 0.96 | 0.86 | 0.29 | 3.34 (0.50, 4.80) |
| | $\beta = 1.0$ | − 0.43 | 0.86 | 0.30 | 0.14 | − 0.96 | 0.86 | 0.29 | 3.34 (0.50, 4.80) |
| *Simple randomization* | | | | | | | | | |
| $\sigma_x = 0.5$ | $\beta = 0.0$ | − 0.66 | 0.87 | 0.28 | 0.13 | − 0.92 | 0.87 | 0.28 | 4.50 (0.49, 9.42) |
| | $\beta = 0.5$ | − 0.66 | 0.87 | 0.28 | 0.13 | − 0.92 | 0.87 | 0.28 | 4.50 (0.49, 9.42) |
| | $\beta = 1.0$ | − 0.66 | 0.87 | 0.28 | 0.13 | − 0.92 | 0.87 | 0.28 | 4.50 (0.49, 9.42) |
| $\sigma_x = 1$ | $\beta = 0.0$ | − 0.66 | 0.87 | 0.28 | 0.13 | − 0.92 | 0.87 | 0.28 | 4.50 (0.49, 9.42) |
| | $\beta = 0.5$ | − 0.66 | 0.87 | 0.28 | 0.13 | − 0.92 | 0.87 | 0.28 | 4.50 (0.49, 9.42) |
| | $\beta = 1.0$ | − 0.66 | 0.87 | 0.28 | 0.13 | − 0.92 | 0.87 | 0.28 | 4.50 (0.49, 9.42) |
| $\sigma_x = 2$ | $\beta = 0.0$ | − 0.66 | 0.87 | 0.28 | 0.13 | − 0.92 | 0.87 | 0.28 | 4.50 (0.49, 9.42) |
| | $\beta = 0.5$ | − 0.66 | 0.87 | 0.28 | 0.13 | − 0.92 | 0.87 | 0.28 | 4.50 (0.49, 9.42) |
| | $\beta = 1.0$ | − 0.66 | 0.87 | 0.28 | 0.13 | − 0.92 | 0.87 | 0.28 | 4.50 (0.49, 9.42) |

*Note: SD* standard deviation, *%Bias* percent bias, *Cov* coverage of the 95% confidence interval, *Type 1 Error* proportion of type 1 errors under a nominal type 1 error rate of 0.05, *Balance* covariate balance

**Table 12** Simulation results for the main effect of treatments and their interaction with 12 clusters, based on an analysis model that controls for cluster-level covariates

| | | **Main effects** | | | | **Interaction** | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Covariate** | **Degree** | **%Bias** | **Cov** | **Power** | **Type 1** | **%Bias** | **Cov** | **Power** | **Mean balance** |
| **SD** | **Confounding** | | | | **Error** | | | | **(min, max)** |
| *Covariate-constrained randomization: top 10% of balance scores* | | | | | | | | | |
| $\sigma_x = 0.5$ | $\beta = 0.0$ | – 0.16 | 0.95 | 0.98 | 0.05 | 0.36 | 0.95 | 0.98 | 1.26 (0.11, 1.92) |
| | $\beta = 0.5$ | – 0.16 | 0.95 | 0.98 | 0.05 | 0.36 | 0.95 | 0.98 | 1.26 (0.11, 1.92) |
| | $\beta = 1.0$ | – 0.16 | 0.95 | 0.98 | 0.05 | 0.36 | 0.95 | 0.98 | 1.26 (0.11, 1.92) |
| $\sigma_x = 1$ | $\beta = 0.0$ | – 0.16 | 0.95 | 0.98 | 0.05 | 0.36 | 0.95 | 0.98 | 1.26 (0.11, 1.92) |
| | $\beta = 0.5$ | – 0.16 | 0.95 | 0.98 | 0.05 | 0.36 | 0.95 | 0.98 | 1.26 (0.11, 1.92) |
| | $\beta = 1.0$ | – 0.16 | 0.95 | 0.98 | 0.05 | 0.36 | 0.95 | 0.98 | 1.26 (0.11, 1.92) |
| $\sigma_x = 2$ | $\beta = 0.0$ | – 0.17 | 0.95 | 0.98 | 0.05 | 0.36 | 0.95 | 0.98 | 1.26 (0.11, 1.92) |
| | $\beta = 0.5$ | – 0.17 | 0.95 | 0.98 | 0.05 | 0.36 | 0.95 | 0.98 | 1.26 (0.11, 1.92) |
| | $\beta = 1.0$ | – 0.17 | 0.95 | 0.98 | 0.05 | 0.36 | 0.95 | 0.98 | 1.26 (0.11, 1.92) |
| *Covariate-constrained randomization: top 20% of balance scores* | | | | | | | | | |
| $\sigma_x = 0.5$ | $\beta = 0.0$ | – 0.06 | 0.95 | 0.98 | 0.05 | 0.18 | 0.95 | 0.98 | 1.54 (0.19, 2.24) |
| | $\beta = 0.5$ | – 0.06 | 0.95 | 0.98 | 0.05 | 0.18 | 0.95 | 0.98 | 1.54 (0.19, 2.24) |
| | $\beta = 1.0$ | – 0.06 | 0.95 | 0.98 | 0.05 | 0.18 | 0.95 | 0.98 | 1.54 (0.19, 2.24) |
| $\sigma_x = 1$ | $\beta = 0.0$ | – 0.05 | 0.95 | 0.98 | 0.05 | 0.18 | 0.95 | 0.98 | 1.54 (0.19, 2.24) |
| | $\beta = 0.5$ | – 0.05 | 0.95 | 0.98 | 0.05 | 0.18 | 0.95 | 0.98 | 1.54 (0.19, 2.24) |
| | $\beta = 1.0$ | – 0.05 | 0.95 | 0.98 | 0.05 | 0.18 | 0.95 | 0.98 | 1.54 (0.19, 2.24) |
| $\sigma_x = 2$ | $\beta = 0.0$ | – 0.06 | 0.95 | 0.98 | 0.05 | 0.18 | 0.95 | 0.98 | 1.54 (0.19, 2.24) |
| | $\beta = 0.5$ | – 0.06 | 0.95 | 0.98 | 0.05 | 0.18 | 0.95 | 0.98 | 1.54 (0.19, 2.24) |
| | $\beta = 1.0$ | – 0.06 | 0.95 | 0.98 | 0.05 | 0.18 | 0.95 | 0.98 | 1.54 (0.19, 2.24) |
| *Covariate-constrained randomization: top 50% of balance scores* | | | | | | | | | |
| $\sigma_x = 0.5$ | $\beta = 0.0$ | – 0.11 | 0.95 | 0.96 | 0.05 | – 0.24 | 0.95 | 0.96 | 2.09 (0.23, 3.06) |
| | $\beta = 0.5$ | – 0.11 | 0.95 | 0.96 | 0.05 | – 0.24 | 0.95 | 0.96 | 2.09 (0.23, 3.06) |
| | $\beta = 1.0$ | – 0.11 | 0.95 | 0.96 | 0.05 | – 0.24 | 0.95 | 0.96 | 2.09 (0.23, 3.06) |
| $\sigma_x = 1$ | $\beta = 0.0$ | – 0.11 | 0.95 | 0.96 | 0.05 | – 0.24 | 0.95 | 0.96 | 2.09 (0.23, 3.06) |
| | $\beta = 0.5$ | – 0.11 | 0.95 | 0.96 | 0.05 | – 0.24 | 0.95 | 0.96 | 2.09 (0.23, 3.06) |
| | $\beta = 1.0$ | – 0.11 | 0.95 | 0.96 | 0.05 | – 0.24 | 0.95 | 0.96 | 2.09 (0.23, 3.06) |
| $\sigma_x = 2$ | $\beta = 0.0$ | – 0.11 | 0.95 | 0.96 | 0.05 | – 0.24 | 0.95 | 0.96 | 2.09 (0.23, 3.06) |
| | $\beta = 0.5$ | – 0.11 | 0.95 | 0.96 | 0.05 | – 0.24 | 0.95 | 0.96 | 2.09 (0.23, 3.06) |
| | $\beta = 1.0$ | – 0.11 | 0.95 | 0.96 | 0.05 | – 0.24 | 0.95 | 0.96 | 2.09 (0.23, 3.06) |
| *Simple randomization* | | | | | | | | | |
| $\sigma_x = 0.5$ | $\beta = 0.0$ | 0.00 | 0.95 | 0.93 | 0.05 | 0.10 | 0.95 | 0.93 | 2.99 (0.30, 8.60) |
| | $\beta = 0.5$ | 0.00 | 0.95 | 0.93 | 0.05 | 0.10 | 0.95 | 0.93 | 2.99 (0.30, 8.60) |
| | $\beta = 1.0$ | 0.00 | 0.95 | 0.93 | 0.05 | 0.10 | 0.95 | 0.93 | 2.99 (0.30, 8.60) |
| $\sigma_x = 1$ | $\beta = 0.0$ | 0.00 | 0.95 | 0.93 | 0.05 | 0.10 | 0.95 | 0.93 | 2.99 (0.30, 8.60) |
| | $\beta = 0.5$ | 0.00 | 0.95 | 0.93 | 0.05 | 0.10 | 0.95 | 0.93 | 2.99 (0.30, 8.60) |
| | $\beta = 1.0$ | 0.00 | 0.95 | 0.93 | 0.05 | 0.10 | 0.95 | 0.93 | 2.99 (0.30, 8.60) |
| $\sigma_x = 2$ | $\beta = 0.0$ | 0.00 | 0.95 | 0.93 | 0.05 | 0.10 | 0.95 | 0.93 | 2.99 (0.30, 8.60) |
| | $\beta = 0.5$ | 0.00 | 0.95 | 0.93 | 0.05 | 0.10 | 0.95 | 0.93 | 2.99 (0.30, 8.60) |
| | $\beta = 1.0$ | 0.00 | 0.95 | 0.93 | 0.05 | 0.10 | 0.95 | 0.93 | 2.99 (0.30, 8.60) |

*Note: SD* standard deviation, *%Bias* percent bias, *Cov* coverage of the 95% confidence interval, *Type 1 Error* proportion of type 1 errors under a nominal type 1 error rate of 0.05, *Balance* covariate balance

**Abbreviations**

| | |
|---|---|
| CRT | Cluster randomized trial |
| CR | Covariate-constrained randomization |
| BEGIN | Behavioral Nudges for Diabetes Prevention study |
| T2D | Type 2 diabetes |
| ICC | Intra-cluster correlation |
| SD | Standard deviation |
| Var | Variance |
| Cov | Coverage of the 95% confidence interval |
| MSE | Mean squared error |
| CI | Confidence interval |

**Acknowledgements**

The authors wish to thank the three reviewers whose comments greatly improved the manuscript.

**Authors' contributions**

JS and MO conceived and designed the work. JS and ZL performed simulations and data analysis and wrote the manuscript. MO critically reviewed the manuscript for important intellectual content. All authors read and approved the final manuscript.

**Funding**

This work was supported by the National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health (R18-DK123375). The funder had no role in the BEGIN study design; the collection, analysis, and interpretation of the data; the writing of this report; or in the decision to submit the article for publication.

**Availability of data and materials**

The data analyzed in this manuscript are listed in Table 2.

## Declarations

**Ethics approval and consent to participate**

The BEGIN trial received approval by the Northwestern University Institutional Review Board and is registered under the protocol NCT04869917 at ClinicalTrials.gov.

**Consent for publication**

Not applicable.

**Competing interests**

The authors declare that they have no competing interests.

Received: 19 December 2023 Accepted: 21 August 2024
Published online: 06 September 2024

**References**

1. Murray DM, Taljaard M, Turner EL, George SM. Essential ingredients and innovations in the design and analysis of group-randomized trials. Annu Rev Public Health. 2020;41(1):1–19.
2. Giraudeau B, Ravaud P. Preventing bias in cluster randomised trials. PLoS Med. 2009;6(5):e1000065.
3. Ivers NM, Halperin IJ, Barnsley J, Grimshaw JM, Shah BR, Tu K, et al. Allocation techniques for balance at baseline in cluster randomized trials: a methodological review. Trials. 2012;13(1):1–9.
4. Dziak JJ, Nahum-Shani I, Collins LM. Multilevel factorial experiments for developing behavioral interventions: power, sample size, and resource considerations. Psychol Methods. 2012;17(2):153.
5. Moerbeek M, van Schie S. How large are the consequences of covariate imbalance in cluster randomized trials: a simulation study with a continuous outcome and a binary covariate at the cluster level. BMC Med Res Methodol. 2016;16(1):1–10.
6. Li F, Lokhnygina Y, Murray DM, Heagerty PJ, DeLong ER. An evaluation of constrained randomization for the design and analysis of group-randomized trials. Stat Med. 2016;35(10):1565–79.
7. Raab GM, Butcher I. Balance in cluster randomized trials. Stat Med. 2001;20(3):351–65.
8. Yu H, Li F, Gallis JA, Turner EL. cvcrand: A Package for Covariate-constrained Randomization and the Clustered Permutation Test for Cluster Randomized Trials. The R J. 2019;9(2):191–204.
9. Zhou Y, Turner EL, Simmons RA, Li F. Constrained randomization and statistical inference for multi-arm parallel cluster randomized controlled trials. Stat Med. 2022;41(10):1862–83.
10. Watson SI, Girling A, Hemming K. Design and analysis of three-arm parallel cluster randomized trials with small numbers of clusters. Stat Med. 2021;40(5):1133–46.
11. Ciolino JD, Diebold A, Jensen JK, Rouleau GW, Koloms KK, Tandon D. Choosing an imbalance metric for covariate-constrained randomization in multiple-arm cluster-randomized trials. Trials. 2019;20(1):1–10.
12. Moulton LH. Covariate-based constrained randomization of group-randomized trials. Clin Trials. 2004;1(3):297–305.
13. Al-Jaishi A, Dixon S, Garg AX. Experimental Designs and Randomization Schemes: Covariate-Constrained Randomization. In: Rethinking Clinical Trials: A Living Textbook of Pragmatic Clinical Trials. Bethesda: NIH Pragmatic Trials Collaboratory; 2024.
14. Collins LM, Dziak JJ, Li R. Design of experiments with multiple independent variables: a resource management perspective on complete and reduced factorial designs. Psychol Methods. 2009;14(3):202.
15. Vargas MC, Pineda GJ, Talamantes V, Toledo MJL, Owen A, Carcamo P, et al. Design and rationale of behavioral nudges for diabetes prevention (BEGIN): A pragmatic, cluster randomized trial of text messaging and a decision aid intervention for primary care patients with prediabetes. Contemp Clin Trials. 2023;130:107216.
16. Menke A, Casagrande S, Geiss L, Cowie CC. Prevalence of and trends in diabetes among adults in the United States, 1988–2012. JAMA. 2015;314(10):1021–9.
17. Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med. 2002;346(6):393–403.
18. Hing E, Uddin S. Visits to primary care delivery sites: United States, 2008. 47. US Department of Health and Human Services, Centers for Disease Control and Prevention; 2010.
19. Kugler KC, Dziak JJ, Trail J. Coding and interpretation of effects in analysis of data from a factorial experiment. In: Collins LM, Kugler KC, editors. Optimization of Behavioral, Biobehavioral, and Biomedical Interventions: Advanced Topics. Cham: Springer International Publishing; 2018. p. 175–205.
20. Kalish LA, Begg CB. Treatment allocation methods in clinical trials: a review. Stat Med. 1985;4(2):129–44.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.