A simulation study to evaluate the effect of constrained randomization for the design and analysis of stepped wedge cluster randomized trials

Peiyan Gao

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2020

Committee:

James P. Hughes

Patrick J. Heagerty

Program Authorized to Offer Degree:

Biostatistics - Public Health

University of Washington

Abstract

A simulation study to evaluate the effect of constrained randomization for the design and analysis of stepped wedge cluster randomized trials

Peiyan Gao

Chair of the Supervisory Committee:

Professor, James P. Hughes

Biostatistics

In this study, we evaluated the effect of constrained randomization on testing the treatment effect in terms of type I error and power with data generated from a stepped wedge cluster-randomized design, under the presence of cluster-level covariates. We considered two cases, one with a single binary covariate and the other with a mixture of continuous and categorical covariates. For case one we used stratified randomization to achieve perfect covariate balance whereas for case two we constrained the randomization space by setting scores based on different balance criteria. For each case we consider eight different scenarios, and apply model-based and permutation-based inference, both adjusted and unadjusted for covariates. We found that the type I error is close to the nominal level most of the time except for permutation inference with constrained randomization and unconstrained analysis, where it drops down towards zero. In general, we see that constrained randomization can slightly increase power when covariates are also included in the analysis phase, and such increase is more visible in case one with a single binary covariate than case two with multiple covariates. Overall, although we discovered some advantages of doing constrained randomization in terms of power, such gain is only marginal and controlling for covariates in the analysis phase is still considered to be a more effective way to attain higher testing power under stepped wedge setting.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

In a stepped wedge cluster randomized (SW-CRT) design, all clusters are randomized into sequences [1] (or time waves [2]). The intervention is initialized in each sequence at different times and will persist until the end of the trial once introduced [3]. Therefore, all clusters start the trial in the control condition and receive the treatment by the end of the trial. A stepped wedge design is called balanced if there are equal number of clusters in each sequence [4], and complete if all clusters have outcomes collected at all time periods. A schematic example of a balanced and complete SW-CRT design is shown in Figure 1.

Stepped wedge cluster-randomized trials have gained increased popularity in recent years in public health research. For instance, Moulton et al. [5] used a randomized SW design to evaluate the effect of training health personnel to conduct TB testing on reducing tuberculosis incidence among the HIV clinic population in Rio de Janeiro; Ji et al. [6] investigated the effectiveness of community-based health insurance(CBHI) schemes to protect individuals against large financial shocks using a SW-CRT design with data from a CBHI roll-out in rural Burkina Faso; Stern et al. [7] conducted a SW-CRT trial to evaluate the clinical and cost effectiveness of enhanced multidisciplinary teams versus 'usual care' for the treatment of pressure ulcers in long term care facilities in Ontario, Canada. The advantages of stepped wedge to parallel cluster randomized trials consist of the following : first, by rolling out the intervention across a period of time, stepped wedge trials can alleviate the financial and logistic burden which would otherwise incur if all participants need to receive the treatment simultaneously. Second, it can be unethical to not assign treatment to some groups in the parallel setting, whereas SW trials ensures everyone will get the treatment by the end [8]. Moreover, SW trials also allow us to study the treatment effect over a period of time by modelling the secular trend of the outcome [1].

There are also some widely discussed issues associated with the design and analysis of SW-CRT, which could potentially hamper its effective implementation and interpretability [9]. The issue we are going to investigate throughout this paper, which is also a common issue in traditional parallel cluster randomized trials, is cluster-level imbalance. For individually randomized trials, simple randomization by randomly permuting treatment assignments rarely results in severe imbalance provided the sample size is large, which tends to balance the difference in individual-level factors between treatment and control [10], hence covariate imbalance rarely draws much concern under the individual level. However, the situation becomes different when each

cluster of individuals receive treatment or control as a whole, explained as follows.

For parallel cluster-randomized trials, observations within a cluster are considered as a single unit as all of them receive the same treatment condition and share common cluster-level characteristics. Therefore, the randomization space becomes much smaller, which dramatically increases the chance of obtaining a randomization with substantial imbalance in one or more cluster-level characteristics across arms [11]. This can make the final estimate of treatment effect difficult to interpret due to confounding by imbalanced characteristics that are prognostic of the outcome [11]. Analysis-based adjustment accounts for such confounding by adjusting for covariates in the model, and often leads to higher precision and power [12]. Alternatively, much of the literature emphasizes the importance of design-based adjustment, which aims to ensure covariate balance by restricting the randomization space to a smaller subset, and some studies have demonstrated additional gain in terms of precision and power by applying constrained randomization. Stratification and pair-matching are commonly used techniques to handle anticipated cluster-level confounders [10, 11, 12, 13, 14] in parallel cluster randomized trials. However, with only a few groups available and multiple relevant covariates to balance, stratification may be infeasible as it tends to form sparse and uneven strata, making the randomization process difficult. Pair-matching can also become hard if there are many factors to balance. Regarding this, Moulton proposed the idea of covariate-based constrained randomization, that is, to set a measurable criterion based on covariate balance and to select a randomization from those allocation schemes meeting this specific criterion [11]. Fan et al. [10, 12] studied both continuous and binary outcomes with multiple binary covariates, and adapted the balance criteria proposed by Raab and Butcher [15] to select the final allocation scheme. Fan et al. found that constrained randomization provides further improvement to the power of both model and permutation based tests even when cluster-level covariates are already controlled for in the analysis phase [10, 12].

For the stepped wedge design, the concept of covariate balance becomes less intuitive, as we can not simply classify each cluster into treatment or control arm since all of them receive both control and treatment. A possible way to evaluate balance in SW setting is to consider each treatment sequence separately, and evaluate the mean difference of cluster-level covariates between the sequences. Lew et al. [2] defined the terms mean balance and sequential balance to seek minimization of covariate differences across all sequences and the trend of covariate values against the cross-over time respectively. Detailed definitions for mean and sequential balance will be

given in the methodology section. The sequential balance score was used to select the final allocation scheme for a Behavioral Health Interdisciplinary Program-Collaborative Care Model(BHIP-CCM) implementation trial, and the author showed that each individual site characteristic became more balanced than simple randomization. Moulton [5] et al. set a restriction on sequences such that the sum of covariate values weighted by the number of months on intervention must be within a certain range of that on control, and this restricted set was used to randomly select the final allocation scheme for the THRio trial. Despite these examples, there seems to be a lack of studies that systematically evaluate the influence of constrained randomization on the inference and testing results under the stepped wedge setting.

In this study, we explore different forms of design-based adjustment for cluster-level covariates, including stratified randomization as well as constrained randomization using different balance scores, and compare them with simple randomization in terms of type I error rate and power. Various data-generating scenarios are used for simulation with varying number of clusters, values of random cluster effect, the existence or absence of random treatment effect as well as the number and type of cluster-level covariates. The linear mixed effect model is fit to the data as model-based inference, whereas the robust inferential procedure derived by Hughes et al. [16] is used as permutation-based inference to estimate and test the treatment effect. It is our particular interest to see whether by controlling for cluster-level imbalance at randomization, higher testing power can be achieved while maintaining the nominal type I error level. In section 2 we illustrate the randomization strategies and procedures, the modelling fitting process as well as different simulation scenarios. In section 3, we present our results comparing the testing level and power among different combination of methods. In section 4, we discuss the implications of our results, comment on further research to be done and make concluding remarks.

2 Methodology

2.1 Outcome data generation

We aim to simulate normally distributed outcomes with an identity link under a cross-sectional stepped wedge setting where new subjects are observed at each time point and the outcome is collected only once from each subject [3, 17]. We also require additional cluster-level covariates to be included, hence we extend the linear

mixed effect model proposed by Hughes et al. [9] to generate our data for analysis:

$$Y_{ijk} = \mu + \beta_j + x_{ij}\theta + \mathbf{z_i}^T\alpha + u_i + v_{ij} + c_i x_{ij} + e_{ijk} \tag{1}$$

$$\begin{pmatrix} u_i \\ c_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^2 & \rho\tau\eta \\ \rho\tau\eta & \eta^2 \end{pmatrix}\right).$$

$$v_{ij} \sim N(0, \gamma^2); e_{ijl} \sim N(0, \sigma^2)$$

where $i = 1...I$ indexes the cluster, $j = 1...J$ indexes the time period and k indexes each individual observation. In our simulations we generate 100 observations under each cluster-time unit so $k = 1...100$. We also require $\frac{I}{J-1}$ to be an integer to allow for equal occurences of each intervention sequence, as illustrated in Figure 1. μ is the fixed intercept term, i.e. the baseline mean for a cluster with covariate profile $\mathbf{z} = \mathbf{0}$ at time 1 under control. β_j with $j = 1...J$ represents the fixed secular trend, and x_{ij} indicates the condition of intervention (1 if treatment, 0 if control) and θ denotes the fixed treatment effect. $\mathbf{z_i} = [z_{i1}...z_{iS}]^T$ stores the values of S different covariates for cluster i, either discrete or continuous, which are assumed to be fixed across time periods, and α is the corresponding coefficient vector. No individual-level covariate is used to generate the outcome $\mathbf{Y}$. The random effects are represented by u_i, v_{ij} and c_i and each follows a Gaussian distribution with mean zero and variance τ^2, γ^2, η^2 respectively. We assume the correlation between the random cluster effect u_i and the random treatment effect c_i is ρ and the correlation between the cluster-specific random time effect v_{ij} and both u_i and c_i are zero. The last term e_{ijk} is the individual-level Gaussian error with mean 0 and standard deviation σ.

		Time period				
		1	**2**	**3**	**4**	**5**
Cluster	**1**	0	1	1	1	1
	2	0	0	1	1	1
	3	0	0	0	1	1
	4	0	0	0	0	1

Figure 1: Schematic representation of a balanced and complete stepped wedge design with 4 clusters and 5 time periods, where 0 indicates control condition and 1 represents treatment. All four clusters receive control at t=1 and each switches to treatment respectively, from t=2 to 5.

2.2 Randomization Procedure

With the covariate matrix $\mathbf{Z} = [\mathbf{z_1}, \mathbf{z_2}...\mathbf{z_I}]$ specified, we need to decide the allocation scheme of treatment sequences to each cluster, i.e. the randomization procedure. We can use either simple or constrained randomization, with the latter expected to provide design-based adjustment on those cluster-level covariates [12]. For simple randomization, the $\frac{I}{J-1}$ copies of each treatment sequence (if balanced design is used) are randomly distributed across the I clusters without taking into account their covariate values, and the randomization space will be the whole possible permutation space. For constrained randomization, certain restrictions will be set on the randomization space in order to balance the cluster-level characteristics among sequences, and such restrictions can take various forms depending on the balance criterion we want to achieve. In the following sections, we will describe four different constraint criteria used for our simulation and analysis, including stratified randomization, treatment vs control balance, mean balance and sequential balance. The latter three are similar in that they all set up a score to measure the level of balance for each allocation scheme. We will also illustrate the procedure to select the final allocation scheme when the latter three randomization strategies are used.

2.2.1 Stratified randomization

Stratified randomization is an easy and efficient way to provide adjustment for factors prognostic to the response, and is widely implemeted under parallel design trials to prevent imbalance between treatment and control groups hence to reduce confounding [13]. In stepped wedge design, every cluster is treated as a single unit, and clusters will be assigned to different strata based on their covariate values $\mathbf{z_i}$. After that treatment sequences will be randomly allocated to each cluster within each stratum separately. Stratification can achieve perfect balance when there is only a single categorical factor with L levels and L satisfies $\frac{I}{L(J-1)} = c$ (c is an integer). This setting allows us to assign c complete sets of the (J-1) intervention sequences to each covariate level, within which simple randomization is performed, so covariate values can be perfectly balanced among treatment sequences and between cluster-time units under treatment versus control.

2.2.2 Treatment vs control balance score

We modify the balance score (B) proposed by Raab and Butcher [15], which was originally applied in the case of parallel cluster randomized trials, to make it compatible with the stepped wedge setting:

$$Imb_{tc} = \sum_{l=1}^{S} \Big(\frac{\sum_{i=1}^{I}(t_i - 1)z_{il}^*}{\sum_{i=1}^{I}(t_i - 1)} - \frac{\sum_{i=1}^{I}(J - t_i + 1)z_{il}^*}{\sum_{i=1}^{I}(J - t_i + 1)} \Big)^2 \tag{2}$$

Where S is the number of cluster-level covariates which are prognostic of the final outcome. Unlike the B score by Raab and Buther which uses the original scale of covariates, here each of the S covariates is rescaled to have zero mean and unit variance across the I clusters so as to eliminate scale heterogeneity in evaluating the balance level, and z_{il}^* is the value of the l^{th} covariate for cluster i after rescaling. Also, t_i indexes the first time period for cluster i to start receiving treatment. Basically, score (2) is constructed in a way that each cluster-time unit is treated separately analogous to each independent cluster under the parallel setting. As can be seen from (2), the first term is actually the mean of the l^{th} covariate values from all cluster-time units assigned to control, whereas the second term is the mean of those assigned to treatment. Unlike Raab and Butcher, we do not set additional weight to each covariate since their values are standardized already and we assume the associations of covariates with the outcome are equal. For balanced stepped wedge design where the number of cluster-time units assigned to treatment is equal to those assigned to control, the formula for score (2) could be further simplified as follows:

$$Imb_{tc}(\mathbf{Z}_{cont.}) = \sum_{l=1}^{S} \Big(\sum_{i=1}^{I}(2t_i - J - 2)z_{il}^* \Big)^2 \tag{3}$$

For categorical covariates, we modify the imbalance score proposed by De Hoop et al. [18] in a similar way as we did for the continuous case, and the expression can be simplified as follows for balanced design:

$$Imb_{tc}(\mathbf{Z}_{cat.}) = \sum_{l=1}^{S} \sum_{\tau} f_{l\tau} \Big(\sum_{i=1}^{I}(2t_i - J - 2)I^*(z_{il} = \tau) \Big)^2 \tag{4}$$

Here in equation (4) the squared differences are taken not only across cluster-level covariates but also across all levels for each covariate, where τ denotes the τ^{th} level of covariate l. We also normalize each level of covariates, which was not done by De Hoop et al. I^* here is the normalized version of the indicator function I, which has value 1 if the level of z_{il} is τ and 0 otherwise. We normalize it here by subtracting its mean across all

clusters, which is just the proportion of clusters with level τ for z_l, then dividing by its corresponding standard deviation. The reason to perform such normalization is similar as in the continuous case, that we want to balance the weight of each level of covariates in terms of their contributions to the final score by putting them into the common scale. An issue here, however, is that categorical covariates with more levels naturally contribute more terms to the balance score. To tackle this, we re-weight each level by its frequency $f_{l\tau}$ satisfying $\sum_\tau f_{l\tau} = 1$ for every $l = 1...S$, so that each covariate still has a total weight 1 hence they contribute equally to the balance score.

For the case with a mixture of continuous and categorical features, it suffices to simply add the scores from (3) and (4) to get the final score without any further reweighting as all of them are already standardized.

2.2.3 Sequential balance score

We adapt the method of sequential balance proposed by Lew et al. [2], which aims to minimize the time trend of cluster-level characteristics with respect to sequences that clusters are assigned to under stepped wedge setting. Formally speaking, the trend is defined by the slope β of the fitted regression line, $z_{il} = \alpha + \beta t_i$, where t_i is the time at which cluster i switches from control to intervention , and z_{il} is the value of the l^{th} covariate. A flat line with the slope estimate $\hat{\beta} = 0$ indicates there is no association between the covariate value of a cluster and how early it starts receiving treatment, hence perfect sequential balance is achieved for covariate l. It was shown by [2] that $\hat{\beta} = 0$ if and only if $\sum_{i=1}^{I} z_{il}(t_i - \bar{t}) = 0$, where z_{il} and t_i has the same meaning as above, and $\bar{t}$ is the mean of t_i across all clusters. The sequential balance score is built by either taking the absolute value or the square of the summation term, and then summing over all the S covariates. Here we narrow down our focus to the square setting, and the score is represented as follows for continuous covariates:

$$Imb_{seq}(\mathbf{Z}_{cont.}) = \sum_{l=1}^{S} \left(\sum_{i=1}^{I} z_{il}^{*}(t_i - \bar{t}) \right)^2 \tag{5}$$

Again we use the standardized version $\mathbf{z_l^*}$ for $l = 1...S$ and assume equal predictive strength of covariates so there is no need to set additional weight for each covariate. For categorical covariates, the same principle of constructing the Imb_{tc} score (4) is applied again here, i.e. to treat each level of a covariate separately and use indicator function to represent its value. Again we weight each level by its frequency as done in (4) and by Lew

et al, and yield the following expression:

$$Imb_{seq}(\mathbf{Z}_{cat.}) = \sum_{l=1}^{S} \sum_{\tau} f_{l\tau} \left(\sum_{i=1}^{I} I^*(z_{il} = \tau)(t_i - \bar{t}) \right)^2 \tag{6}$$

Under a balanced and complete design, it can be shown that the Imb_{tc} score and the Imb_{seq} score only differs by a constant multiplicative factor for both continuous and categorical covariates. Therefore, they will assign exactly the same ranks to each allocation sequences, so we can treat them as a unitary score for analysis.

2.2.4 Mean balance score

Another more stringent constraint criterion mentioned in [2] is called mean balance. Under stepped wedge setting, it aims to minimize the sum of deviations from covariate means of clusters at each sequence to the grand mean, with expression as follows for continuous covariates:

$$Imb_{mean}(\mathbf{Z}_{cont.}) = \sum_{l=1}^{S} \sum_{t=2}^{J} (\bar{z}_{l,t} - \bar{z}_l)^2 \tag{7}$$

Where $\bar{z}_{l,t}$ is the mean of $\mathbf{z_l}$ among clusters that switch at time t and $\bar{z}_l$ is the grand mean of $\mathbf{z_l}$ among all clusters. If standardization is performed on each column of $\mathbf{Z}$ then equation (7) can be reformulated as follows, with the term $\bar{z}_l$ disappears and $\bar{z}^*_{l,t}$ is the standardized value of $\bar{z}_{l,t}$.

$$Imb_{mean}(\mathbf{Z}_{cont.}) = \sum_{l=1}^{S} \sum_{t=2}^{J} (\bar{z}^*_{l,t})^2 \tag{8}$$

The mean balance score for categorical covariates is constructed in similar ways as the treatment v.s. control score and sequential balance score as mentioned previously, that each factor of $\mathbf{z_l}$ is treated separately and we use the indicator function after re-scaling, plus reweighting each level by its frequency $f_{l\tau}$.

$$Imb_{mean}(\mathbf{Z}_{cat.}) = \sum_{l=1}^{S} \sum_{\tau} f_{l\tau} \sum_{t=2}^{J} \left(\bar{I}^*(\mathbf{z}_{l,t} = \tau) \right)^2 \tag{9}$$

where $\bar{I}^*(\mathbf{z}_{l,t} = \tau)$ is the mean of the rescaled indicator function at level τ of the l^{th} covariate among clusters that switch at t.

It is easily seen that perfect mean balance implies perfect sequential balance but not vice versa. For instance, consider the setting with 3 clusters and 4 time periods. We denotes the clusters as A, B, C, which start receiving treatment at time points 2, 3, 4 respectively, and we assume there's only one prognostic covariate z with value equals to 100 for all three clusters. In this situation, both perfect mean balance and perfect sequential balance

are satisfied simultaneously. However, if **z** has values 100, 40, 100 for clusters A, B, C instead, then sequential balance is still preserved as the slope of **z** against sequences is still 0, whereas mean balance is not preserved as all three values differ from the grand mean which is 80. In practice covariate values could rarely be constant across all clusters and perfect balance can not be achieved, hence these two balance criteria are expected to result in different ranking regimes.

2.2.5 Allocation of treatment sequences

Once the balance score is specified, we can label each of the sequence allocation schemes with their scores, sort them and then select the final candidate set. The first step is to enumerate all possible allocations. This is feasible when the number of clusters and time periods are relatively small. For instance, when $I = 8$ and $J = 5$, there will be two complete sets of four intervention sequences, hence in total $\binom{8}{2\,2\,2\,2} = 2520$ possible allocation schemes. However, when I and J become larger it will become formidable to enumerate them altogether. In such cases, we adapt the simulation method proposed by Li et al [10] to randomly select 20000 randomization schemes and remove duplicates among them to obtain the initial candidate set which mimics the entire randomization space. We then select those schemes located within the first q^{th} percentile ordered by the pre-specified balance score in ascending order, and take them altogether as the final candidate set, from which we will randomly pick one, and assign treatment sequences to each cluster accordingly.

The parameter q is set in advance with range (0,1] which controls the final candidate set size and hence the tightness of the constraint. $q = 1$ corresponds to simple randomization as we are selecting from all the candidate schemes without any deletions, whereas smaller value of q indicates tighter constraint as fewer schemes are included with the least imbalance scores. By varying the value of q, we are able to evaluate the effects of design-based adjustment with various levels of constraint on the final outcomes of our interest, i.e. the empirical type I error rate and power for testing the treatment effect θ.

2.3 Inference on the treatment effect

We apply model-based and permutation-based inference to estimate and test the fixed treatment effect θ. For model-based inference, we always fit the correct model that is used to generate the data (except for the inclusion or exclusion of the covariate term as mentioned later). Since our primary focus is on the joint effect of

randomization and analysis procedure, we want to eliminate issues of model misspecification from our initial evaluation. For model-based inference, we implement the R function lmer [19]. This function fits the linear mixed-effect model and finds the optimal $\hat{\theta}$ which minimizes the restricted maximum likelihood criterion. Then an F-test is used with Satterthwaite denominator degrees of freedom to test the significance of $\hat{\theta}$. For adjusted analysis, we account for the covariate vector $\mathbf{z}$ and include it as fixed-effect terms into the lmer function, hence the model to be fitted is the same as model (1) which is used to generate the data. For unadjusted analysis, we still fit the data as generated by model (1) but do not account for covariates in the model fitting process, that is we would include all fixed and random effect terms as specified in model (1) except $\mathbf{z}$ into lmer.

For permutation-based inference, we use the method derived by Hughes et al [16], which does not require the correct model structure to be specified and therefore is considered more robust. Based on a permutation argument, it yields the following estimator of treatment effect θ:

$$\tilde{\theta} = \frac{\sum_{ij} Y_{ij}(x_{ij} - \bar{x}_j)}{N \sum_j \bar{x}_j(1 - \bar{x}_j)} \tag{10}$$

It was shown that $\tilde{\theta}$ is unbiased even if the true covariance structure of $\mathbf{Y}$ is ignored. Besides, by using similar permutation arguments on the variance of $\tilde{\theta}$, the following unbiased variance estimate was also obtained:

$$\begin{aligned} V_\theta^1(\tilde{\theta}) = \{\sum_i [\sum_j (Y_{ij} - x_{ij}\theta)^2 \bar{x}_j(1 - \bar{x}_j) + 2 \sum_{j<j'} (Y_{ij} - x_{ij}\theta)(Y_{ij'} - x_{ij'}\theta)\bar{x}_j(1 - \bar{x}_{j'})] \\ - \frac{2}{N-1} \sum_{i<i'} \sum_{j,j'} (Y_{ij} - x_{ij}\theta)(Y_{i'j'} - x_{i'j'}\theta)\bar{x}_{min(j.j')}(1 - \bar{x}_{max(j.j')})\}/ \\ (N \sum_j \bar{x}_j(1 - \bar{x}_j))^2 \end{aligned} \tag{11}$$

When testing the hypothesis $H_0 : \theta = \theta_o$, we can plug in θ_o for the unknown θ in the expression of $V_\theta^1(\tilde{\theta})$, and use the Z statistics of $\frac{\tilde{\theta} - \theta_o}{V_{\theta_o}^1}$ to test H_0, assuming the central limit theorem holds.

Here, we refer to the direct use of $\tilde{\theta}$ and $V_\theta^1(\tilde{\theta})$ for testing θ as the unstratified analysis, since it doesn't account for the heterogeneity in fixed effect terms introduced by the covariate vector $\mathbf{z}$. However, the validity of the permutation-based method does require values of fixed effect terms (except for the treatment effect) to be constant within each time period across clusters, as argued in [16]. To do stratified analysis, permutations should be done within each covariate stratum separately, in which the assumption of within-period constancy of fixed effects holds, and the final estimate of θ and its variance are the weighted average of those within-strata

estimates, with the closed-form expression listed below:

$$\tilde{\theta} = \frac{\sum_{hij} Y_{hij}(x_{hij} - \bar{x}_{hj})}{\sum_h N_h \sum_j \bar{x}_{hj}(1 - \bar{x}_{hj})} \tag{12}$$

$$V_\theta^1(\tilde{\theta}) = \{\sum_{hi}[\sum_j (Y_{hij} - x_{hij}\theta)^2 \bar{x}_{hj}(1 - \bar{x}_{hj}) + 2\sum_{j<j'}(Y_{hij} - x_{hij}\theta)(Y_{hij'} - x_{hij'}\theta)\bar{x}_{hj}(1 - \bar{x}_{hj'})]$$
$$-\frac{2}{N_h - 1}\sum_h \sum_{i<i'} \sum_{j,j'} (Y_{hij} - x_{hij}\theta)(Y_{hi'j'} - x_{hi'j'}\theta)\bar{x}_{h,min(j.j')}(1 - \bar{x}_{h,max(j,j')})\}/ \tag{13}$$
$$(\sum_h N_h \sum_j \bar{x}_{hj}(1 - \bar{x}_{hj}))^2$$

where Y_{hij} represents the observation on the i'th cluster in stratum h at time j, and similarly for x_{hij} and $\bar{x}_{hj}$, and N_h is the number clusters in strata h. For the rest of the paper we will use adjusted and unadjusted analyses to indicate whether the permutation is stratified within each covariate stratum to calculate $\tilde{\theta}$ so that the notation is consistent under both model and permutation-based inference.

2.3.1 Simulation procedures and scenarios

Simulations are repeated 5000 times for each scenario we consider, allowing the type I error and power estimates to vary within ± 0.006 and ± 0.014 respectively due to Monte Carlo variation. The simulation process is outlined here: first we generate the covariate matrix **Z** according to a pre-specified distribution, then the constraints, if any, are imposed to obtain the final candidate set of possible randomizations, **C**. After these preparation steps are finished, we randomly draw one scheme from **C** to determine the values of x_{ij} for each cluster, then the outcome **Y** is generated according to model (1) with all necessary parameters specified by the scenario. Lastly we apply adjusted and unadjusted analyses under both model-based and permutation-based inference to yield the estimated treatment effects, their standard errors and associated testing results. After 5000 such simulations, the empirical distribution of our estimates with the associated Type I error rates/power are then obtained.

We first investigate the case with single binary covariate (i.e. $\mathbf{z_i}$ has only one element taking values 0 or 1 and L=2) and we always set the number of clusters at each level to be equal for each simulation. The values I and J will be set to ensure perfect balance can be achieved by stratifying on each level of **z**, as illustrated in section 1.2.1. We then consider the case with three cluster-level covariates $[\mathbf{Z_1}, \mathbf{Z_2}, \mathbf{Z_3}]$, where $\mathbf{Z_1}$ is continuous and randomly drawn from $Unif(-1, 1)$, and $\mathbf{Z_2}, \mathbf{Z_3}$ are categorical factors following $Multinoulli(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and $Bernoulli(\frac{1}{2})$ respectively. This setting is an analogy of a typical stepped wedge trial in reality, where each

Table 1: Simulation scenarios for both settings. Data are generated according to model (1) with varying parameters as listed below.

Simulation scenarios	I - number of clusters	J - number of time periods	τ- SD of random cluster effect	γ - SD of random cluster-time effect	η - SD of random treatment effect	ρ - correlation between τ and η	σ - SD of random noise	θ_1 - treatment effect under H_0
1	8	5	0.1	0.1	0	0	1	0.25
2	8	5	0.1	0.1	0.1	0.1	1	0.25
3	8	5	0.3	0.1	0	0	1	0.25
4	8	5	0.3	0.1	0.1	0.1	1	0.25
5	24	5	0.1	0.1	0	0	1	0.15
6	24	5	0.1	0.1	0.1	0.1	1	0.15
7	24	5	0.3	0.1	0	0	1	0.15
8	24	5	0.3	0.1	0.1	0.1	1	0.15

clinic is enrolled as a cluster and has a mixture of continuous and categorical features, for instance, the number of beds, size (large,medium,small) and location (rural,urban). Under both cases, We will consider scenarios (as summarised in Table 1) with $I = (8, 24)$, $\tau = (0.1, 0.3)$ and $\eta = (0, 0.01)$ in model (1). when η is non-zero we set ρ, the correlation between random treatment and random cluster effect, to be 0.1. The scenarios are set in this way for easy comparisons between cases of different cluster numbers, magnitudes of intra-class correlation(ICC) and the existence of random treatment effects. Besides, we create a linear time trend by setting $\beta = (0, 1, 2, 3, 4)$, and the values of J, μ, γ, σ are set to 5, 0, 0.1, 1 respectively, and will be kept fixed across scenarios. The treatment effect θ is set to 0.25 when $I = 8$ and 0.15 when $I = 24$, to make the power lies within reasonable ranges for easy comparisons between scenarios. For the first case, we fix α to be 1, and compare simple vs stratified randomization and apply both model-based and permutation-based inference on θ. For the second case, α is a vector of length 4: $[1, 1, 2, 1]$, corresponding to $\mathbf{Z_1}$, and the dummy representations of $\mathbf{Z_2}$ and $\mathbf{Z_3}$ respectively. For each simulation with the set of covariates been drawn, we consider a grid values of q ranging from 0.1 to 1 corresponding to different candidate set sizes to compare simple randomization with the constraint strategies based on balance scores as illustrated in section 2.2. To estimate and test for θ, we apply model-based inference, both adjusted and unadjusted, and unadjusted permutation inference only, as in this case it is impossible to stratify clusters by covariate values.

3 Results

3.1 Single binary covariate

Table 2 summarizes the results on type 1 error rates of testing $H_0 : \theta = 0$ for scenarios 1 to 8, with model and permutation-based inference, adjusted and unadjusted analyses been conducted under both simple and stratified

randomization. For model-based inference, adjusted analysis tends to be slightly more anti-conservative than its unadjusted counterpart most of the times, though the type I errors still lie within the 95% error margin centered around the nominal value except for adjusted analysis in scenario 1, and such inflation in type I error is only marginal and not considered to be devastating. For permutation-based inference under scenarios 1 to 4, we constantly observe conservative type I errors for adjusted analyses except for scenario 2 under stratified randomization, nonetheless all adjusted analyses approach the nominal level under the other four scenarios. Another noticeable pattern is that unadjusted analysis with stratified randomization always yields zero type 1 error under permutation-based inference, which indicates the testing under this setting fails and will be discarded for the following power analysis.

Table 3 shows the corresponding power for testing $H_0 : \theta = 0$ versus $H_0 : \theta = \theta_a$ under the same setting as table 2. Several distinct patterns can be found for model-based inference. First, although the difference is only marginal, adjusted analysis does consistently achieve higher power than unadjusted analysis under each pair of comparison within the same scenario and randomization scheme. Another pattern occurs with the power comparison between simple and stratified randomization: with covariates having been adjusted in the analysis phase, stratified randomization provides additional power gain across all scenarios, except for scenario 4 where simple randomization yields slightly higher power(0.76 vs 0.756). However, the increase in power is smaller than the gain by switching from unadjusted to adjusted analysis. Another pattern is that stratified randomization with unadjusted analysis yields lower power than simple randomization with adjusted analysis in all scenarios, and sometimes the power is even lower than simple randomization with unadjusted analysis.

For permutation-based inference, the highest power is achieved under stratified randomization and adjusted analysis in all scenarios. For unadjusted analysis, although the power stays at 0 under all scenarios with stratified randomization, this should not be drawn much attention as the test is already shown to be invalid by table 2. However, it is worth noting that in every scenario, there is also a dramatic drop in power by switching from adjusted to unadjusted analysis under simple randomization, even if the test maintains the nominal type 1 error rate. Finally, with other settings the same, model-based inference always achieves higher power than permutation-based approach, and such gap becomes larger as τ rises from 0.1 to 0.3 due to the significant power loss under permutation-based inference with adjusted analysis.

Figure 2 shows the box plots of $\hat{\theta}$ obtained by all combinations of settings under scenario 1 to 4 with $\theta = 0.25$. All settings yield unbiased estimates of θ, as denoted by the red dashed line. The variance of $\hat{\theta}$ is low across scenarios with model-based inference, and close to each other with adjusted and unadjusted analyses. For permutation-based inference, the distribution of $\hat{\theta}$ gets more disperse with adjusted analysis and simple randomization in scenarios 3 and 4, which matches with the power loss at corresponding entries in table 3. It can also be seen that the variance of $\hat{\theta}$ with simple randomization and unadjusted analysis is much higher than the others, which is also consistent with the zero power shown in table 3.

Table 2: Type I error rates for testing the null hypothesis $H_0 : \theta = 0$. Data are simulated from equation (1) with a single binary covariate under scenarios 1 to 8, with simple vs stratified randomization, adjusted vs adjusted analyses, model-based vs permutation-based inference been performed. The covariate coefficient α is fixed to be 1.

Scenario	**Adjusted for covariate in analysis**	**Model-based inference**		**Permutation-based inference**	
		Simple randomization	**Stratified randomization**	**Simple randomization**	**Stratified randomization**
1	Yes	0.063	0.061	0.039	0.038
	No	0.052	0.048	0.045	0
2	Yes	0.052	0.052	0.038	0.047
	No	0.044	0.044	0.048	0
3	Yes	0.055	0.053	0.034	0.033
	No	0.053	0.048	0.045	0
4	Yes	0.049	0.047	0.030	0.035
	No	0.046	0.044	0.046	0
5	Yes	0.054	0.051	0.049	0.052
	No	0.052	0.043	0.046	0
6	Yes	0.051	0.059	0.049	0.053
	No	0.050	0.053	0.053	0
7	Yes	0.052	0.046	0.050	0.043
	No	0.051	0.044	0.044	0.002
8	Yes	0.050	0.057	0.045	0.047
	No	0.050	0.054	0.054	0

3.2 A mixture of continuous and categorical covariates

Figures 3 and 4 summarize the results on type I errors for testing $H_0 : \theta = 0$ under all scenarios with a mixture of continuous and categorical cluster-level covariates, by varying the level of constraint q using sequential (treatment v.s. control) balance and mean balance as the constraint criterion respectively. Model-based inference, both adjusted and unadjusted, and unstratified permutation inference are applied to estimate and test for θ. Several notable patterns exist here. In particular, model-based inference maintains the desired type I error range most of the times except for scenario 1 where adjusted analysis consistently displays type 1 error inflation. In

Table 3: Power for testing the null hypothesis $H_0 : \theta = 0$ versus $H_a : \theta = \theta_1$. Data are simulated from equation (1) with a single binary covariate under scenarios 1 to 8, with simple vs stratified randomization, adjusted vs adjusted analyses, model-based vs permutation-based inference been performed. The covariate coefficient α is fixed to be 1.

Scenario	Adjusted for covariate in analysis	Model-based inference		Permutation-based inference	
		Simple randomization	Stratified randomization	Simple randomization	Stratified randomization
1	Yes	0.893	0.911	0.569	0.712
	No	0.848	0.851	0.074	0
2	Yes	0.788	0.792	0.502	0.598
	No	0.759	0.753	0.083	0
3	Yes	0.853	0.858	0.152	0.213
	No	0.847	0.849	0.072	0.005
4	Yes	0.760	0.756	0.152	0.195
	No	0.759	0.751	0.077	0.005
5	Yes	0.934	0.937	0.799	0.810
	No	0.881	0.892	0.110	0
6	Yes	0.857	0.861	0.691	0.717
	No	0.819	0.821	0.105	0
7	Yes	0.890	0.898	0.237	0.250
	No	0.882	0.891	0.097	0.006
8	Yes	0.826	0.831	0.227	0.242
	No	0.820	0.822	0.094	0.009

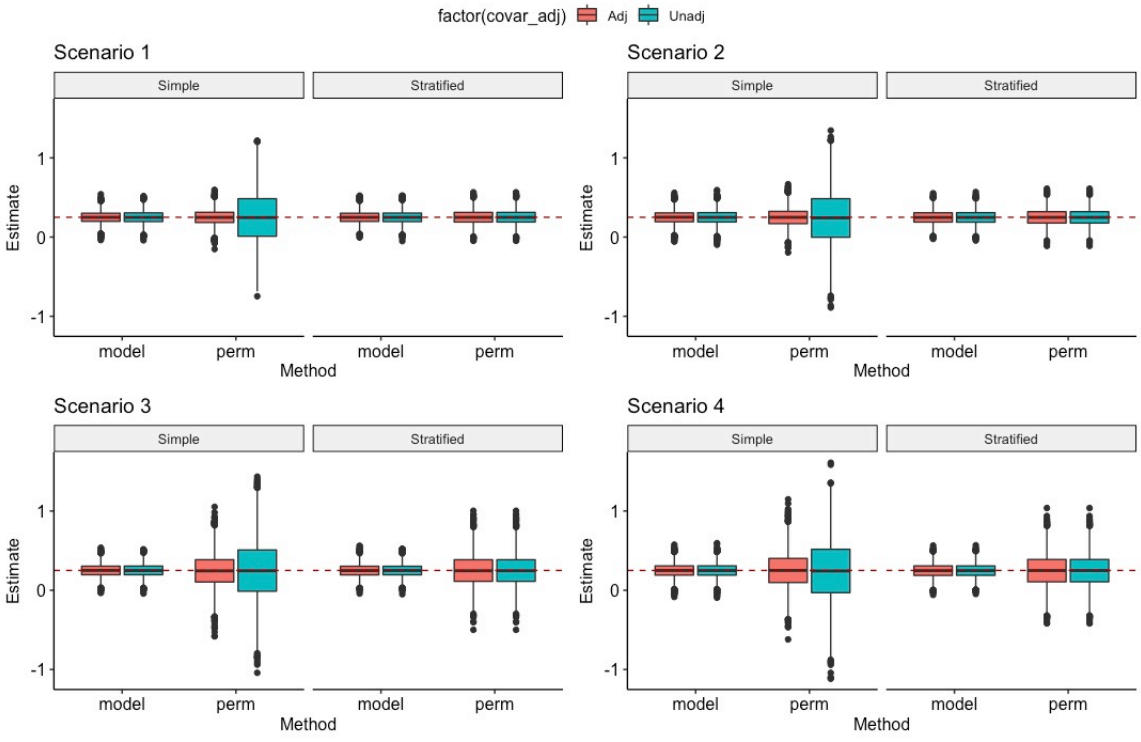

Figure 2: Boxplots of $\hat{\theta}$ obtained by simple vs stratified randomization, model-based vs permutation-based inference and adjusted vs unadjusted analyses for scenarios 1 to 4, with the actual $\theta = 0.25$ as pointed by the red dashed line

general, adjusted analysis tends to be slightly more anti-conservative than its unadjusted counterpart, especially for scenario 1. Nonetheless such disparity becomes less evident with larger values of I and τ as well as the

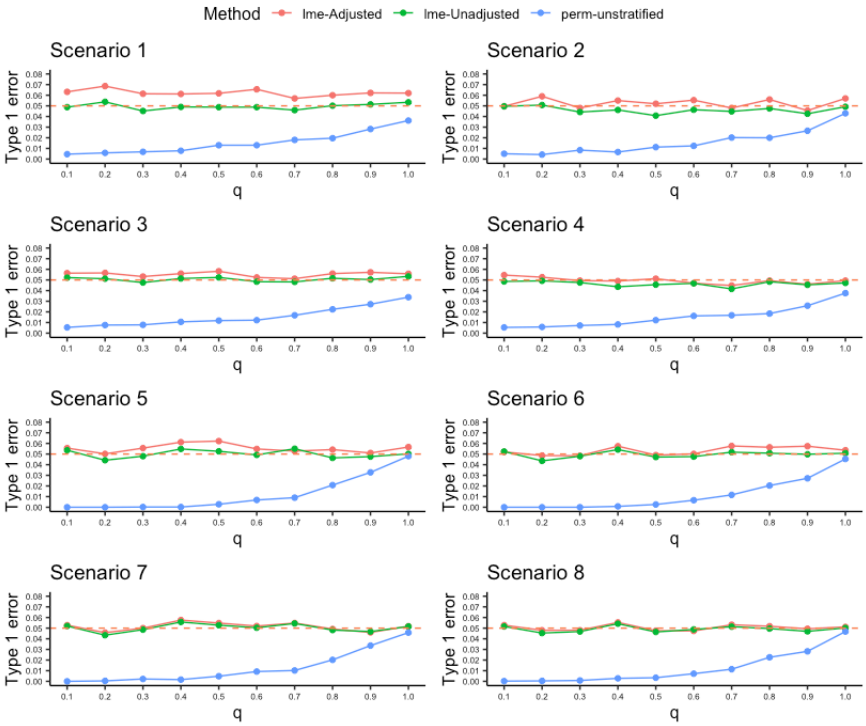

Figure 3: Type I error rate for testing the null hypothesis $H_0 : \theta = 0$. Data are simulated from equation (1) with a mixture of continuous and categorical covariates under scenarios 1 to 8, model-based inference (labeled as lme), both adjusted and unadjusted for covariates, as well as unstratified permutation-based inference are applied. Sequential balance (or equivalently, the treatment vs Control balance) score is used to sort the randomization sequences, with the level of constraint q ranging from 0.1 (the tightest) to 1 (simple randomization) stepped by 0.1

introduction of random treatment effect η, and we can see there is almost no difference in scenarios 7 and 8 as the two curves nearly overlap. Interestingly, there is no clear pattern of type 1 error change with respect to the level of constraint q. For unstratified permutation-based inference with no constraint (q=1), the type I error stays below the nominal level at scenario 1 and 3 with smaller values of I and τ, but such deflation gets less at scenario 2 and 4 with larger value of τ, and almost vanishes at scenario 5 to 8 with larger value of I. However, the type I error gradually drops down as q decreases for all scenarios. Similar to the first case with single binary covariate, this behavior indicates the failure of using the proposed permutation inference for testing θ in the presence of multiple covariates. In general, there is not much impact of the balance scores used on the general pattern as displayed in figure 3 and 4, except the curves for unstratified permutation inference seems to drop more quickly with decreasing values of q in figure 3.

Figures 5 and 6 summarize the results on power corresponding to Figure 3 and 4 for model based inference. Power of unstratified permutation-based inference is not included as its type I error was already shown to be invalid. Again there is no marked difference between the balance scores used. Also, by looking at scenarios

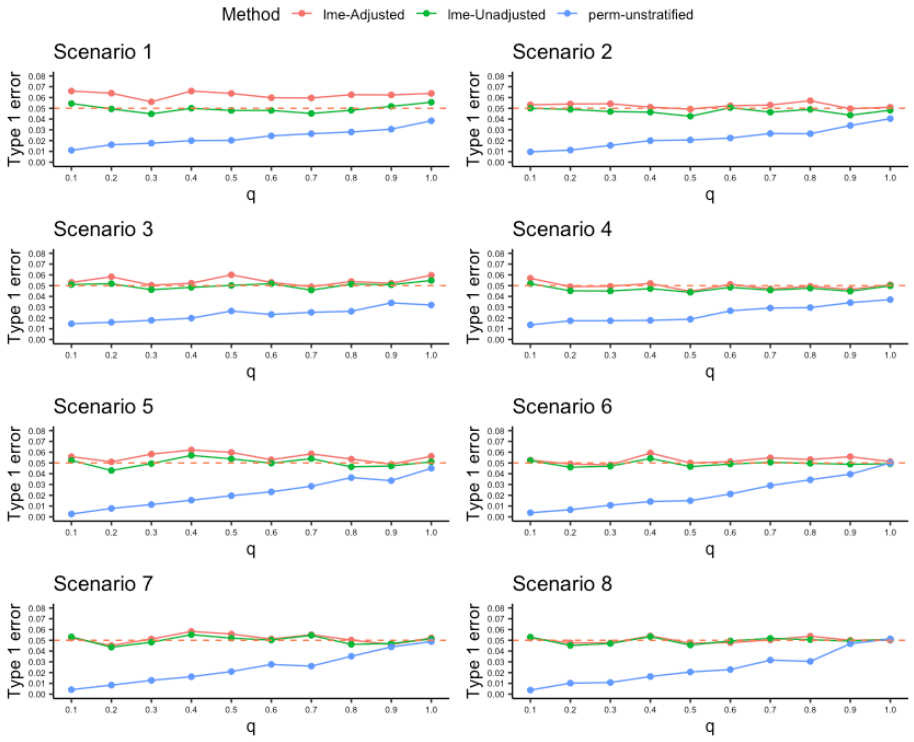

Figure 4: Type I error rate for testing the null hypothesis $H_0 : \theta = 0$. Data are simulated from equation (1) with a mixture of continuous and categorical covariates under scenarios 1 to 8, model-based inference (labeled as lme), both adjusted and unadjusted for covariates, as well as unstratified permutation-based inference are applied. Mean balance score is used to sort the randomization sequences, with the level of constraint q ranging from 0.1 (the tightest) to 1 (simple randomization) stepped by 0.1

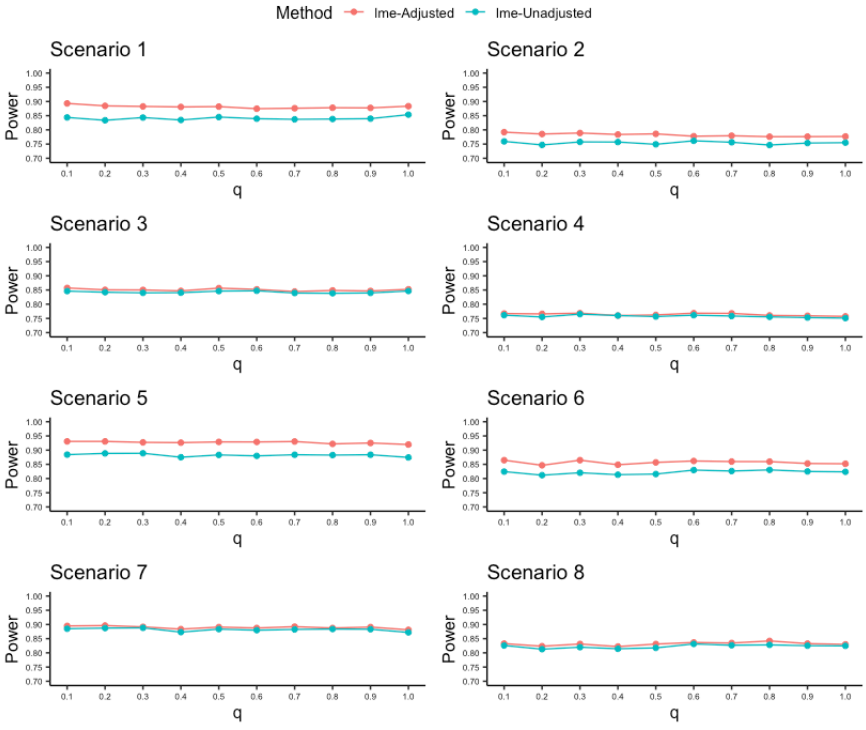

Figure 5: Power for testing the null hypothesis $H_0 : \theta = 0$ versus $H_0 : \theta = \theta_a$. Data are simulated from equation (1) with a mixture of continuous and categorical covariates under scenarios 1 to 8, model-based inference (labeled as lme), both adjusted and unadjusted for covariates are applied. Sequential balance (or equivalently, the treatment vs Control balance) score is used to sort the randomization sequences, with the level of constraint q ranging from 0.1 (the tightest) to 1 (simple randomization) stepped by 0.1

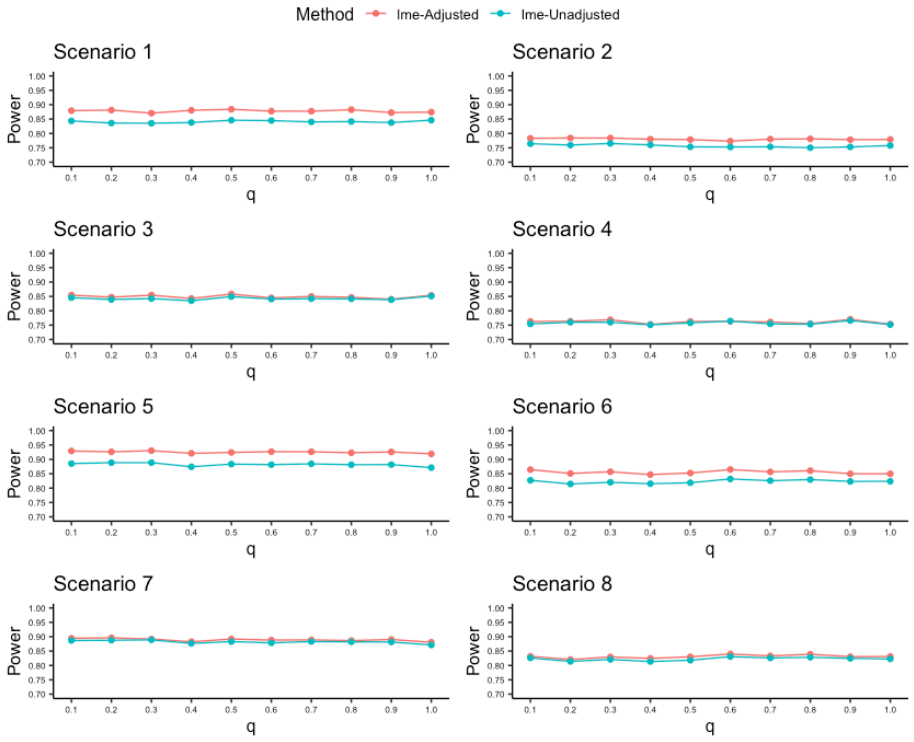

Figure 6: Power for testing the null hypothesis $H_0 : \theta = 0$ versus $H_0 : \theta = \theta_a$. Data are simulated from equation (1) with a mixture of continuous and categorical covariates under scenarios 1 to 8, model-based inference (labeled as lme), both adjusted and unadjusted for covariates, are applied. Sequential balance (or equivalently, the treatment vs Control balance) score is used to sort the randomization sequences, with the level of constraint q ranging from 0.1 (the tightest) to 1 (simple randomization) stepped by 0.1

1-4 and 5-8 separately, we observe rather similar patterns in the power curves, except the power is apparently levelled up in the last four scenarios where $I = 24$. For scenarios 1, 2, 5, 6 with smaller τ, adjusted analysis consistently achieves higher power than its unadjusted counterpart at all values of q, though such difference is not dramatic. For the other four scenarios, although adjusted analysis still dominates, such difference in power diminishes as the two curves nearly overlap. We can see that the introduction of η lowers the power by comparing scenarios pair-to-pair horizontally. Also, there is less reduction in power with higher values of τ by comparing scenarios pair-to-pair vertically, and such reduction only occurs under adjusted analysis, which narrows the power gap with the unadjusted counterpart. Finally, although our particular attention lies on the potential association between the level of constraint and power, there is little visible association between them shown in the figures.

In table 4, we list the slopes of fitted regression lines of power against q corresponding to Figure 5 and 6 to further evaluate their association. As can be seen, no matter which balance score is used, slopes are always negative for adjusted analysis (except for scenario 8). However, they are sometimes positive or negative without

Table 4: Estimated slope of power against q(level of constraint) for scenarios 1 to 8, using sequential(treatment v.s. control) and mean balance as the constraint criteria. Results are listed for model-based inference only, with both adjusted and unadjusted analysis.

Scenario	Adjusted for covariate	Estimated slopes	
		Sequential(treatment v.s. control) balance score	Mean balance score
1	Yes	-0.0111	-0.0037
	No	0.0063	0.0042
2	Yes	-0.0174	-0.0058
	No	-0.0023	-0.0120
3	Yes	-0.0056	-0.0051
	No	-0.0015	0.0032
4	Yes	-0.0088	-0.0046
	No	-0.0094	0.0002
5	Yes	-0.0093	-0.0067
	No	-0.0079	-0.0110
6	Yes	-0.0033	-0.0050
	No	0.0114	0.0075
7	Yes	-0.0096	-0.0108
	No	-0.0086	-0.0122
8	Yes	0.0082	0.0090
	No	0.0101	0.0080

observable inclination to either side for unadjusted analysis. This again suggests that constrained randomization is more effective in improving power when covariates are also included in the fitted model. However, it is worth noting that all the estimated slopes are small in magnitude. The one which has the highest abosolute value is -0.0174, indicating that, on average, constraining the randomization space by a factor of 0.1 can only bring around 0.002 increase in power. Therefore the actual effects of the constraints on power are trivial in practical sense.

4 Discussion

Under the stepped wedge setting, we studied the effect of constrained randomization on type I error rates and power using two inferential procedures, one to fit the actual linear mixed effect model and the other to apply the robust permutation inference proposed by Hughes et al. [16]. Results are further sorted by whether cluster-level covariates are adjusted or not. We considered the case with only one binary covariate, where stratified randomization was used and evaluated, as well as the case with a mixture of categorical and continuous covariates, where different balance metrics were used to as the constraint criteria to order the randomization schemes.

For model-based inference with a single binary covariate, the slight type 1 error inflation in scenario 1 with adjusted analysis raises some concerns over the asymptotic property of the F test used when the cluster number

I is small. Nevertheless all the other values are within the 95% error margin of the nominal type I error rate, indicating that F test can still be relied upon to a large extent, at least in this case where we fit the correct model that was used to generate the data. Regarding the power, adjusting for covariates at the analysis phase is shown to be more effective than doing stratification at the design phase, though of course there is usually further gain in power by stratification when covariates have already been adjusted. In contrast, there is no clear advantage of stratification at the design phase when covariates are not controlled for analysis.

Under permutation-based inference, when $I = 8$ and the analysis is stratified by covariate levels, the test turns out to be conservative most of times no matter whether the randomization space is constrained or not. However, there is not much deflation with respect to type I error when I increases to 24, indicating the test is also more reliable with larger cluster number. On the other hand, the 0 type 1 error rates shown in table 2 for all scenarios indicate the test fails when the randomization space is constrained within each covariate stratum but the analysis space is not, and similar results were also obtained by Li et al. [10]. Such test failure is mainly due to the estimation of marginal variance which is much larger than the true variance under the constraint, as illustrated by [10]. This phenomenon draws also our attention that the analysis and randomization space should be matched to guarantee the test is valid, which is also illustrated by Li et al. [10]. In terms of power for adjusted analysis, stratified randomization consistently performs better than simple randomization, and such improvement is greater than it is under model-based inference. This is likely because the constraint in randomization space reduces the variance of $\tilde{\theta}$. Besides, both the introduction of random treatment effect η and increase in the value of random cluster effect τ reduce the power. It is also worth noting the power gets significantly lower for simple randomization and unadjusted analysis. The reason is that the permutation inference utilize between-cluster information to estimate θ, and the variation of fixed effect term across clusters was not included in the analysis, hence got included in the random error, which has the same effect as increasing the error variance. This confirms that, under permutation-based inference, although there can be some advantages for doing stratification at the design phase in terms of power, it is still more important to adjust for covariates at the analysis phase.

For the second case with multiple covariates and model-based inference, the slightly inflated type 1 errors of adjusted analysis compared with its unadjusted counterpart, which can also be found for the case with single

binary covariate, are probably due to the loss in degree of freedom by the inclusion of covariates into model fitting. However, such inflated type I errors converge to nominal level with larger values of I and τ as well as the presence of η.

For permutation-based inference, the only choice is to do unadjusted analysis as it becomes infeasible to stratify clusters by their covariate values when multiple covariates present. Even if we stratify by, say, partitioning the continuous covariate into multiple levels and form strata by taking all possible combinations of the cluster-level covariates, the strata formed can be very sparse which can make the variance estimate of $\tilde{\theta}$ to be unstable. Moreover, the constancy of fixed effects across clusters is still not satisfied since the outcome is prognostic to the continuous values of the covariate instead of its partitioned levels. The failure of unadjusted permutation inference, as shown by the over-conservative type 1 errors in figure 3 and 4, is again due to the mismatch between the randomization and analysis space similar to the first case.

The power curves in figure 5 and 6 turn out to be flat, and we do not see a monotonic increasing trend of power with tighter constraint level for adjusted analysis as shown by Li et al. [10] under parallel cluster-randomized setting. Although table 4 conveys some evidence that additional power can be attained by further constraining the randomization space if covariates are also adjusted for analysis, it should be noted that all the slope estimates are small in magnitude hence the actual gain in power is too marginal to have any strong practical implications. One possible explanation is that, under stepped wedge setting where there are multiple time periods for each cluster, the linear mixed model takes into account both within and between cluster information to estimate θ, whereas only between-cluster information is available under the parallel setting. And since we assume the cluster-level covariates are constant across time, the within-cluster information is unaffected by the covariate imbalance hence less power is lost under the stepped wedge setting. In contrast, the power comparison between adjusted versus adjusted analyses is relatively clearer as the former one consistently yields higher power at all the constraint levels been considered. It is mainly because the unadjusted analysis does not condition on the cluster-level covariates, and the heterogeneity of covariate values among clusters is instead included as a part of the random cluster effect, which has the same effect as increasing the intra-class correlation, hence making the effective sample size to be smaller.

Relatively speaking, the gain in power by balancing covariates at the design stage is more evident with

single binary covariate than with a mixture of covariates. This might due to the fact that the presence of multiple covariates tend to offset the covariate imbalance among sequences, hence making a bad randomization scheme less likely to happen.

This study has several potential limitations. First, we only consider the standard balanced design where there are equal number of crossover(s) from the second to the last time periods. This is not always true in reality, for instance, after all clusters have received treatment, records might be gathered for additional time periods. In such cases the treatment versus control balance and the sequential balance are no longer equivalent and the results can be potentially very different. Secondly, we assume the design to be cross-sectional, in which there is no correlation at the individual level. Further research is required to analyze the case for cohort design, where there is assumed to be the same group of individuals within each cluster throughout the time periods. Another limitation is that we assume equal cluster size $N = 100$. If instead cluster sizes are different and unevenly distributed among treatment sequences, the testing power may be affected in a similar way as in the case with covariate imbalance. Further research could investigate performing constrained randomization by treating cluster size as one of the covariates and take it into the calculation of balance scores, and investigate its effect on power. Another limitation comes with table 4. Although we aim to explore the association between power and the level of constraint q by looking at the slopes of regression lines, their actual association may not be linear, and the effect is also likely to vary in different ranges of q. Both of these might weaken the validity of our analysis. Moreover, we should be aware that constrained randomization reduces covariate imbalance whereas at the same time also reduces the randomness of the trial, which can potentially create ethical issues in certain circumstances. Therefore, it would be more practically plausible to incorporate such trade-off while setting constraints on the randomization space, possibly by setting a lower bound on the proportion of candidate sequences to be selected upon.

This study analyzed various situations by using constrained randomization under the stepped wedge setting, and we believe that it can provide some guidance to researchers on which design and analytical approach to use with data collected from stepped wedge trials under the presence of cluster-level covariates .

BIBLIOGRAPHY

[1] Ren Y, Hughes JP, Heagerty PJ (2019). A Simulation Study of Statistical Approaches to Data Analysis in the Stepped Wedge Design. *Statistics in Biosciences*. https://doi.org/10.1007/s12561-019-09259-x

[2] Lew RA, Miller CJ, Kim BY, Wu H, Stolzmann K, Bauer MS (2019). A method to reduce imbalance for site-level randomized stepped wedge implementation trial designs. *Implementation Science*, 14: Article 46. https://doi.org/10.1186/s13012-019-0893-3

[3] Girling A, Hemming K (2016). Statistical efficiency and optimal design for stepped cluster studies under linear mixed effects models. *Statistics in Medicine*, 35:2149-2166

[4] Hu Y, Hoover DR (2018). Non-randomized and randomized stepped-wedge designs using an orthogonalized least squares framework. *Statistical Methods in Medical Research*, 27(4):1202-1218

[5] Moulton LH, Golub JE, Durovni B, Cavalcante SC, Pacheco AG, Saraceni V, King B, Chaisson RE (2007). Statistical design of THRio: a phased implementation clinic-randomized study of a tuberculosis preventive therapy intervention. *Clinical Trials*, 4(2):190-199

[6] Ji X, Gunther F, Robyn PJ, Small DS (2017). Randomization inference for stepped-wedge cluster-randomized trials: An application to community-based health insurance. *The Annals of Applied Statistics*, 11:1-20

[7] Stern A, Mitsakakis N, Paulden M, Alibhai S, Wong J, Tomlinson G, Brooker AS, Krahn M, Zwarenstein M (2014). Pressure ulcer multidisciplinary teams via telemedicine: a pragmatic cluster randomized stepped wedge trial in long term care. *BMC Health Services Research*, 14:Article 83. https://doi.org/10.1186/1472-6963-14-83

[8] Barker D, McElduff P, D'Este C, Campbell MJ (2016). Stepped wedge cluster randomised trials: a review of the statistical methodology used and available. *BMC Medical Research Methodology*, 16:Article 69. https://doi.org/10.1186/s12874-016-0176-5

[9] Hughes JP, Granston TS, Heagerty PJ (2015). Current issues in the design and analysis of stepped wedge trials. *Contemporary Clinical Trials*, 45:55–60

[10] Li F, Lokhnygina Y, Murray DM, Heagerty PJ, DeLong ER (2016). An evaluation of constrained randomization for the design and analysis of group-randomized trials. *Statistics in Medicine*, 35:1565–79.

[11] Moulton LH (2004). Covariate-based constrained randomization of group-randomized trials. *Clinical Trials*, 1(3):297-305.

[12] Li F, Turner EL, Heagerty PJ, Murraye DM, Vollmerf WM, DeLong ER (2017). An evaluation of constrained randomization for the design and analysis of group-randomized trials with binary outcomes. *Statistics in Medicine*, 36(24): 3791–3806

[13] Kernan WN, Viscoli CM, Makuch RW, Brass LM, Horwitz RI (1999). Stratified Randomization for Clinical Trials. *Journal of Clinical Epidemiology*, 52(1):19-26

[14] Rosenberger WF, Sverdlov O (2008). Handling Covariates in the Design of Clinical Trials. *Statistical Science*, 23(3):404-419

[15] Raab GM, Butcher I (2001). Balance in cluster randomized trials. *Statistics in Medicine*, 20(3): 351–365

[16] Hughes JP, Heagerty PJ, Xia F, Ren Y (2019). Robust inference in the stepped wedge design. *Biometrics*, 76(1):119-130

[17] Zhou X, Liao X, Spiegelman D (2017). "Cross-sectional" stepped wedge designs always reduce the required sample size when there is no time effect. *Journal of Clinical Epidemiology*, 83: 108–109

[18] de Hoop E, Teerenstra S, van Gaal BGI, Moerbeek M, Borm GF (2012). The "best balance" allocation led to optimal balance in cluster-controlled trials. Journal of Clinical Epidemiology. *Journal of Clinical Epidemiology*, 65(2):132–137

[19] Kuznetsova A, Brockhof PB, Christensen RHB (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13). https://doi.org/10.18637/jss.v082.i13

APPENDIX

R code used for this project can be found in the following link:

https://github.com/pygao1/Constrained-randomization