Research Article

Statistics in Medicine

Received 6 June 2013, Accepted 28 May 2015 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6565

Recommendations for choosing an analysis method that controls Type I error for unbalanced cluster sample designs with Gaussian outcomes

Jacqueline L. Johnson,[a] Sarah M. Kreidler,[b*†] Diane J. Catellier,[c] David M. Murray,[d] Keith E. Muller[e] and Deborah H. Glueck[f]

We used theoretical and simulation-based approaches to study Type I error rates for one-stage and two-stage analytic methods for cluster-randomized designs. The one-stage approach uses the observed data as outcomes and accounts for within-cluster correlation using a general linear mixed model. The two-stage model uses the cluster specific means as the outcomes in a general linear univariate model. We demonstrate analytically that both one-stage and two-stage models achieve exact Type I error rates when cluster sizes are equal. With unbalanced data, an exact size α test does not exist, and Type I error inflation may occur. Via simulation, we compare the Type I error rates for four one-stage and six two-stage hypothesis testing approaches for unbalanced data. With unbalanced data, the two-stage model, weighted by the inverse of the estimated theoretical variance of the cluster means, and with variance constrained to be positive, provided the best Type I error control for studies having at least six clusters per arm. The one-stage model with Kenward–Roger degrees of freedom and unconstrained variance performed well for studies having at least 14 clusters per arm. The popular analytic method of using a one-stage model with denominator degrees of freedom appropriate for balanced data performed poorly for small sample sizes and low intracluster correlation. Because small sample sizes and low intracluster correlation are common features of cluster-randomized trials, the Kenward–Roger method is the preferred one-stage approach.

Keywords: cluster randomized; group randomized; Type I error; unbalanced; Gaussian

1. Introduction

In cluster randomized designs, also known as group-randomized designs, scientists randomize clusters of research participants to the intervention of interest [1–3]. Common clusters include communities, schools, or families. Cluster randomized designs often improve efficiency of intervention delivery and reduce the effect of treatment contamination [2, 4]. Observations on participants within a cluster are correlated because of the common experiences and characteristics of research participants within a cluster. Analytic approaches that do not account for within-cluster correlation yield artificially reduced standard errors and inflated Type I error [1, 2, 5].

In a balanced design, in which every cluster is the same size, ordinary least squares methods give a Type I error rate that is exactly the rate specified in the design. However, Type I error inflation may occur when statistical methods appropriate for balanced data are applied to data with unbalanced cluster sizes.

[a]*University of North Carolina, Department of Psychiatry, Chapel Hill, NC, U.S.A.*
[b]*Neptune and Company, Lakewood, CO, U.S.A.*
[c]*RTI International, Research Triangle Park, NC, U.S.A.*
[d]*Biostatistics and Bioinformatics Branch, Division of Epidemiology Statistics, and Prevention Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Rockville, MD, U.S.A.*
[e]*Department of Health Outcomes and Policy, University of Florida, Gainesville, FL, U.S.A.*
[f]*Department of Biostatistics and Informatics, University of Colorado Denver, Aurora, CO, U.S.A.*
[*]*Correspondence to: Sarah Kreidler, Neptune and Company, Garrison Street, Suite, Lakewood,CO, U.S.A.*
[†]*E-mail: skreidler@neptuneinc.org*

Murray [1] suggested that clustered data can be analyzed either via a 'one-stage' or a 'two-stage' model. In the one-stage model, observations on individual research participants are analyzed with a linear mixed model. Within-subject correlation is handled with a random intercept for cluster. In the two-stage analysis, cluster means are analyzed using a general linear model. Because the analysis is performed at the cluster level, the assumption of independence holds. Murray [1] noted that the one-stage and two-stage models appropriately account for within-cluster correlation. Donner and Klar [2] and Murray *et al*. [6] provided full details on statistical methods for cluster randomized trials.

A variety of statistical methods have been applied to clustered data. In a review of 60 group randomized trials, Varnell *et al*. [7] found that 54% of studies reported only appropriate statistical methods, 25% had a mix of appropriate and inappropriate methods, and 20% used inappropriate methods that did not adequately account for within-cluster correlation. For the studies that used appropriate methods, 68% used the one-stage approach, and 32% analyzed cluster means or a similar summary statistic [7]. In cluster randomized trials of cancer screening, Crespi *et al*. [4] found that only 60% of trials published between 2007 and 2010 used appropriate statistical analyses. In addition, the authors suggested that use of appropriate methods for cluster randomized designs increased in the early to mid-2000s, but that usage had declined in recent years.

The paper is organized as follows. In Section 2, we review the one-stage and two-stage models, hypothesis testing approaches, and causes of possible Type I error inflation with unbalanced data. In Section 3, we demonstrate analytically that one-stage and two-stage models yield exact Type I error rates when cluster sizes are balanced. In Section 4, we use a simulation study to compare the Type I error performance of four one-stage and six two-stage hypothesis testing methods across a wide range of study designs. Section 5 provides guidance on selecting an analytic method.

2. Notation, models, and hypothesis testing

2.1. Notation

We define matrix operations for one-stage and two-stage models. Let $\boldsymbol{a}$ denote an $(n \times 1)$ column vector, $\boldsymbol{A} = \{a_{ij}\}, i \in \{1, 2, \ldots n\}, j \in \{1, 2, \ldots m\}$ indicate an $(n \times m)$ matrix with transpose $\boldsymbol{A}' = \{a_{ji}\}$, and $\boldsymbol{D} = \{\boldsymbol{D}_{ij}\}$ a supermatrix [8, p. 3] containing smaller matrices $\boldsymbol{D}_{ij}$. For a matrix $\boldsymbol{A} = [\boldsymbol{a}_1\ \boldsymbol{a}_2\ \ldots\ \boldsymbol{a}_n]$, let $\text{vec}(\boldsymbol{A}) = [\boldsymbol{a}_1'\ \boldsymbol{a}_2'\ \ldots\ \boldsymbol{a}_n']'$. Let $(\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_n)$ denote an ordered list of matrices, and $(\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n)$ denote an ordered list of vectors. Define the list creation operator as $\boxdot_{i=1}^{n} \boldsymbol{A}_i = (\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_n)$. Note that $\boxdot_{i=1}^{n} \boxdot_{j=1}^{m} \boldsymbol{A}_{ij} = (\boldsymbol{A}_{11}, \boldsymbol{A}_{12}, \ldots, \boldsymbol{A}_{1m}, \ldots, \boldsymbol{A}_{n1}, \ldots, \boldsymbol{A}_{nm})$. For arbitrarily sized matrices, $\boldsymbol{A}_i$, define $\text{diag}\left(\boxdot_{i=1}^{n} \boldsymbol{A}_i\right) = \{\boldsymbol{D}_{ij}\}$ with $\boldsymbol{D}_{ij} = \boldsymbol{A}_i$ for $i = j$, and $\boldsymbol{0}$ otherwise. Let $\boldsymbol{1}_n$ denote an $n \times 1$ column vector of ones, and $\boldsymbol{I}_n = \text{diag}\left(\boxdot_{i=1}^{n} 1\right)$. Define the Kronecker product of two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ as $\boldsymbol{A} \otimes \boldsymbol{B} = \{a_{ij}\boldsymbol{B}\}$ [8, p. 7]. Define $\mathcal{V}(\boldsymbol{a})$ as the covariance of $\boldsymbol{a}$.

We consider study designs with one level of clustering and a single fixed effect of interest, which we refer to as treatment. Let $h \in \{1, \ldots, g\}$ index treatment groups, $i \in \{1, \ldots, m_h\}$ index clusters within treatment group h, and $j \in \{1, \ldots, n_{hi}\}$ index observations within treatment cluster i and treatment group h. Denote the total number of clusters by $m = \sum_{h=1}^{g} m_h$, the number of observations in treatment group h by $n_h = \sum_{i=1}^{m_h} n_{hi}$, and the total number of observations by $N = \sum_{h=1}^{g} \sum_{i=1}^{m_h} n_{hi}$.

2.2. The one-stage model for clustered data

The one-stage model is a general linear mixed model with Gaussian errors [9]. Define $\boldsymbol{y}_{hi}$ as the vector of outcomes for individual research participants in treatment h and cluster i, with the complete $(N \times 1)$ outcome vector $\boldsymbol{y} = \text{vec}(\boldsymbol{y}_{hi})$. Using cell mean coding [10], define the $(N \times g)$ design matrix of individual-level treatment effects $\boldsymbol{X} = \text{diag}\left(\boxdot_{h=1}^{g} \boldsymbol{1}_{n_h}\right)$. Let $\boldsymbol{\beta}$ be the $(g \times 1)$ vector of treatment means. Account for within cluster correlation using a single random effect for cluster so that the $(N \times m)$ matrix $\boldsymbol{Z} =$

$\text{diag}\left(\boxdot_{h=1}^{g}\boxdot_{i=1}^{m_h}\mathbf{1}_{n_{hi}}\right)$. Let $\boldsymbol{d}\sim\mathcal{N}_m\left(0,\sigma_c^2\boldsymbol{I}_m\right)$ be the $(m\times 1)$ vector of random cluster offsets, independently distributed from the $(N\times 1)$ vector of residual errors $\boldsymbol{e}\sim\mathcal{N}_N\left(0,\sigma_e^2\boldsymbol{I}_N\right)$. The one-stage model is

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{d}+\boldsymbol{e}. \tag{1}$$

Define the total variance by $\sigma_y^2=\sigma_c^2+\sigma_e^2$ and the within-cluster correlation coefficient by $\rho=\sigma_c^2/\sigma_y^2$. The random effect for cluster induces a compound symmetric form for the variance of $\boldsymbol{y}_{hi}$ so that the covariance of $\boldsymbol{y}$, $\boldsymbol{\Sigma}$, is given by

$$\boldsymbol{\Sigma}=\sigma_c^2\boldsymbol{Z}\boldsymbol{Z}'+\sigma_e^2\boldsymbol{I}_N. \tag{2}$$

The covariance can be equivalently expressed in terms of ρ and σ_y^2 as

$$\boldsymbol{\Sigma}=\sigma_y^2\,\text{diag}\left\{\boxdot_{h=1}^{g}\boxdot_{i=1}^{m_h}\left[\mathbf{1}_{n_{hi}}\mathbf{1}'_{n_{hi}}\rho+\boldsymbol{I}_{n_{hi}}(1-\rho)\right]\right\}. \tag{3}$$

For balanced data, we define $n_{hi}\equiv n\;\forall\;h\in\{1,\ldots,g\}$, $i\in\left\{1,\ldots,m_h\right\}$. In this case, the covariance reduces to

$$\boldsymbol{\Sigma}=\boldsymbol{I}_m\otimes\left[\sigma_c^2\mathbf{1}_n\mathbf{1}'_n+\sigma_e^2\boldsymbol{I}_n\right]. \tag{4}$$

In terms of ρ and σ_y^2, the covariance for balanced data is

$$\boldsymbol{\Sigma}=\boldsymbol{I}_m\otimes\sigma_y^2\left[\mathbf{1}_n\mathbf{1}'_n\rho+\boldsymbol{I}_n(1-\rho)\right]. \tag{5}$$

2.3. The two-stage model for clustered data

The two-stage model is a general linear univariate model with Gaussian errors [8, Chapter 2]. The outcomes are the means for each cluster h and treatment i, denoted $\bar{y}_{hi}$. The two-stage model is obtained by transforming the one-stage model by the $(m\times N)$ matrix

$$\boldsymbol{T}=\text{diag}\left(\boxdot_{h=1}^{g}\boxdot_{i=1}^{m_h}\mathbf{1}'_{n_{hi}}/n_{hi}\right). \tag{6}$$

We denote elements of the two-stage model using the subscript T and corresponding elements from the one-stage model with no subscript. Let $\boldsymbol{y}_T$ be the $(m\times 1)$ vector of cluster means, with $\boldsymbol{y}_T=\boldsymbol{T}\boldsymbol{y}$. Using cell mean coding, define the $(m\times g)$ design matrix of cluster-level treatment effects $\boldsymbol{X}_T=\text{diag}\left(\boxdot_{h=1}^{g}\mathbf{1}_{m_h}\right)$, with $\boldsymbol{X}_T=\boldsymbol{T}\boldsymbol{X}$. Let $\boldsymbol{\beta}$ be the $(g\times 1)$ vector of treatment means. Notice that population treatment means, $\boldsymbol{\beta}$, are equivalent for both the one-stage and two-stage models. Define $\boldsymbol{e}_T$ to be the $(m\times 1)$ vector of residual errors, with $\boldsymbol{e}_T=\boldsymbol{T}\boldsymbol{e}$. The two-stage model is

$$\boldsymbol{y}_T=\boldsymbol{X}_T\boldsymbol{\beta}+\boldsymbol{e}_T. \tag{7}$$

Define $\boldsymbol{\Sigma}_T$ to be the $(m\times m)$ covariance of $\boldsymbol{y}_T$. Because $\boldsymbol{y}_T=\boldsymbol{T}\boldsymbol{y}$, $\boldsymbol{\Sigma}_T=\boldsymbol{T}\boldsymbol{\Sigma}\boldsymbol{T}'$. The covariance can be expressed in terms of σ_c^2 and σ_e^2 as

$$\boldsymbol{\Sigma}_T=\text{diag}\left[\boxdot_{h=1}^{g}\boxdot_{i=1}^{m_h}\left(\sigma_c^2+\sigma_e^2/n_{hi}\right)\right]. \tag{8}$$

The covariance can be equivalently expressed in terms of ρ and σ_y^2 as

$$\boldsymbol{\Sigma}_T=\sigma_y^2\,\text{diag}\left[\boxdot_{h=1}^{g}\boxdot_{i=1}^{m_h}\left\{\left[1+\left(n_{hi}-1\right)\rho\right]/n_{hi}\right\}\right]. \tag{9}$$

For balanced data, we have $n_{hi}\equiv n\;\forall\;h\in\{1,\ldots,g\}$, $i\in\left\{1,\ldots,m_h\right\}$. In this case, the covariance reduces to

$$\boldsymbol{\Sigma}_T=\left(\sigma_c^2+\sigma_e^2/n\right)\times\boldsymbol{I}_m. \tag{10}$$

In terms of ρ and σ_y^2, the covariance for balanced data in the two-stage model is

$$\boldsymbol{\Sigma}_T = \left(\sigma_y^2/n\right)\left[1+(n-1)\,\rho\right]\times \boldsymbol{I}_m. \tag{11}$$

In practice, true values for σ_c^2 and σ_e^2 are unknown. The corresponding estimates, $\hat{\sigma}_c^2$ and $\hat{\sigma}_e^2$, can be obtained by fitting a one-stage model (see Appendix B for details). When restricted maximum likelihood estimation is used to estimate the covariance parameters, $\hat{\sigma}_c^2$ and $\hat{\sigma}_e^2$ are commonly constrained to be positive. However, allowing the variance components to be negative allows $\hat{\rho}$, the estimate of ρ, to be negative. As described by Muller and Stewart [8, p. 37–38], a negative value for $\hat{\rho}$ may be required to ensure that the resulting compound symmetric covariance is positive definite. If ρ were exactly zero, then $\hat{\rho}$ would have a symmetric distribution about zero. Thus, half the time, $\hat{\rho}$ would be negative. If ρ were near zero, a substantial fraction of estimates of ρ would be negative. Therefore, constraining $\hat{\rho}$ to be positive would fail to represent the true sampling distribution of $\hat{\rho}$.

2.4. *The general linear hypothesis*

The general linear hypothesis for the one-stage and two-stage models is $\boldsymbol{\theta} = \boldsymbol{C}\boldsymbol{\beta}$, where $\boldsymbol{C}$ is an $(a \times g)$ matrix of between-subject contrasts, and $\boldsymbol{\theta}$ is the $(a \times 1)$ vector of observed contrast values. We study the two-sided general linear hypothesis, which compares $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ with $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. In most cases, $\boldsymbol{\theta}_0 = \boldsymbol{0}$. We require that $\boldsymbol{X}$ have full rank to ensure that $\boldsymbol{\theta}$ is estimable. We also require that $\boldsymbol{C}$ have full row rank [rank $(\boldsymbol{C}) = a$] to ensure the general linear hypothesis is testable [10].

2.5. *Tests of fixed effects in one-stage model with unbalanced data*

In the one-stage model, restricted maximum likelihood estimation [11] is commonly used to obtain estimates of the covariance, $\hat{\boldsymbol{\Sigma}}$, and estimates of the regression parameters, $\hat{\boldsymbol{\beta}}$. Because $\hat{\boldsymbol{\beta}}$ depends on $\hat{\boldsymbol{\Sigma}}$, an iterative optimization procedure, such as the expectation–maximization algorithm [12], must be used to obtain the estimates. Consequently, no closed form expression exists for $\mathcal{V}\left(\hat{\boldsymbol{\beta}}\right)$. Analysts typically use the estimated variance $\hat{\mathcal{V}}\left(\hat{\boldsymbol{\beta}}\right) = \left(\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}\right)^{-1}$ to form the Wald statistic

$$T = \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \left\{\boldsymbol{C}\left(\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{C}'\right\}^{-1}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)/a, \tag{12}$$

where $\hat{\boldsymbol{\theta}} = \boldsymbol{C}\hat{\boldsymbol{\beta}}$. Kackar and Harville [13] and Dempster *et al*. [14] previously demonstrated that $\hat{\mathcal{V}}\left(\hat{\boldsymbol{\beta}}\right)$ underestimates the variance of $\hat{\boldsymbol{\beta}}$.

Because the exact distributions of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ are unknown, the exact distribution of T is unknown. Under the null, the distribution of T is approximated by a noncentral $F\left(a, \nu_2, \omega\right)$ distribution. In the method proposed by Kenward and Roger [15], T is multiplied by a scale factor to account for the additional variability in $\mathcal{V}\left(\hat{\boldsymbol{\beta}}\right)$ introduced by estimating $\boldsymbol{\Sigma}$. Satterthwaite-style [16] degrees of freedom are then computed for the scaled statistic. Another common choice for denominator degrees of freedom is $(m - g)$, by analogy to the result for balanced data.

Four common testing approaches for the one-stage model are as follows: (i) the Wald test with Kenward and Roger denominator degrees of freedom [15] and variance components constrained positive; (ii) the Wald test with Kenward and Roger denominator degrees of freedom [15] and unconstrained variance components; (iii) the Wald test with denominator degrees of freedom $(m - g)$ and variance components constrained positive; and (iv) the Wald test with denominator degrees of freedom $(m - g)$ and unconstrained variance components. Regardless of the testing approach, the distribution of T is not exact and, consequently, the Type I error rate may not equal α.

2.6. *Tests of fixed effects in two-stage models with unbalanced data*

For the two-stage model with unbalanced data, unequal cluster sizes violate the assumption of homoscedasticity required for the general linear univariate model. Therefore, we make the less restrictive assumption that $\boldsymbol{\Sigma}_T \approx \sigma^2\boldsymbol{W}^{-1}$ [17]. Model estimates are obtained via weighted least squares with $\hat{\boldsymbol{\beta}}_{\mathrm{WLS}} = \left(\boldsymbol{X}_T'\boldsymbol{W}\boldsymbol{X}_T\right)^{-1}\left(\boldsymbol{X}_T'\boldsymbol{W}\boldsymbol{y}_T\right)$. Covariance estimates are obtained using the weighted restricted maximum likelihood estimate for σ^2,

$$\hat{\sigma}^2_{\text{REML}} = \boldsymbol{y}_T' \left[\boldsymbol{W} - \boldsymbol{W}\boldsymbol{X}_T \left(\boldsymbol{X}_T' \boldsymbol{W} \boldsymbol{X}_T \right)^{-1} \boldsymbol{X}_T' \boldsymbol{W} \right] \boldsymbol{y}_T / (m - g). \tag{13}$$

The variance of the corresponding general linear hypothesis, $\hat{\boldsymbol{\theta}}_{\text{WLS}} = \boldsymbol{C}\hat{\boldsymbol{\beta}}_{\text{WLS}}$, is $\hat{\mathbb{V}}\left(\hat{\boldsymbol{\theta}}_{\text{WLS}}\right) = \hat{\sigma}^2_{\text{REML}} \boldsymbol{C} \left(\boldsymbol{X}_T' \boldsymbol{W} \boldsymbol{X}_T \right)^{-1} \boldsymbol{C}'$. When $\boldsymbol{W}$ is known, a uniformly most powerful, size-α test is given by

$$T_T = \left(\hat{\boldsymbol{\theta}}_{\text{WLS}} - \boldsymbol{\theta}_0 \right)' \left[\hat{\mathbb{V}} \left(\hat{\boldsymbol{\theta}}_{\text{WLS}} \right) \right]^{-1} \left(\hat{\boldsymbol{\theta}}_{\text{WLS}} - \boldsymbol{\theta}_0 \right) / a. \tag{14}$$

For the unbalanced two-stage model, the weight matrix $\boldsymbol{W}$ is typically unknown. When $\boldsymbol{W}$ is unknown, T_T has an approximate F distribution. Three common weighting schemes for cluster means in the two-stage model [18] are: (1) no weighting, (2) weighting by cluster size, and 3) weighting by the inverse of the estimated variance of the sample means. Another approach is to obtain estimates of covariance parameters from a one-stage model fit and form the estimated theoretical variance of the cluster means. This weighting scheme preserves the original scale of the outcomes. Finally, some analysts, in error, weight the cluster means by the inverse of the cluster size. Regardless of the choice of weights, the distribution of T_T is not exact and, consequently, the Type I error rate may not equal α. Analytic approaches for the one-stage and two-stage models are summarized in Table I.

Table I. Testing approaches for one-stage and two-stage models.

Method	Model	Details
1	One-stage	Denominator degrees of freedom as described by Kenward and Roger, with variance components constrained to be positive
2	One-stage	Denominator degrees of freedom as described by Kenward and Roger, with unconstrained variance components
3	One-stage	Denominator degrees of freedom $(m - g)$, with variance components constrained to be positive
4	One-stage	Denominator degrees of freedom $(m - g)$, with unconstrained variance components
5	Two-stage	Unweighted (e.g., analysis of variance, t-test) $\boldsymbol{W} = \boldsymbol{I}_m$
6	Two-stage	Weighting by cluster size $\boldsymbol{W} = \text{diag}\left(\boxdot_{h=1}^{g} \boxdot_{i=1}^{m_h} n_{hi} \right)$
7	Two-stage	Incorrect weighting by the inverse of the cluster size $\boldsymbol{W} = \text{diag}\left(\boxdot_{h=1}^{g} \boxdot_{i=1}^{m_h} 1/n_{hi} \right)$
8	Two-stage	Weighting by the inverse of the estimated variance of the sample means $\boldsymbol{W} = \text{diag}\left(\boxdot_{h=1}^{g} \boxdot_{i=1}^{m_h} \left[n_{hi} \left(n_{hi} - 1 \right) \right] / \left[\boldsymbol{y}_{hi}' \boldsymbol{y}_{hi} - n_{hi} \bar{y}_{1,hi} \right] \right)$
9	Two-stage	Weighting by the estimated theoretical variance of the cluster means, with variance constrained positive $\boldsymbol{W} = \text{diag}\left[\boxdot_{h=1}^{g} \boxdot_{i=1}^{m_h} \left(\hat{\sigma}_c^2 + \hat{\sigma}_e^2 / n_{hi} \right)^{-1} \right]$
10	Two-stage	Weighting by the estimated theoretical variance of the cluster means, with unconstrained variance $\boldsymbol{W} = \text{diag}\left[\boxdot_{h=1}^{g} \boxdot_{i=1}^{m_h} \left(\hat{\sigma}_c^2 + \hat{\sigma}_e^2 / n_{hi} \right)^{-1} \right]$

3. Theoretical results

We demonstrate that Type I error rates are exact for one-stage and two-stage models with balanced data. For balanced data, the following two theorems hold. Proofs appear in Appendix A.

Theorem 1
For balanced data such that $n_{hi} \equiv n \; \forall \; h \in \{1, \ldots, g\}$, $i \in \{1, \ldots, m_h\}$, and with $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$, $\widehat{\boldsymbol{\theta}} = \boldsymbol{C}\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{y}'\left[\boldsymbol{I}_m - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right]\boldsymbol{y}/\left(m-g\right)$, and $\widehat{\mathcal{V}}\left(\widehat{\boldsymbol{\theta}}\right) = \widehat{\boldsymbol{\Sigma}}\left[\boldsymbol{C}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{C}'\right]$, the Wald test

$$T = \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \left[\widehat{\mathcal{V}}\left(\widehat{\boldsymbol{\theta}}\right)\right]^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)/a \tag{15}$$

has exact Type I error rate α.

Theorem 2
For balanced data such that $n_{hi} \equiv n \; \forall \; h \in \{1, \ldots, g\}$, $i \in \{1, \ldots, m_h\}$, and with $\widehat{\boldsymbol{\beta}}_T = \left(\boldsymbol{X}_T'\boldsymbol{X}_T\right)^{-1}\boldsymbol{X}_T'\boldsymbol{y}_T$, $\widehat{\boldsymbol{\theta}}_T = \boldsymbol{C}\widehat{\boldsymbol{\beta}}_T$, $\widehat{\sigma}_T^2 = \boldsymbol{y}_T'\left[\boldsymbol{I}_m - \boldsymbol{X}_T\left(\boldsymbol{X}_T'\boldsymbol{X}_T\right)^{-1}\boldsymbol{X}_T'\right]\boldsymbol{y}_T/\left(m-g\right)$, $\widehat{\mathcal{V}}\left(\widehat{\boldsymbol{\theta}}_T\right) = \widehat{\sigma}^2\left[\boldsymbol{C}\left(\boldsymbol{X}_T'\boldsymbol{X}_T\right)^{-1}\boldsymbol{C}'\right]$, and $T_T = \left(\widehat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\right)' \left[\widehat{\mathcal{V}}\left(\widehat{\boldsymbol{\theta}}_T\right)\right]^{-1} \left(\widehat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\right)/a$, the hypothesis test for the one-stage model is equivalent to the hypothesis test for the two-stage model and

$$\boldsymbol{T}_T = \boldsymbol{T}. \tag{16}$$

Theorem 1 implies that an exact size α test can be formed for the one-stage model. As a result, the Type I error rate will be exactly the α level specified in the study design. *Theorem 2* implies that the balanced one-stage model and balanced two-stage model produce equivalent hypothesis tests. Therefore, the balanced two-stage model will also have a Type I error rate exactly equal to α. For unbalanced data, *Theorems* 1 and 2 no longer hold.

4. Simulation study

4.1. Simulation methods

We performed a simulation study to compare the Type I error performance of four analytic methods for the unbalanced one-stage model and six methods for the unbalanced two-stage model for cluster randomized designs (Table I). The analytic methods were applied to a cluster randomized design with two treatment groups and one level of clustering, under several scenarios of cluster size imbalance. The majority of cluster-randomized trials reviewed by Varnell *et al.* [7] and Murray *et al.* [6] used a two group design, indicating the study design is commonly used in practice.

One-stage methods were fit using PROC GLIMMIX, and two-stage methods were fit using PROC GLM with SAS version 9.3 [19]. We used the double-dogleg optimization method with PROC GLIMMIX because of its superior convergence performance. By default, PROC GLIMMIX produces variance estimates constrained to be positive. For methods requiring unconstrained variance estimates, we used the PARMS statement in PROC GLIMMIX to specify a lower bound of -1 for all variance estimates. The bound of -1 effectively provided unconstrained variance estimates for the specific combinations of intracluster correlation and standard deviations used in our simulation study. In general, the NOBOUND option on the PARMS statement provides fully unconstrained variance estimates.

4.2. Scenarios of cluster size imbalance

For each analytic method, empirical Type I error was calculated for 9,090 different experimental scenarios. We restricted our attention to designs with two treatment groups, because this covers approximately 88% of all group-randomized trials reviewed by Murray *et al.* [20].

Each scenario was characterized by six parameters: the number of clusters per treatment group, denoted as m_1 and m_2, the average cluster sizes of each treatment group, denoted as $\bar{n}_1$ and $\bar{n}_2$, the ratio of maximum to minimum cluster sizes for each treatment group, denoted as r_1 and r_2, and the within-cluster correlation, ρ. We assumed that ρ was equal in both treatment groups. The values of m_1 and m_2 were $\left(m_1, m_2\right) \in \{(2,4), (3,4), (4,4), (6,8), (7,8), (8,8), (14,16), (15,16), (16,16)\}$. The values of $\bar{n}_1$ and

$\bar{n}_2$ were $\{8, 16, 32, 64, 128\}$. Values of r_1 and r_2 were $\{1, 2, 4, 8\}$. The values of ρ were $\{0.001, 0.01, 0.1\}$. All values were chosen to reflect common study designs used in cluster-randomized trials.

A full factorial design would yield 18,000 possible study designs. We restricted our simulation to combinations of the parameters, which produced unique study designs. Unique cases included those for which $\left(m_1 < m_2\right)$, $\left(m_1 = m_2 \text{ and } \bar{n}_1 < \bar{n}_2\right)$ or $\left(m_1 = m_2\ , \bar{n}_1 = \bar{n}_2, \text{ and } r_1 \leq r_2\right)$.

4.3. Calculation of empirical Type I error rate

We calculated the empirical Type I error rate for each scenario of cluster imbalance and analytic method. First, cluster sizes were generated as follows. For each treatment group, we selected m_h evenly spaced points, p_i, on the interval $\{0.025, 0.975\}$. Cluster sizes were then computed as

$$n_{hi} = \Phi^{-1}\left(p_i\right) \times \left[\bar{n}_h r_h / 2\left(1 + r_h\right)\right]^2 + \bar{n}_h, \tag{17}$$

rounded to the nearest integer. Equation 17 assumes that cluster sizes follow a doubly truncated normal distribution, with mean $\bar{n}_h$ and variance proportional to the mean. Second, we simulated 20,000 replicates of $\boldsymbol{e}^* = \boldsymbol{Zd} + \boldsymbol{e}$ with variance as given in Equation 2. We then computed $\boldsymbol{y}$ as in Equation 1, with $\boldsymbol{\beta}$ selected such that there was no difference between the treatment groups. For each replicate, we tested the null hypothesis of no difference between the treatment groups. The empirical Type I error for each scenario and analytic method was calculated as the proportion of replicates in which the null hypothesis was rejected.

4.4. Four metrics for evaluating the Type I error performance

We expect a method with good Type I error properties to achieve the nominal Type I error rate of 0.05. For an exact test, the null and non-null distributions of the test statistic are known without error, and one can show analytically that the analytic method will behave reasonably for any data set. However, for the 10 analytic methods considered in this paper, none are exact. Thus, any evaluation of Type I error, or other characteristics of the analytic method is dependent on the choice of experimental scenarios.

We use four metrics to assess the Type I error performance of the analytic methods over all 9,090 experimental scenarios. First, we give a box-plot of the observed Type I error values. Second, we provide a lookup table. The lookup table allows scientists to investigate the Type I error performance of the analytic methods in scenarios that best match their intended design. Third, we use a general linear multivariate model to assign significance to the comparison of the analytic methods. Fourth, we identify methods that reliably generated a Type I error rate between 0.04 and 0.06, which defines a 20% tolerance around the desired rate of 0.05. We refer to this metric as the '20% tolerance' metric.

We use the general linear multivariate model for several reasons. First, in large samples, we expect that the Type I error of an exact size α test will follow an approximately normal distribution, as shown in Appendix A. Because our simulation included 9,090 experimental scenarios, the assumption of normality should hold. Second, because each experimental scenario is independent, the general linear model assumption of independence is valid. Lastly, the results of the 10 analytic methods are correlated, because the analytic methods are applied to the same data sets. The general linear multivariate model allows us to account for this correlation.

We considered, and rejected, using the Kolmogorov–Smirnov test [21] to compare the analytic methods. The Kolmogorov–Smirnov test would have allowed us to compare the observed Type I error distributions of the 10 tests with a theoretical binomial distribution. However, for large sample sizes, the Kolmogorov–Smirnov test [22] tends to reject for minor deviations from the theoretical distribution. Given our simulation sample size of 9,090, we decided that the Kolmogorov–Smirnov would be ineffective at identifying methods with reasonable Type I error properties.

The choice of which metric to use depends on the view of the researcher. Researchers who want a gestalt recommendation will be best served by looking at the box plot in Figure 1. If a scientist knows what the anticipated cluster sizes and correlation will be for a specific design, the scientist should consult the lookup table and choose a method that has the best Type I error properties for her specific study design. Researchers who want a method that performs well for a general class of study designs should examine the results of the modeling approach. Finally, some investigators prefer methods with Type I error rates that do not 'exceed the nominal size by 20% or more' [23]. For those investigators, we suggest the '20% tolerance metric'.

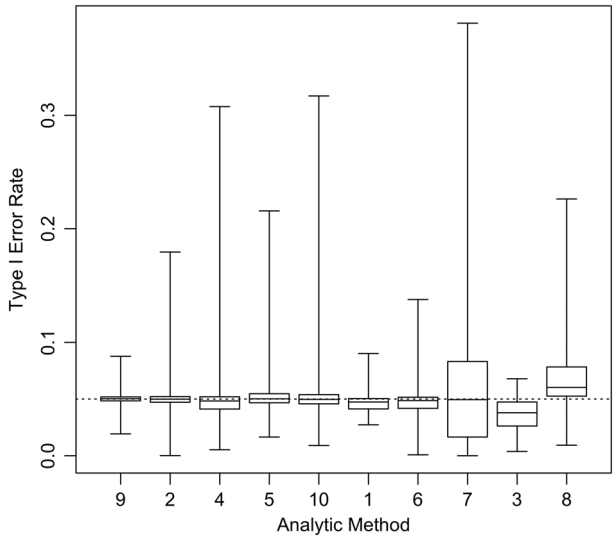

Figure 1. Empirical Type I error for each test, sorted by increasing mean difference from 0.05.

4.5. Methods for the four metrics

Box-plot: We produced a box-plot (Figure 1) showing the Type I error rate results for the 10 analytic methods and all experimental scenarios. The upper and lower hinges of the box represent the 75th and 25th quantiles of the distribution. Error bars extend to the maximum and minimum observed Type I error rate for the analytic method.

Lookup table: The lookup table is available in the SampleSizeShop.org website. A set of dropdown menus allows users to select experimental design parameters. To access the table, either click the hotlink above, or go to SampleSizeShop.org and click on the 'Downloads' tab.

Hypothesis testing: We used a general linear multivariate model to compare the empirical Type I error rates observed in the simulation against the desired Type I error rate of 0.05. The outcomes for the general linear multivariate model were the difference scores between the observed empirical Type I error rates for each analytic method and 0.05. Predictors included intracluster correlation, ρ, the ratio of maximum to minimum cluster sizes for treatment, r_1, and control, r_2, and the total sample size N.

We selected 16 sets of predictor values at which to evaluate the Type I error rates. We included all possible combinations of $\rho \in \{0.1, 0.001\}$, $\left(r_1, r_2\right) \in \{(8,8), (8,1), (1,1), (4,4)\}$, and $N \in \{100, 1000\}$. We chose these values to reflect a variety of correlations, sample sizes, and cluster size imbalance parameters. For each set of values, we performed an omnibus test and then, if the omnibus test were significant, 10 step-down tests. The omnibus test evaluated the null hypothesis that every analytic method achieved an equal Type I error rate of 0.05. The 10 step-down tests were used to identify the analytic method, which showed the deviation. For each set of predictor values, we corrected for multiple comparisons using an α-spending approach. We used an α level of 0.04 for each omnibus test, and Bonferroni corrected $\alpha = 0.01/10 = 0.001$ for the respective step-down tests.

20% Tolerance: To provide a systematic summary of the results available from the lookup table, we examined all 9,090 cells for all 10 methods to identify conditions under which each method had a Type I error rate between 0.04 and 0.06.

4.6. Results for the four metrics

Box-plot: The Type I error performance for each analytic method across all scenarios are shown in Figure 1. Over 9090 scenarios, an exact size α test will achieve an average Type I error rate of 0.05, with 75% quartile at 0.050001 and 25% quartile at 0.049999. Figure 1 suggests that Methods 9, 1, and 2 have a similar distribution to that of an exact size α test. Although Method 2 had appropriate quartiles, it achieved Type I error values as high as 0.18 in some cases.

For Method 9, any inaccuracy in Type I error appears to be independent of the misspecification of ρ. Across all 9090 scenarios, for Method 9, $\hat{\rho} - \rho$ is no more than 0.04 and no smaller than -0.01, with the absolute error decreasing rapidly as $m_1\bar{n}_1 + m_2\bar{n}_2$ increases and as ρ increases.

Column	Filter
Intracluster correlation (rho)	rho
Number of clusters in treatment group 1 (m1)	From min to max
Number of clusters in treatment group 2 (m2)	From min to max
Average cluster size in treatment group 1 (nbar1)	From min to max
Average cluster size in treatment group 2 (nbar2)	From min to max
The ratio of maximum to minimum cluster size in treatment group 1 (r1)	From min to max
The ratio of maximum to minimum cluster size in treatment group 2 (r2)	From min to max
	Clear Filters

Show 10 entries Search:

rho	m1	m2	nbar1	nbar2	r1	r2	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7	Method 8	Method 9	Method 10
0.001	2	4	8	8	1	1	0.0308	0.0499	0.0051	0.0499	0.0499	0.0499	0.0499	0.0687	0.0499	0.0499
0.001	2	4	8	8	1	2	0.0309	0.019	0.0052	0.0557	0.0492	0.0514	0.0518	0.0666	0.0515	0.0653
0.001	2	4	8	8	1	4	0.0321	0.0095	0.005	0.0521	0.0415	0.0503	0.0469	0.0631	0.0504	0.0725
0.001	2	4	8	8	1	8	0.0294	0.0065	0.0058	0.0479	0.0386	0.0509	0.05	0.0617	0.0508	0.0737
0.001	2	4	8	8	2	1	0.0273	0.0417	0.006	0.0891	0.0495	0.0491	0.0694	0.0686	0.0491	0.103
0.001	2	4	8	8	2	2	0.0313	0.0024	0.006	0.1799	0.0491	0.0515	0.0702	0.0682	0.0516	0.1853
0.001	2	4	8	8	2	4	0.0314	0.0019	0.0061	0.0952	0.0448	0.0494	0.0731	0.0645	0.0493	0.1159
0.001	2	4	8	8	2	8	0.0297	0.0009	0.0054	0.0843	0.0378	0.0506	0.0695	0.0638	0.0506	0.1098
0.001	2	4	8	8	4	1	0.0294	0.0793	0.0054	0.106	0.0502	0.0517	0.1416	0.0748	0.0514	0.1326
0.001	2	4	8	8	4	2	0.0275	0.0744	0.0052	0.1305	0.0479	0.049	0.1453	0.0726	0.0486	0.1544

Showing 1 to 10 of 9,090 entries First Previous 1 2 3 4 5 Next Last

Figure 2. Lookup table results for Method 9.

Table II. Predicted Type I error of analytic methods for designs with $\rho = 0.001$ and $N = 100$.

r_1	8	8	1	4
r_2	8	1	1	4
Method				
1	0.0413 (0.0410, 0.0416)	0.0413 (0.0410, 0.0416)	0.0411 (0.0409, 0.0413)	0.0412 (0.0410, 0.0414)
2	* 0.0506 (0.0496, 0.0517)	0.0570 (0.0560, 0.0580)	0.0531 (0.0523, 0.0539)	0.0520 (0.0514, 0.0526)
3	0.0224 (0.0219, 0.0229)	0.0214 (0.0209, 0.0219)	0.0204 (0.0200, 0.0208)	0.0213 (0.0210, 0.0216)
4	0.0547 (0.0536, 0.0559)	* 0.0488 (0.0478, 0.0499)	0.0418 (0.0409, 0.0427)	0.0473 (0.0467, 0.0480)
5	0.0515 (0.0506, 0.0523)	0.0553 (0.0545, 0.0561)	0.0556 (0.0549, 0.0562)	0.0538 (0.0533, 0.0543)
6	0.0554 (0.0546, 0.0562)	0.0516 (0.0509, 0.0523)	0.0472 (0.0466, 0.0478)	* 0.0507 (0.0503, 0.0512)
7	0.1218 (0.1192, 0.1244)	0.0871 (0.0848, 0.0895)	0.0385 (0.0364, 0.0405)	0.0742 (0.0727, 0.0757)
8	0.1057 (0.1040, 0.1074)	0.0925 (0.0909, 0.0940)	0.0695 (0.0681, 0.0708)	0.0850 (0.0840, 0.0860)
9	0.0525 (0.0522, 0.0528)	0.0510 (0.0508, 0.0513)	0.0491 (0.0489, 0.0493)	0.0505 (0.0504, 0.0507)
10	0.0689 (0.0678, 0.0701)	0.0608 (0.0598, 0.0619)	* 0.0506 (0.0497, 0.0515)	0.0585 (0.0578, 0.0592)

* Statistically similar to an exact size $\alpha = 0.05$ test.

Lookup table: To illustrate the use of the lookup table, we retrieved the Type I error rate for Method 9 for a study design with low intracluster correlation, four clusters per treatment group and 30 to 40 observations per cluster. We accessed the lookup table in the SampleSizeShop.org website. We filtered the results to match our study design as shown in Figure 2. The column labeled 'Method 9' showed that empirical Type I error values for the method ranged from 0.0492 to 0.0526. Thus, Method 9 would be a good analytic approach for the planned study. In contrast, Type I error rates for Method 10 ranged from 0.0492 to 0.1859, indicating that Type I error inflation could occur were we to perform the analysis with Method 10.

Hypothesis testing: For the 16 sets of fixed values for ρ, r_1, r_2, and N, we rejected the null hypothesis that every analytic method achieved an equal Type I error rate of 0.05. The results of the step-down tests and predicted Type I error rates are summarized in Tables II through V.

There was no clear winner among the analytic methods. For designs with highly unbalanced data in both treatment groups ($r_1 = 8, r_2 = 8$), Method 2 provided the best Type I error performance, regardless of sample size or intracluster correlation. For designs with $r_1 = 8$, $r_2 = 8$, and $\rho = 0.1$, Method 6 performed

Table III. Predicted Type I error of analytic methods for designs with $\rho = 0.001$ and $N = 1000$.

r_1	8	8	1	4
r_2	8	1	1	4
Method				
1	0.0429 (0.0426, 0.0432)	0.0428 (0.0426, 0.0431)	0.0426 (0.0424, 0.0428)	0.0427 (0.0426, 0.0429)
2	* 0.0498 (0.0488, 0.0507)	0.0561 (0.0553, 0.0570)	0.0522 (0.0515, 0.0530)	0.0512 (0.0507, 0.0517)
3	0.0295 (0.0289, 0.0300)	0.0285 (0.0280, 0.0289)	0.0275 (0.0271, 0.0279)	0.0283 (0.0281, 0.0286)
4	0.0539 (0.0528, 0.0550)	0.0480 (0.0470, 0.0489)	0.0410 (0.0402, 0.0418)	0.0465 (0.0460, 0.0471)
5	0.0515 (0.0507, 0.0523)	0.0553 (0.0546, 0.0560)	0.0555 (0.0549, 0.0562)	0.0538 (0.0534, 0.0542)
6	0.0519 (0.0512, 0.0527)	0.0482 (0.0475, 0.0488)	0.0437 (0.0432, 0.0443)	0.0472 (0.0469, 0.0476)
7	0.1099 (0.1075, 0.1123)	0.0752 (0.0731, 0.0774)	0.0266 (0.0247, 0.0284)	0.0623 (0.0611, 0.0635)
8	0.0962 (0.0946, 0.0979)	0.0830 (0.0815, 0.0844)	0.0600 (0.0588, 0.0612)	0.0755 (0.0747, 0.0763)
9	0.0519 (0.0517, 0.0522)	0.0505 (0.0502, 0.0507)	0.0486 (0.0484, 0.0487)	* 0.0500 (0.0499, 0.0501)
10	0.0648 (0.0637, 0.0659)	0.0567 (0.0557, 0.0577)	0.0465 (0.0456, 0.0473)	0.0543 (0.0538,0.0549)

* Statistically similar to an exact size $\alpha = 0.05$ test.

Table IV. Predicted Type I error of analytic methods for designs with $\rho = 0.1$ and $N = 100$.

r_1	8	8	1	4
r_2	8	1	1	4
Method				
1	0.0512 (0.0509, 0.0515)	0.0511 (0.0508, 0.0514)	0.0509 (0.0507, 0.0512)	0.0510 (0.0508, 0.0513)
2	* 0.0493 (0.0482, 0.0505)	0.0557 (0.0547, 0.0568)	0.0518 (0.0509, 0.0527)	* 0.0508 (0.0500, 0.0515)
3	0.0424 (0.0418, 0.0429)	0.0414 (0.0408, 0.0419)	0.0404 (0.0399, 0.0409)	0.0412 (0.0408, 0.0416)
4	0.0592 (0.0580, 0.0605)	0.0533 (0.0522, 0.0545)	0.0463 (0.0453, 0.0473)	0.0519 (0.0510, 0.0527)
5	0.0480 (0.0471, 0.0490)	0.0518 (0.0510, 0.0527)	0.0521 (0.0514, 0.0529)	* 0.0504 (0.0497, 0.0510)
6	* 0.0492 (0.0484, 0.0500)	0.0454 (0.0447, 0.0462)	0.0410 (0.0403, 0.0417)	0.0445 (0.0440, 0.0451)
7	0.1144 (0.1116, 0.1172)	0.0797 (0.0772, 0.0823)	0.0311 (0.0288, 0.0333)	0.0668 (0.0649, 0.0686)
8	0.0904 (0.0885, 0.0923)	0.0772 (0.0754, 0.0789)	0.0541 (0.0526, 0.0557)	0.0697 (0.0684, 0.0709)
9	0.0530 (0.0527, 0.0533)	0.0515 (0.0513, 0.0518)	* 0.0496 (0.0494, 0.0499)	0.0511 (0.0509, 0.0513)
10	0.0667 (0.0654, 0.0679)	0.0586 (0.0574, 0.0598)	* 0.0484 (0.0473, 0.0494)	0.0562 (0.0554, 0.0571)

* Statistically similar to an exact size $\alpha = 0.05$ test.

Table V. Predicted Type I error of analytic methods for designs with $\rho = 0.1$ and $N = 1000$.

r_1	8	8	1	4
r_2	8	1	1	4
Method				
1	0.0527 (0.0524, 0.0530)	0.0527 (0.0524, 0.0529)	0.0525 (0.0522, 0.0527)	0.0526 (0.0524, 0.0528)
2	* 0.0485 (0.0474, 0.0496)	0.0549 (0.0539, 0.0559)	* 0.0510 (0.0501, 0.0518)	* 0.0499 (0.0492, 0.0506)
3	* 0.0494 (0.0489, 0.0500)	0.0484 (0.0479, 0.0490)	0.0474 (0.0470, 0.0479)	0.0483 (0.0479, 0.0486)
4	0.0584 (0.0572, 0.0596)	0.0525 (0.0514, 0.0536)	0.0455 (0.0446, 0.0464)	* 0.0510 (0.0503, 0.0518)
5	0.0480 (0.0471, 0.0489)	0.0518 (0.0510, 0.0526)	0.0521 (0.0514, 0.0528)	* 0.0503 (0.0498, 0.0509)
6	0.0457 (0.0450, 0.0465)	0.0420 (0.0412, 0.0427)	0.0375 (0.0369, 0.0382)	0.0411 (0.0406, 0.0415)
7	0.1025 (0.0998, 0.1051)	0.0678 (0.0654, 0.0703)	0.0192 (0.0170, 0.0213)	0.0549 (0.0532, 0.0565)
8	0.0809 (0.0791, 0.0827)	0.0677 (0.0660, 0.0693)	0.0446 (0.0432, 0.0461)	0.0602 (0.0591, 0.0613)
9	0.0525 (0.0522, 0.0528)	0.0510 (0.0507, 0.0513)	0.0491 (0.0488, 0.0493)	0.0505 (0.0504, 0.0507)
10	0.0625 (0.0613, 0.0637)	0.0544 (0.0533, 0.0555)	0.0442 (0.0432, 0.0452)	0.0520 (0.0513,0.0528)

* Statistically similar to an exact size $\alpha = 0.05$ test.

well with $N = 100$, and Method 3 performed well with $N = 1000$. In designs with highly unbalanced data in only one treatment group ($r_1 = 8, r_2 = 1$), Type I error performance was less consistent, although Method 4 showed good Type I error performance with $\rho = 0.001$ and $N = 100$. With moderate imbalance ($r_1 = 4, r_2 = 4$), Methods 2 and 5 performed well for $\rho = 0.1$. Methods 4, 6, and 9 performed well with moderate imbalance but only for specific sample size and intracluster correlation combinations. For balanced data ($r_1 = 1, r_2 = 1$), Methods 2, 9, and 10 tended to have the best Type I error characteristics.

20% Tolerance: Every method yielded a Type I error rate between 0.04 and 0.06 for some experimental scenarios. The Type I error rate was more likely to fall in this range for scenarios involving (1) more clusters per arm (14 to 16 were better than 6 to 8, which were better than 2 to 4), (2) a larger value for ρ (0.1 was better than 0.01, which was better than 0.001), and a larger average cluster size (128 was better than 64, etc.). Perhaps most striking was that Methods 2, 9, and 10 generated Type I error rates in this range whenever the number of clusters per arm was at least 14, regardless of the levels of the other factors.

Also striking were the results for Methods 3 and 4, which are commonly used by investigators who employ group-randomized or cluster- randomized designs. Method 3 was very conservative when ρ is small, as has been reported previously [24]. Because small ρ values are common in cluster-randomized designs, Method 3 cannot be recommended. The simulation confirmed that Method 4 performed much better than Method 3. However, Method 4 did not perform well with only two to four clusters per arm and did not perform well consistently with six to eight clusters per arm. These results suggest that Method 4 cannot be recommended for smaller studies. Finally, we must note that none of the methods performed well across all combinations of the factors examined in the simulations, even by the 20% tolerance metric. Even Method 9, which generated a Type I error rate in this range more often than any of the other methods, performed poorly in small studies (two to four clusters per arm) with modest to large values of ρ $(0.01 - 0.1)$.

5. Discussion

Our simulation study indicates that the choice of test in unbalanced, cluster randomized trials can dramatically impact Type I error rates. None of the methods tested matched an exact size α test in all situations. Based on hypothesis testing from the multivariate model, Method 2 performed well in seven of 16 test cases. Recall that Method 2 was the one-stage model using Kenward and Roger [15] denominator degrees of freedom and allowing negative variance components. Based on the box-plot, Method 9 had the best Type I error distribution across all simulated scenarios. Method 9 was the two-stage model weighted by the inverse of the estimated theoretical variance of the cluster means, with variance components constrained positive.

Based on the 20% tolerance metric, Methods 2, 9, and 10 performed well in studies with at least 14 clusters per arm. Method 10 was identical to Method 9, but with variance unconstrained. Investigators could safely use Methods 2, 9, or 10 to analyze data from a study with at least 14 clusters per arm, regardless of the average cluster size, the value of ρ, or the degree of imbalance. Method 9 performed better than the other models in smaller studies and reliably generated a Type I error rate between 0.04 and 0.06 in studies with at least six clusters per arm and 16 participants per cluster. No model performed well in studies with only two to four clusters per arm. Such studies would have very limited power based on very limited degrees of freedom and are not recommended except for pilot studies. In practice, Methods 3 and 4 are used frequently by investigators designing cluster-randomized trials. Given the poor performance of the methods for small values of ρ and smaller sample sizes, the method of Kenward and Roger is recommended for studies using one-stage analyses.

Although our simulation included a large set of scenarios of imbalance, the set may not be representative of the true distribution of cluster randomized trial designs that appear in the literature. In future research, it may be more valuable to sample published cluster randomized trial designs. A sample would provide values for cluster size and ratio of maximum cluster to minimum cluster sizes that reflect true scientific practice. Also, although the methods presented allow for more than two treatment groups, the simulation study only covered cases with two treatment groups. In practice, with multiple treatment groups, one would typically conduct an omnibus test, followed by step-down, pairwise comparisons of the groups. In addition, our results are currently limited to Gaussian outcomes. Further research is needed to extend the work to binary or count data. Finally, we made strong assumptions of equal intracluster correlations.

A majority of group-randomized trials include only two treatment groups [20] and no step-down testing, and thus multiple comparison correction is not required. Analysis of variance (ANOVA) tests and associated multiple comparison methods are readily available for studies with three or more groups [25, 26]. Although we did not investigate Type I error specifically for multiple comparisons methods, extensive analytic and simulation results for linear model hypothesis testing with multiple comparisons, in our opinion, obviate the need for separate treatment. Finally, we were always careful to cast all results in terms of general ANOVA models. We simulated studies with two treatment groups to simplify reporting the results.

As a general purpose approach, we recommend Method 9 for studies having at least six clusters per arm. Methods 2 and 10 are also recommended for studies having at least 14 clusters per arm. For scientists with detailed knowledge of the pattern of imbalance in their study design, we suggest using the online lookup table to select an appropriate analytic method. We hope that this paper guides scientists in selecting appropriate statistical methods for cluster randomized trials.

Appendix A: Proofs

Lemma 1
With $\boldsymbol{A}\ (m \times N) = \text{diag}\left(\boxdot_{h=1}^{g} \boxdot_{i=1}^{m_h} \mathbf{1}'_{n_{hi}}/n_{hi}\right)$, $\boldsymbol{A}\boldsymbol{y} = \boldsymbol{y}_T$. In particular, for the balanced case with $n_{hi} \equiv n\ \forall$ $h \in \{1, \ldots, g\}$, $i \in \{1, \ldots, m_h\}$, $\boldsymbol{A}_b = \frac{1}{n}\text{diag}\left(\boxdot_{h=1}^{g} \boxdot_{i=1}^{m_h} \mathbf{1}'_n\right)$, $\boldsymbol{A}'_b\boldsymbol{A}_b = \frac{1}{n^2}\text{diag}\left(\boxdot_{h=1}^{g} \boxdot_{i=1}^{m_h} \mathbf{1}_n\mathbf{1}'_n\right)$ and $\boldsymbol{A}_b\boldsymbol{y} = \boldsymbol{y}_T$.

Proof of Lemma 1
By inspection.

Lemma 2
For the one-stage model with $n_{hi} \equiv n\ \forall\ h \in \{1, \ldots, g\}$, $i \in \{1, \ldots, m_h\}$, $\widehat{\boldsymbol{\beta}}_{WLS} = \widehat{\boldsymbol{\beta}}$.

Proof of Lemma 2
With balanced data, $n_{hi} \equiv n\ \forall\ h \in \{1, \ldots, g\}$, $i \in \{1, \ldots, m_h\}$, $\boldsymbol{X} = \text{diag}\left(\boxdot_{h=1}^{g} \mathbf{1}_{m_h n}\right)$. In addition, $\boldsymbol{\Sigma}$ reduces to

$$\boldsymbol{\Sigma} = \boldsymbol{I}_m \otimes \left\{\sigma_y^2\left[\mathbf{1}_n\mathbf{1}'_n\rho + \boldsymbol{I}_n(1-\rho)\right]\right\}. \tag{A.1}$$

Thus, $\widehat{\boldsymbol{\Sigma}} = I_m \otimes \left\{\widehat{\sigma}_y^2\left[1_n 1'_n\widehat{\rho} + I_n\left(1-\widehat{\rho}\right)\right]\right\}$ and

$$\widehat{\boldsymbol{\Sigma}}^{-1} = \boldsymbol{I}_m \otimes \frac{1}{\widehat{\sigma}_y^2\left(1-\widehat{\rho}\right)}\left\{\boldsymbol{I}_n - \frac{\widehat{\rho}}{\left[1+(n-1)\widehat{\rho}\right]}\mathbf{1}_n\mathbf{1}'_n\right\}. \tag{A.2}$$

Then $\boldsymbol{X}'\widehat{\boldsymbol{\Sigma}}^{-1} = \left\{\widehat{\sigma}_y^2\left[1+(n-1)\widehat{\rho}\right]\right\}^{-1}\boldsymbol{X}'$, and

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{WLS} &= \left(\boldsymbol{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{y} \\
&= \left(\left\{\widehat{\sigma}_y^2\left[1+(n-1)\widehat{\rho}\right]\right\}^{-1}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left\{\widehat{\sigma}_y^2\left[1+(n-1)\widehat{\rho}\right]\right\}^{-1}\boldsymbol{X}'\boldsymbol{y} \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} \\
&= \widehat{\boldsymbol{\beta}}.
\end{aligned} \tag{A.3}$$

Proof of Theorem 1
Based on *Lemma 2*, we can obtain estimates for treatment means in the balanced one-stage model with $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Y}\right)$. Thus, by standard linear model theory $\widehat{\boldsymbol{\beta}} \sim \mathcal{N}\left[\boldsymbol{\beta}, \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right]$ and $\widehat{\boldsymbol{\theta}} \sim \mathcal{N}_a\left[\boldsymbol{\theta}, \sigma^2\boldsymbol{C}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{C}'\right]$, with $\sigma^2 = \left(\sigma_y^2/n\right)\{1+(n-1)\rho\}$. Thus, T (Equation 15) is a uniformly most powerful size α test for the general linear hypothesis [8, p. 51].

Proof of Theorem 2
We first show that (i) $\widehat{\boldsymbol{\theta}}_T = \widehat{\boldsymbol{\theta}}$ and (ii) $\widehat{\mathbb{V}}\left(\widehat{\boldsymbol{\theta}}_T\right) = \widehat{\mathbb{V}}\left(\widehat{\boldsymbol{\theta}}\right)$. With $\boldsymbol{X}$ and $\boldsymbol{y}$ defined in Equation (1), and $\boldsymbol{X}_T$ and $\boldsymbol{y}_T$ defined in Equation (7), and $\widehat{\boldsymbol{\beta}}$ defined as in *Theorem 1*, $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_T$. Thus (i) holds.

Substituting the expression for $\hat{\sigma}_T^2$ into the expression for $\hat{\mathcal{V}}\left(\hat{\theta}_T\right)$, and invoking *Lemma 1* for the balanced case,

$$\begin{aligned}\hat{\mathcal{V}}\left(\hat{\theta}_T\right) &= \frac{1}{m-g}\left\{\boldsymbol{y}'\boldsymbol{A}_b'\left[\boldsymbol{I}_m - \boldsymbol{A}_b\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{A}_b'\boldsymbol{A}_b\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{A}_b'\right]\boldsymbol{A}_b\boldsymbol{y}\boldsymbol{C}\left(\boldsymbol{X}'\boldsymbol{A}_b'\boldsymbol{A}_b\boldsymbol{X}\right)^{-1}\boldsymbol{C}'\right\} \\ &= \left\{\boldsymbol{y}'\left[\boldsymbol{I}_m - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right]\boldsymbol{y}/\left(m-g\right)\right\}\left\{\boldsymbol{C}\left(\boldsymbol{X}'\boldsymbol{X}/n\right)^{-1}\boldsymbol{C}'\right\} \\ &= \hat{\mathcal{V}}\left(\hat{\theta}\right).\end{aligned} \tag{A.4}$$

Thus (ii) holds. Because (i) and (ii) hold and θ_0 is constant, then $T_T = T$.

A.1. Distribution of Type I error values

We derive a normal approximation to the distribution of Type I error for the simulation described in Section 4. Let α be the Type I error rate desired in the study design. Let Q be the number of experimental scenarios considered. Let N be the number of hypothesis decisions performed for each imbalance scenario. Define R_s to be the number of rejections of the null hypothesis in the N hypothesis decisions for imbalance scenario s. Assuming a true null hypothesis, we would expect R_s to be binomial with $R_s \sim \mathcal{B}(N, \alpha)$ for an exact size α test. The empirical estimate of Type I error for a scenario s is R_s/N. The normal approximation to the binomial suggests that $R_s/N \dot{\sim} \mathcal{N}\left[\alpha, \alpha(1-\alpha)/N\right]$. The empirical estimate of the size of the test is $\hat{\alpha} = \sum_{q=1}^{Q}\left(R_s/N\right)/Q$. Gaussian theory suggests that

$$\hat{\alpha}_t \dot{\sim} \mathcal{N}\left[\alpha, \alpha(1-\alpha)/\left(NQ^2\right)\right]. \tag{A.5}$$

Appendix B: Example SAS code for method 9

The following SAS code implements the two-stage model weighted by the inverse of the estimated theoretical variance of the cluster means (Method 9 in Table I). The indicator for treatment group (control or treatment) is *treatment*, the response variable is called *outcome*, and cluster is specified by the *cluster* variable. The estimates $\hat{\sigma}_c^2$ and $\hat{\sigma}_e^2$ are obtained from PROC GLIMMIX. The values are output to the COV data set and stored as the variables CVEST and EVEST.

```
/* Calculate cluster means */
PROC MEANS DATA=clusteredDataSet NWAY;
  CLASS cluster treatment;
  VAR outcome;
  OUTPUT OUT=means MEAN=mean N=size;
RUN;
/*
 * Get estimates of the within cluster variance
 * and error variance
 */
ODS OUTPUT COVPARMS=COV;
PROC GLIMMIX data=clusteredDataSet;
  CLASS cluster treatment;
  MODEL outcome = treatment / DDFM=KR;
  RANDOM INTERCEPT / SUBJECT=cluster;
  NLOPTIONS TECHNIQUE=DBLDOG;
RUN;
/*
* Move the covariance estimates onto
* a single row
*/
DATA COV1;
  RETAIN CVEST EVEST;
```

```
  SET COV;
  * within cluster covariance;
  IF COVPARM="Intercept" THEN CVEST=ESTIMATE;
  * residual covariance;
  IF COVPARM="Residual" THEN EVEST=ESTIMATE;
  IF CVEST ne . & EVEST ne . THEN output;
  KEEP CVEST EVEST;
RUN;
/* Compute the variance weights */
DATA clusterMeansCovar;
  MERGE means;
  IF _n_ eq 1 then SET cov1;
  EVMIXED = 1/(CVEST + EVEST/size);
RUN;
/*
 * Run Test 9 -- two-stage model with means weighted by the
 * theoretical variance of each
 * cluster mean and variance parms constrained to be positive
 *
 */
/** Covariance parameters used to compute weights come from Test 1
**/
PROC GLM DATA=clusterMeansCovar;
  CLASS treatment;
  WEIGHT evmixed;
  MODEL mean = treatment / clparm;
  ESTIMATE 'TRT DIFF' treatment 1 -1;
RUN;
```

Acknowledgements

A portion of this paper was submitted to the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the DrPH in Biostatistics for J. L. Johnson.

D. M. Murray completed work on the research for this paper before accepting a position at the National Institutes of Health.

Supported by the National Institute of Dental and Craniofacial Research under award NIDCR 1 R01 DE020832-01A1, Multilevel and Longitudinal Study Sample Size Tools for Behavioral Scientists (12/09/10-11/30/14), led by Keith Muller and Deborah Glueck

References

1. Murray DM. *Design and Analysis of Group-Randomized Trials*. Oxford University Press: US, 1998.
2. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold: New York, 2000.
3. Hayes RJ, Moulton LH. *Cluster Randomised Trials* 1st ed. Chapman and Hall/CRC: Boca Raton, 2009.
4. Crespi CM, Maxwell AE, Wu S. Cluster randomized trials of cancer screening interventions: are appropriate statistical methods being used? *Contemporary Clinical Trials* 2011; **32**(4):477–484.
5. Cornfield J. Randomization by group: a formal analysis. *American Journal of Epidemiology* 1978; **108**(2):100–102.
6. Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. *American Journal of Public Health* 2004; **94**(3):423–432.
7. Varnell SP, Murray DM, Janega JB, Blitstein JL. Design and analysis of group-randomized trials: a review of recent practices. *American Journal of Public Health* 2004; **94**(3):393–399.
8. Muller KE, Stewart PW. *Linear Model Theory: Univariate, Multivariate, and Mixed Models*. John Wiley and Sons: Hoboken, New Jersey, 2006.
9. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; **38**(4):963–974.
10. Muller KE, Fetterman BA. *Regression and ANOVA: An Integrated Approach Using SAS Software*. SAS Institute: Cary, NC, 2002.
11. Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data* (1st edn). Springer: New York, 2009.
12. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 1977; **39**(1):1–38.

13. Kackar RN, Harville DA. Approximations for standard errors of estimators of fixed and random effect in mixed linear models. *Journal of the American Statistical Association* 1984; **79**(388):853–862.
14. Dempster AP, Rubin DB, Tsutakawa RK. Estimation in covariance components models. *Journal of the American Statistical Association* 1981; **76**(374):341–353.
15. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 1997; **53**(3):983–997.
16. Satterthwaite F. Synthesis of variance. *Psychometrika* 1941; **6**(5):309–316.
17. McCulloch CE, Searle SR, Neuhaus JM. *Generalized, Linear, and Mixed Models* (2nd edn). Wiley-Interscience, 2008.
18. Kerry SM, Bland JM. Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. *Statistics in Medicine* 2001-02; **20**(3):377–390.
19. SAS Institute Inc. *SAS® 9.3 Language Reference: Concepts* Second Edition. SAS Institute Inc.: Cary, NC, 2010.
20. Murray DM, Pals SL, Blitstein JL, Alfano CM, Lehman J. Design and analysis of group-randomized trials in cancer: a review of current practices. *Journal of the National Cancer Institute* 2008; **100**(7):483–491.
21. Smirnov N. Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics* 1948; **19**(2):279–281.
22. Gleser LJ. Exact power of goodness-of-fit tests of Kolmogorov type for discontinuous distributions. *Journal of the American Statistical Association* 1985; **80**(392):954–958.
23. Gail MH, Mark SD, Carroll RJ, Green SB, Pee D. On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine* 1996; **15**(11):1069–1092.
24. Murray DM, Hannan PJ, Baker WL. A Monte Carlo study of alternative responses to intraclass correlation in community trials. Is it ever possible to avoid Cornfield's penalties? *Evaluation Review* 1996-06; **20**(3):313–337.
25. Maxwell SE, Delaney HD. *Designing Experiments and Analyzing Data: A Model Comparison*. Routledge: Mahwah, NJ, 2003.
26. Kutner MH, Nachtsheim CJ, Neter John, Li William. *Applied Linear Statistical Models* (5th edn). McGraw-Hill/Irwin: Boston, 2004.