RESEARCH

Type I Error Control for Cluster Randomized Trials Under Varying Small Sample Structures

Joshua R Nugent and Ken P Kleinman*

Abstract

Background: Linear mixed models (LMM) are a common approach to analyzing data from cluster randomized trials (CRTs). Inference on parameters can be performed via Wald tests or likelihood ratio tests (LRT), but both approaches may give incorrect Type I error rates in common finite sample settings. The impact of different combinations of cluster size, number of clusters, intraclass correlation coefficient (ICC), and analysis approach on Type I error rates has not been well studied. Reviews of published CRTs find that small sample sizes are not uncommon, so the performance of different inferential approaches in these settings can guide data analysts to the best choices.

Methods: Using a random-intercept LMM stucture, we use simulations to study Type I error rates with the LRT and Wald test with different degrees of freedom (DF) choices across different combinations of cluster size, number of clusters, and ICC.

Results: Our simulations show that the LRT can be anti-conservative when the ICC is large and the number of clusters is small, with the effect most pronouced when the cluster size is relatively large. Wald tests with the between-within DF method or the Satterthwaite DF approximation maintain Type I error control at the stated level, though they are conservative when the number of clusters, the cluster size, and the ICC are small.

Conclusions: Depending on the structure of the CRT, analysts should choose a hypothesis testing approach that will maintain the appropriate Type I error rate for their data. Wald tests with the Satterthwaite DF approximation work well in many circumstances, but in other cases the LRT may have Type I error rates closer to the nominal level.

Keywords: Linear mixed models; Wald test; Likelihood ratio test; Type I error

Background

In cluster-randomized trials (CRTs), also called group randomized trials, subjects are organized in groups. These groups, rather than the subjects directly, are randomized to the trial interventions [1]. In these studies, outcomes within a cluster – for example, patients within hospitals or students within classrooms – are almost certainly correlated with one another. This clustering complicates data analysis because the common regression assumption that observations are independent is violated. When the response variable of interest is continuous, linear mixed models (LMMs), which require that observations are independent only after conditioning on cluster membership, are a common approach to the data analysis. CRTs are a widely used experimental design (see for example [2**?** , 3]), and LMMs are an attractive option for data analysis. Some reasons for this attractiveness are that LMMs are robust to certain missing data mechanisms and can flexibly accommodate nested levels of clustering and/or varying cluster sizes [4]. Generalized linear mixed models (GLMMs) extend the approach to non-Gaussian data, such as binary, count, or multinomial outcomes.

When fitting LMMs to CRT data, inference on parameters depends on asymptotic results, and in settings where the number of clusters is small they can generate Type I error rates well above or below the nominal level [5]. All frequentist null hypothesis testing theory depends on tests having the nominal size – a test with a nominal 5% error rate should produce false rejections 5% of the time. If not, data analysts in a CRT could be led to inappropriate conclusions, for example, producing too many false positives or false negatives when evaluating a treatment effect.

*Correspondence: ken.kleinman@gmail.com
Department of Biostatistics and Epidemiology, School of Public Health and Health Sciences, University of Massachusetts Amherst, 715 North Pleasant Street, 01003 Amherst, Massachusetts USA
Full list of author information is available at the end of the article

Unfortunately, small cluster counts are not uncommon in the literature, because it is often more expensive to add more clusters to a study than more individuals to a cluster. Despite common heuristics such as 'at least 30 units at each level of analysis' [6], CRTs often have as few as 20 clusters. Another review of 100 CRTs [7] found 37% with fewer than 20 clusters and minimal reporting of any small-sample corrections employed.

Some limited investigations of the problems with (G)LMM small sample inference have been conducted. Pinheiro and Bates [5] examined a very restricted parameter space, while Schluchter and Elashoff [8] reviewed the issue from a slightly different angle, examining approaches for longitudinal data with different covariance structures, which have different interpretations than a typical CRT. Several studies [9–12] suggested improving small-sample inference by applying the Bartlett correction [13], also under a smaller set of parameters than we apply here. However, as far as we are aware there is no simple way for data analysts to implement the Bartlett correction in SAS or R.

Other studies [14–16] examine issues around small numbers of clusters, but include both random intercepts and slopes, which may not be a structure that all CRTs utilize. Closer to our setting in this article, Leyrat et al. [17] evaluated the power and Type I error rates of different degrees of freedom (DF) choices for LMMs with Wald hypothesis tests for CRT designs under various design factors. They found both conservative and anti-conservative results, depending on the DF method chosen. Kahan et al. [7] reviewed small sample issues, but limited investigation to a small set of parameters and methods. Johnson et al. [18] examined LMM Type I error rates, but only for Wald tests with two DF choices, and did not break down their results by design factors. In the GLMM context, for binary outcomes only, Li and Redden [19] examined Type I error rates under different DF choices and found that the rates varied widely by method and design factors.

The work discussed above either does not break down the small-sample problems by design factor combinations (the effect of the ICC may vary depending on the number of clusters and cluster size, for example), does not compare results to the likelihood ratio test, and/or examines a limited set of data-generating parameters. Our work aims to add to this literature by examining in more detail the Type I error control of several LMM inference approaches in a variety of plausible CRT scenarios. We examine both likelihood ratio test and Wald test results, including different DF choices for the latter. We also vary cluster size, number of clusters, and intracluster correlation coefficient, looking at how results vary under the different approaches. We hope to provide enough detail to alert data analysts to the situations that may lead to incorrect Type I error rates with LMMs, and give guidance on which methods have the best error control given those factors.

Methods

We performed a Monte Carlo simulation study to examine the Type I error control of different LMM inference approaches under varying, plausible CRT circumstances. First, we describe the statistical model in question and the difficulties with small-sample inference, then we outline our specific study design. For all data analysis in this article, we used the SAS/STAT 15.1 (SAS Institute Inc., Cary, NC) and R 3.6.0 (R Foundation for Statistical Computing) software packages.

Model

We consider here a version of the linear mixed-effects model of Laird and Ware [20]:

$$Y_{ij} = \mathbf{X}_{ij}^T\boldsymbol{\beta} + \mathbf{Z}_{ij}^T\boldsymbol{b}_i + \epsilon_{ij} \quad (1)$$

where Y_{ij} is a continuous response variable for individual j in cluster i, X_{ij}^T are that individual's covariates for a vector of fixed effect regression parameters $\boldsymbol{\beta}$, Z_{ij}^T are the cluster-level values for a vector of random effects $\boldsymbol{b}_i$ for cluster i, and ϵ_{ij} is the residual error of the observation. In our case, matching common practice in CRTs, we restricted the random-effects structure to include only a random intercept term, so the term $Z_{ij}^T\boldsymbol{b}_i$ reduces to b_{0i}. We let $\epsilon_{ij} \sim N(0, \sigma^2)$ for all individuals, and cluster-level variance b_{0i} was distributed $N(0, \sigma_b^2)$, with b_{0i} independent of ϵ_{0i}. We further assumed that cluster size is uniform for all clusters, and that there are two treatment arms with an equal number of clusters in each arm, modeled with an indicator variable $x_i \in \{0, 1\}$ for control or treatment arm, with β_1 being the treatment effect. Thus, for the remainder of the article, our model is:

$$Y_{ij} = \beta_0 + \beta_1 x_i + b_{0i} + \epsilon_{ij} \quad (2)$$

Impact of clustering on inference

In a CRT, there are typically two assumed sources of variability in outcomes: between-cluster, denoted here as σ_b^2, and within-cluster, denoted as σ^2. The marginal variance of $y_{ij} = \sigma_b^2 + \sigma^2$. One way of quantifying the amount of clustering is via the *intracluster correlation*

coefficient (ICC) ρ, defined as $\frac{\sigma_b^2}{\sigma_b^2+\sigma^2}$, or the proportion of total variance due to the cluster-level variability. If one were to incorrectly analyze the data using a linear model rather than a linear mixed model, standard errors for the coefficient estimates would have to be adjusted, since observations are correlated in violation of the model assumptions. An approximation of this adjustment, the *design effect* [21], is a multiplier for the sampling variance of the treatment effect estimator. It is defined as $[(n-1)\rho+1]$, where n is the number of subjects per cluster. For example, with 10 observations per cluster and an ICC of .01, the design effect is 1.09, meaning that the treatment effect coefficient standard errors would have to be multiplied by roughly $\sqrt{1.09} \approx 1.04$ to account for clustering. However, with 100 observations per cluster and the same ICC, the standard error multiplier increases to $\sqrt{2} \approx 1.41$, and for 1000 observations per cluster it increases to $\sqrt{11} \approx 3.31$, meaning that even a very small ICC can drastically change inferences when the cluster size is large. This approximation demonstrates the necessity of accounting for between-cluster variation in the data analysis, even if the ICC is expected to be small.

Inference with LMM fixed effect estimators

Two ways of fitting a linear mixed model are by maximum likelihood (ML) and restricted maximum likelihood (REML), and most major statistical software packages can perform estimation by either method. Inference about $\hat{\beta}_1$ can be made using the likelihood ratio test (LRT) if fitting via ML, or by a Wald test if fitting via REML. A third test based on the maximum likelihood, the score test, is rarely used in this setting and is not discussed here. The LRT compares the log-likelihood of a model without β_1 (ℓ_0) to a model that includes it (ℓ_1), and the test statistic $\lambda = -2(\ell_0 - \ell_1)$ has a χ^2_p distribution, asymptotically, with degrees of freedom p the difference in parameter dimension between the two models. In our case, as in many CRTs, there is one treatment effect parameter, so $p = 1$. In general, the LRT is recommended over the Wald test, as its asymptotic properties are superior [22]. Unfortunately, the χ^2 distribution may be a poor approximation of the distribution of λ when the amount of information in a sample, for example, cluster count, is small.

Alternatively, a Wald test statistic under the null hypothesis $H_0 : \beta_1 = 0$ can be generated by dividing the estimated treatment effect by its standard error: $t^* = \hat{\beta}_1 / SE(\hat{\beta}_1)$. This value can then be compared to a central t distribution. Unfortunately, for many designs, it is unclear what the appropriate degrees of freedom (DF) for that distribution should be [23]. Choices include:

- Residual: $N - p$, where N is the total number of observations and p is the number of fixed-effects coefficients to be estimated in the model. In the CRT design assumed here, $p = 2$. Since the number of observations is usually much larger than the number of parameters in the model, this will generate similar results to the 't as z' approach described below.
- Between-within: The residual DF are partitioned into between-subject and within-subject groups, equivalent in this case to a one-way ANOVA decomposition, meaning $DF = K - 2$, where K is the number of clusters.
- Satterthwaite approximation: This method, generalizing the ideas of Satterthwaite [24], is quite complex, but it essentially uses the variance of the β_1 estimate in its calculation of the DF. For more detail, see McCulloch et al. [25], Ch. 6.
- Kenward-Roger approximation: This method [26] inflates the fixed and random effects variance-covariance matrix, and calculates Satterthwaite DF based on these inflated values. Under our model with one treatment effect, it generates DF equivalent to the Satterthwaite approximation.
- Infinite ('t as z'): The statistic is compared to a standard normal distribution, equivalent to a t distribution with infinite DF.

Alternative inferential approaches

The Wald and likelihood ratio tests are not the only options for generating confidence intervals and performing inference in CRTs. Bayesian methods have been implemented with mixed models [27, 28], though under the study designs considered here, these reports showed no major improvements over frequentist approaches in small-sample settings, so we chose not to include Bayesian methods in this analysis. Alternatively, confidence intervals for LMM fixed effects can be generated by a parametric, semi-parametric, or non-parametric bootstrap. All are computationally intensive and require careful implementation due to the clustered nature of the original sample, so we chose not to investigate those approaches, though the parametric boostrap has been recommended by some authors [29].

Data generation

We generated clustered, balanced data sets from the null model

$$y_{ij} = b_{0i} + \epsilon_{ij} \tag{3}$$

for clusters $i = 1, 2, ..., K$ and individuals $j = 1, 2, ..., N$ within each cluster. The random intercept

b_{0i} for cluster i was distributed $\sim N(0, \sigma_b^2)$, and the residual error term $\epsilon_{ij} \sim N(0, \sigma^2)$. b_{0i} and ϵ_{ij} were generated as independent pseudorandom variates. We also generated values of x_{ij} such that for clusters $i = 1, ...K/2$, $x_{ij} = 0$, and for $i = K/2 + 1, ...K$, $x_{ij} = 1$. This variable represents the treatment indicator, though it was not used in the data generation, as there is no treatment effect under the null hypothesis.

For each data set, we then fit the model shown in equation (2) using SAS PROC MIXED and the **lme4** and **lmerTest** packages in R. The coefficient of interest in these fitted models, $\hat{\beta}_1$, represents the estimated treatment effect.

We gathered p-values for the $\hat{\beta}_1$ coefficients using the LRT and the Wald test using the various DF options. We assessed the rejection rate under each test for the null hypothesis that $\beta_1 = 0$ with $\alpha = .05$. Since the data-generating mechanism had a true β_1 value of zero, this estimates the TIE rate for the nominal $\alpha = .05$ level.

We performed our analysis on 10,000 simulated data sets for all possible combinations of the following data-generating parameters:

- total number of clusters $K \in \{10, 20, 40, 100\}$, divided evenly among the two treatment arms
- subjects per cluster $N \in \{3, 10, 20, 50\}$
- $\sigma_b^2 \in \{0.001, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$
- $\sigma^2 = 1$

The Wald test statistic is scaled by the standard error of $\hat{\beta}_1$. Tthat standard error is proportional to the square root of the total outcome variance through the ICC. Therefore, different magnitudes of σ_b^2 and σ^2 that gave the same ICC will produce the same test statistics. We tested different magnitudes for σ_b^2 and σ^2 that produced the same ICC and confirmed this. This allowed us to simplify our analysis by fixing σ^2 at 1 and only varying σ_b^2.

Determining p-values

Both PROC MIXED and **lme4** report $\hat{\beta}_1$ estimates, their associated standard errors, and t^* statistics. This allows for easy testing of the $\hat{\beta}_1$ coefficient via a Wald test, fitting with REML. The t^* statistics generated were compared to t distributions with three choices of DF: between-within, Satterthwaite/Kenward-Roger, and residual, as described earlier. We then collected the p-values and calculated TIE rates under the three DF choices.

Both software packages also allow for model fitting using ML, allowing for model comparison and p-value determination for $\hat{\beta}_1$ via the LRT. First, a null model (4) was fit, with the only fixed effect being an intercept term:

$$y_{ij} = \beta_0 + b_{0i} + \epsilon_{ij} \quad (4)$$

Second, a model with an added fixed effect for x_{ij}, as in model (2). The doubled difference in maximized log-likelihood was compared to a χ_1^2 distribution since there was a one-parameter difference in model dimension. P-values from the χ_1^2 distribution were collected and TIE rates calculated.

Results

Both software packages generated identical $\hat{\beta}_1$ estimates and standard errors when fitting with REML, and identical differences in likelihoods when fitting with ML. Reported results are from SAS. In addition, since the Kenward-Roger and Satterthwaite approximations were indistinguishable in this setting, they are both labeled as "approximate."

Results are displayed in Figure 1. Under all approaches, departures from the nominal α level were most pronounced when the number of clusters is small.

When the number of observations per cluster is small, and there is a relatively small ICC, the LRT demonstrated appropriate TIE control. Regardless of the number of observations per cluster, the LRT is anti-conservative as the ICC rises. However, the anti-conservatism of the LRT was most apparent with smaller ICC when the number of observations per cluster was larger. Even with as many as 40 clusters and 50 observations per cluster, the LRT was noticeably anti-conservative once the ICC rose above .1. Worse, even when the ICC was very small (.01, .02), the LRT was anti-conservative with as few as 20 clusters of 50 observations per cluster.

As for the Wald tests, the between-within DF option led to conservative TIE rates when the ICC was small and/or the cluster size was small, but maintained the appropriate TIE rate with large clusters or a large ICC. The residual DF choice was less conservative in the case of a small ICC, but produced anti-conservative results as the ICC increased, and was more anti-conservative when the cluster size was large. Notably, depending on how the model is fit, the default method for determining DF in SAS may be 'containment', which under this study design leads to SAS assigning residual DF. Since this choice leads to the most anti-conservative results, it may be a concern for SAS analysts. The Satterthwaite approximation for our simulation estimated the DF as equal to the between-within DF in some cases and to residual DF in other cases, depending on the data set. This is why the TIE rates labeled "approximate" in Figure 1 are bounded by those other two options.

We also tested the the effect of an ICC of .09 generated with $\sigma_b^2 = 1$ and $\sigma^2 = 10$ rather than the values discussed above. The results did not differ notably, which suggests that this pattern of TIE rate inflation with the LRT, as with the Wald test, is insensitive to the absolute size of the σ_b^2 and σ^2 values, only their relative size.

Finally, given the balanced nature of our data and the lack of other covariates, we could have used a t-test on the cluster means of each treatment arm to perform a hypothesis test. Using this approach, we achieved close to the nominal .05 alpha level in all cases. However, since most CRTs include covariates, a t-test would be inappropriate, and hence these results are omitted from the plot. The Wald test with the between-within DF choice is almost equivalent to this t-test [30], the only difference being that the LMM estimates two variances ($\hat{\sigma}_b^2$ and $\hat{\sigma}^2$), while the t-test only estimates their sum, leading to slighly different inferences.

Discussion

To our knowledge, the effect of different combinations of design factors and analysis approach on TIE rates have not been examined comprehensively in previous reports. Our results show that none of the approaches meet the nominal alpha level in all cases examined, and the departures from the nominal level are directionally different based on the approach and data structure. Hence, there is no one-size-fits all recommendation for data analysts in these small-sample cases.

The likelihood ratio test, based on an asymptotic χ^2 distribution, does not perform well in these finite-sample cases, especially when the clusters contain many observations. This extends other studies that found the LRT to be anti-conservative [5, 31] in smaller explorations of the possible parameter combinations.

Alternatively, with a Wald test, some choices of DF, such as between-within or the data-adaptive Satterthwaite, can avoid anti-conservatism. However, a tradeoff exists, as they are too conservative when the ICC, the number of clusters, and/or cluster size is small.

We tested the interactions between our design factors, using a three-way ANOVA within each analysis type with the TIE rate as the outcome, breaking the 10,000 simulations of each condition into 10 sets of 1,000. Most of these three-way interactions were statistically significant, and given the strong patterns see in Figure 1, we expect that we could show significance of all the interactions if we grew the number of simulations arbitrarily.

The results here suggest that data analysts should choose an approach that best suits their data. For example, if the ICC is expected to be small and the number of observations per cluster is small, the likelihood ratio test should perform well. For cases where the number of observations per cluster is large, a Wald test with the Satterthwaite DF approximation is better, though it can be conservative in some situations.

One perhaps unsatisfying conclusion is that analysts may want to generate their own small simulation studies to evaluate different approaches before fitting their final data models, since they will likely know the model structure, number of clusters, and cluster size by that point.

Finally, we caution analysts to be careful when using defaults settings in software. For example, with Wald tests, SAS PROC MIXED may default to the poorly-performing residual DF choice, and the **lmerTest** package in R defaults to the Satterthwaite approximation, which may be too conservative in some cases.

It is unclear how aware data analysts may be about the small-sample problems that may arise in making inference from mixed models. A review of LMM applications in education and social sciences [32] found minimal reporting of estimation and inference methods and assumptions, and that cluster sizes could be as low as 2 and the number of clusters as low as 8. Our own review, and that of Kahan et al. [7], confirmed that small cluster counts are not unusual in biomedical settings as well. Therefore, we hope this will provide analysts with some recommendations of which approaches control Type I error at appropriate rates under different circumstances, and we encourage more reporting of DF choices and analytic methods in CRT publications.

Given that small sample sizes are not uncommon in CRT literature, there is need for more investigation of which methods control Type I error in other contexts. One limitation of our result is that we did not include any scenarios with repeated measures (for example, baseline, post-treatment, and follow-up), which are common in biomedical settings, and deserve similar scrutiny. Additionally, more parameters could have been added to the simulations, such as unbalanced cluster sizes or varying ICC by treatment arm. Another potential avenue for exploration, following on the work of Li and Redden [19], would be to examine TIE rates for Poisson outcomes under these study conditions and add comparisons to the LRT under both binary and count outcomes. Type II errors may also be a concern for researchers, and investigating the role of different analytic methods on these could be an area for future work. Finally, the impact of these data/approach effects on statistical power should be determined so that analysts can make appropriate sample size calculations during the design phase of a CRT.

Conclusions

Competing interests
The authors declare that they have no competing interests.

Author's contributions
JN performed the simulations and drafted the text. KK supervised the research and edited the manuscript.

Acknowledgements
Support for this work for provided by NIH/NIGMS grant R01GM121370.

References

1. Hayes, R.J., Moulton, L.H.: Cluster Randomised Trials, 2 edition edn. Chapman and Hall/CRC, Boca Raton (2017)
2. Moon, R.Y., Hauck, F.R., Colson, E.R., Kellams, A.L., Geller, N.L., Heeren, T., Kerr, S.M., Drake, E.E., Tanabe, K., McClain, M., Corwin, M.J.: The Effect of Nursing Quality Improvement and Mobile Health Interventions on Infant Sleep Practices: A Randomized Clinical Trial. JAMA **318**(4), 351–359 (2017). doi:10.1001/jama.2017.8982
3. Huang, S.S., Septimus, E., Kleinman, K., Moody, J., Hickok, J., Avery, T.R., Lankiewicz, J., Gombosev, A., Terpstra, L., Hartford, F., Hayden, M.K., Jernigan, J.A., Weinstein, R.A., Fraser, V.J., Haffenreffer, K., Cui, E., Kaganov, R.E., Lolans, K., Perlin, J.B., Platt, R.: Targeted versus Universal Decolonization to Prevent ICU Infection. New England Journal of Medicine **368**(24), 2255–2265 (2013). doi:10.1056/NEJMoa1207290
4. Fitzmaurice, G.M., Laird, N.M., Ware, J.H.: Applied Longitudinal Analysis, 2 edition edn. Wiley, ??? (2012)
5. Pinheiro, J., Bates, D.: Mixed-Effects Models in S And S-PLUS. Springer, New York (2009)
6. Kreft, I.G.G.: Introducing Multilevel Modeling, 1 edition edn. SAGE Publications Ltd, London ; Thousand Oaks, Calif (1998)
7. Kahan, B.C., Forbes, G., Ali, Y., Jairath, V., Bremner, S., Harhay, M.O., Hooper, R., Wright, N., Eldridge, S.M., Leyrat, C.: Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters: a review, reanalysis, and simulation study. Trials **17**(1) (2016). doi:10.1186/s13063-016-1571-2. Accessed 2019-10-06
8. Schluchter, M.D., Elashoff, J.T.: Small-sample adjustments to tests with unbalanced repeated measures assuming several covariance structures. Journal of Statistical Computation and Simulation **37**(1-2), 69–87 (1990). doi:10.1080/00949659008811295
9. Zucker, D.M., Lieberman, O., Manor, O.: Improved small sample inference in the mixed linear model: Bartlett correction and adjusted likelihood. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **62**(4), 827–838 (2000). doi:10.1111/1467-9868.00267
10. Melo, T.F.N., Ferrari, S.L.P., Cribari-Neto, F.: Improved testing inference in mixed linear models. Computational Statistics & Data Analysis **53**(7), 2573–2582 (2009). doi:10.1016/j.csda.2008.12.007
11. Manor, O., Zucker, D.M.: Small sample inference for the fixed effects in the mixed linear model. Computational Statistics & Data Analysis **46**(4), 801–817 (2004). doi:10.1016/j.csda.2003.10.005
12. Stein, M.C., da Silva, M.F., Duczmal, L.H.: Alternatives to the usual likelihood ratio test in mixed linear models. Computational Statistics & Data Analysis **69**, 184–197 (2014). doi:10.1016/j.csda.2013.08.002
13. Bartlett, M.S., Fowler, R.H.: Properties of sufficiency and statistical tests. Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences **160**(901), 268–282 (1937). doi:10.1098/rspa.1937.0109
14. Luke, S.G.: Evaluating significance in linear mixed-effects models in R. Behavior Research Methods **49**(4), 1494–1502 (2017). doi:10.3758/s13428-016-0809-y
15. Maas, C.J.M., Hox, J.J.: Sufficient Sample Sizes for Multilevel Modeling. Methodology **1**(3), 86–92 (2005). doi:10.1027/1614-2241.1.3.86
16. Bell, B., Morgan, G., Schoeneberger, J., Loudermilk, L., Kromrey, J., Ferron, J.: Dancing the Sample Size Limbo with Mixed Models: How Low Can You Go? SAS Global Forum **4** (2010)
17. Leyrat, C., Morgan, K.E., Leurent, B., Kahan, B.C.: Cluster randomized trials with a small number of clusters: which analyses should be used? International Journal of Epidemiology **47**(1), 321–331 (2018). doi:10.1093/ije/dyx169
18. Johnson, J.L., Kreidler, S.M., Catellier, D.J., Murray, D.M., Muller, K.E., Glueck, D.H.: Recommendations for choosing an analysis method that controls Type I error for unbalanced cluster sample designs with Gaussian outcomes. Statistics in Medicine **34**(27), 3531–3545 (2015). doi:10.1002/sim.6565
19. Li, P., Redden, D.T.: Comparing denominator degrees of freedom approximations for the generalized linear mixed model in analyzing binary outcome in small sample cluster-randomized trials. BMC Medical Research Methodology **15** (2015). doi:10.1186/s12874-015-0026-x
20. Laird, N.M., Ware, J.H.: Random-effects models for longitudinal data. Biometrics **38**(4), 963–974 (1982)
21. Kish, L.: Survey Sampling. Wiley-Interscience, New York (1965)
22. Cox, D., Hinkley, D.: Theoretical Statistics. Chapman & Hall/CRC, Boca Raton (1979)
23. Bates, D., Mächler, M., Bolker, B., Walker, S.: Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software **67**(1), 1–48 (2015). doi:10.18637/jss.v067.i01
24. Satterthwaite, F.E.: An Approximate Distribution of Estimates of Variance Components. Biometrics Bulletin **2**(6), 110–114 (1946). doi:10.2307/3002019
25. McCulloch, C.E., Searle, S.R., Neuhaus, J.M.: Generalized, Linear, and Mixed Models, 2 edition edn. Wiley-Interscience, Hoboken, N.J (2008)
26. Kenward, M.G., Roger, J.H.: Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. Biometrics **53**(3), 983–997 (1997). doi:10.2307/2533558
27. Browne, W.J., Draper, D.: A comparison of Bayesian and likelihood-based methods for fitting multilevel models. Bayesian Analysis **1**(3), 473–514 (2006). doi:10.1214/06-BA117
28. Baldwin, S.A., Fellingham, G.W.: Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. Psychological Methods **18**(2), 151–164 (2013). doi:10.1037/a0030642
29. Ukyo, Y., Noma, H., Maruo, K., Gosho, M.: Improved Small Sample Inference Methods for a Mixed-Effects Model for Repeated Measures Approach in Incomplete Longitudinal Data Analysis. Stats **2**(2), 174–188 (2019). doi:10.3390/stats2020013
30. Moerbeek, M., van Breukelen, G.J.P., Berger, M.P.F.: A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies. Journal of Clinical Epidemiology **56**(4), 341–350 (2003). doi:10.1016/S0895-4356(03)00007-6
31. Halekoh, U., Hojsgaard, S.: A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models: The R Package pbkrtest. Journal of Statistical Software **59**(1), 1–32 (2014). doi:10.18637/jss.v059.i09
32. Dedrick, R.F., Ferron, J.M., Hess, M.R., Hogarty, K.Y., Kromrey, J.D., Lang, T.R., Niles, J.D., Lee, R.S.: Multilevel modeling: A review of methodological issues and applications. Review of Educational Research **79**(1), 69–102 (2009). doi:10.3102/0034654308325581

Figures

Figure 1 Relationship between Type I error rate and design factors.