COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2021.1982976

Use of between-within degrees of freedom as an alternative to the Kenward–Roger method for small-sample inference in generalized linear mixed modeling of clustered count data

Vincent S. Staggs[a,b] and Keith Feldman[b,c]

[a]Biostatistics & Epidemiology Core, Children's Mercy Kansas City, Kansas City, MO, USA; [b]Department of Pediatrics, University of Missouri-Kansas City School of Medicine, Kansas City, MO, USA; [c]Health Services & Outcomes Research, Children's Mercy Kansas City, Kansas City, MO, USA

ABSTRACT

Clustered count data, common in health-related research, are routinely analyzed using generalized linear mixed models. There are two well-known challenges in small-sample inference in mixed modeling: bias in the naïve standard error approximation for the empirical best linear unbiased estimator, and lack of clearly defined denominator degrees of freedom. The Kenward–Roger method was designed to address these issues in linear mixed modeling, but neither it nor the simpler option of using between-within denominator degrees of freedom has been thoroughly examined in generalized linear mixed modeling. We compared the Kenward–Roger and between-within methods in two simulation studies. For simulated cluster-randomized trial data, coverage rates for both methods were generally close to the nominal 95% level and never outside 93-97%, even for 5 clusters with an average of 3 observations each. For autocorrelated longitudinal data, between-within intervals were more accurate overall, and under some conditions both the original and improved Kenward–Roger methods behaved erratically. Overall, coverage for Kenward–Roger and between-within intervals was generally adequate, if often conservative. Based on the scenarios examined here, use of between-within degrees of freedom may be a suitable or even preferable alternative to the Kenward–Roger method in some analyses of clustered count data with simple covariance structures.

ARTICLE HISTORY
Received 6 July 2021
Accepted 14 September 2021

KEYWORDS
Generalized linear mixed model; Mixed model; Multilevel model; Hierarchical model; Clustered data

1. Introduction

Advances in collection and digitalization of medical and health-related data have expanded the sources, volume, and types of information available for use in research. Count data are common in health care research and are often clustered organizationally (e.g., participants within sites of cluster-randomized trials), temporally (e.g., repeated measurements for participants in longitudinal studies), or both. Due to practical limitations, the number and size of clusters in these studies are not always large. To carry out appropriate inferential analyses of such data, we must consider both the clustering, which may violate the usual modeling assumption of independent observations, and the non-Gaussian distribution of the counts, which tend to yield non-Gaussian model residuals.

CONTACT Vincent S. Staggs vstaggs@cmh.edu Biostatistics & Epidemiology Core, Children's Mercy Kansas City, 2400 Gillham Road, Kansas City, MO 64108, USA.

Generalized linear mixed models (GLMMs) have become popular for modeling clustered data that violate standard linear model assumptions of independent, Gaussian errors. In mixed modeling, however, there are two well-known challenges in inference for fixed effects when samples are not large. The first is approximating the standard error of the empirical best linear unbiased estimator (EBLUE), which cannot generally be written in closed form. The standard naïve approximation of this standard error is computed by substituting estimates of the variance parameters for their true values. Because sampling variability in the variance parameter estimates is ignored, however, the naïve approximation is biased downward for intrinsically linear covariance structures, resulting in anticonservative confidence intervals and p-values (Kenward and Roger 2009). In linear mixed models, the error of the naïve approximation does converge to zero as the number of clusters goes to infinity (Demidenko 2004), but as with other asymptotic methods, it can be difficult to determine how large a sample is needed to rely on asymptotic properties. Complicating this issue, the effects of increasing sample size at Level 1 (number of observations per cluster) and at Level 2 (number of clusters) are not the same.

The second challenge is choosing denominator degrees of freedom for F tests of fixed effects and, equivalently, the appropriate degrees of freedom for the t distribution cutoff used in constructing Wald-type confidence intervals (Littell 2002). The residual degrees of freedom used for inferences in ordinary least squares regression ($N - k - 1$ for total sample size N and k explanatory variables) are anticonservative when used for inference for fixed effects in mixed models. Asymptotically, of course, the t and F distributions converge to Gaussian and χ^2 distribution, respectively, and approximating degrees of freedom for mixed models becomes unnecessary with a sufficiently large number of clusters. But with sample sizes encountered in practice, the choice of degrees of freedom can have meaningful effects on inference, biasing both confidence limits and p-values.

1.1. Kenward–Roger method

To address these challenges, Kenward and Roger proposed a method that combines a standard error approximation based on a Taylor series expansion with a Satterthwaite-type degrees of freedom approximation (Kenward and Roger 1997). Their method has been shown to perform well in linear mixed model contexts, where Level 1 errors are assumed to be Gaussian (McNeish and Stapleton 2016; Luke 2017; Staggs 2017; Francq, Lin, and Hoyer 2019). Kenward and Roger later developed an improved version of their method to account for bias in covariance parameter estimates, which can negatively affect performance of their original method for models with non-linear covariance structures (Kenward and Roger 2009). Results of the original and improved Kenward–Roger methods differ only for models with non-linear covariance structures. Following SAS, we denote these methods KR and KR2, respectively, when they need to be differentiated. Neither has been thoroughly studied for analysis of clustered count data using GLMMs.

In early work designed to compare methods of analysis for clustered count data, Bell and Grunwald simulated longitudinal count data under various covariance structures and compared Type I error rates for (1) unadjusted tests from Poisson GLMMs fit with maximum likelihood, and (2) Kenward–Roger-based tests from Poisson GLMMs fit with residual pseudo-likelihood (akin to residual or restricted maximum likelihood) (Bell and Grunwald 2011). Although cluster sizes as small as 3 were considered, the smallest number of clusters was 10. When the covariance structure specified for the GLMM was correct, Type I error rates for the two test types were very similar and generally conservative (0.02–0.05). When the GLMM covariance structure was *incorrect*, error rates remained conservative except for autocorrelated data with clusters of size 10. Because the unadjusted tests were based on (unrestricted) maximum likelihood estimates, which are only unbiased asymptotically, and the KR tests were based on residual (restricted) pseudo-likelihood estimates, differences between the unadjusted and adjusted test types were conflated with differences between the unrestricted and restricted likelihood methods, making it difficult to separate effects of test type and estimation method.

Stroup reported good performance for the Kenward–Roger method in simulation studies where GLMMs were fit to clustered count data under a limited range of conditions (Stroup 2018; Stroup 2015). Specifically, Stroup simulated clustered negative binomial data with skewed, Gamma-distributed, random cluster intercepts. Because random cluster effects in GLMMs are assumed to be Gaussian, the simulated data allowed Stroup to assess the method's robustness against violation of this assumption, but not the method's performance with data generated from the correct model. In addition, compound symmetry was the only covariance structure Stroup examined. Durán Pacheco et al. also simulated and analyzed clustered negative binomial data but did not compare options for estimating degrees of freedom or consider fewer than 10 clusters, average cluster sizes smaller than 30, or covariance structures other than compound symmetry (Durán et al. 2009). In more recent work, the Kenward–Roger method performed very well in Poisson mixed modeling of count data with as few as 10 clusters having compound symmetric covariance structure; cluster sizes smaller than 12 and nonlinear covariance structures were not examined (McNeish 2019).

Most recently, Jackson et al. simulated count data for different cluster-randomized trial scenarios, varying the number of clusters from 6 to 18 and cluster sizes from 37 to 127 (Jackson et al. 2021). Various approaches to analyzing the data were compared, including Poisson mixed modeling with pseudo-likelihood estimation, both with and without application of the Kenward-Roger adjustment. Not surprisingly, unadjusted tests were prone to inflated Type I error rates, whereas the Kenward-Roger tests were not, though under one study design Kenward–Roger Type I error rates were excessively conservative (0.92% and 1.86%). Non-linear covariance structures were beyond the scope of the study.

1.2. Between-within degrees of freedom

An alternative to approximating both the EBLUE standard error and denominator degrees of freedom with the Kenward–Roger method is to use the unadjusted naïve EBLUE standard error approximation with the between-within method for computing degrees of freedom (Schluchter and Elashoff 1990). Under the between-within method, degrees of freedom are allocated based on whether an explanatory variable varies only *between* clusters (a Level 2 variable) or varies across observations *within* clusters (a Level 1 variable). Assuming a two-level model with a fixed intercept, the between-within denominator degrees of freedom for cluster-level effects is defined as *number of clusters – number of (fixed) Level 2 parameters—1*. For quantitative Level 1 variables, between-within denominator degrees of freedom are computed by subtracting the degrees of freedom for Level 2 effects from the residual degrees of freedom, which simplifies to *total sample size – number of clusters – number of (fixed) Level 1 parameters*. The formula is more complicated for Level 1 categorical variables with multiple levels, but in the simpler case here, the between- and within-cluster denominator degrees of freedom sum to the residual degrees of freedom. Whereas clustering is simply ignored with residual degrees of freedom, the between-within method takes the number of clusters and level of the explanatory variables into account, so we can expect its results to be less anticonservative than those based on residual degrees of freedom. Used in conjunction with the (biased) naïve EBLUE standard error approximation, however, the between-within method may be anticonservative relative to the Kenward–Roger method.

Like the Kenward–Roger method, use of between-within degrees of freedom in GLMMs is not well-studied. In a simulation study involving GLMM analysis of clustered binary data, Li and Redden found that the between-within method generally outperformed the Kenward–Roger method (Li and Redden 2015). For 10 or 20 clusters, the Kenward–Roger method yielded conservative Type I error rates in tests for a Level 2 fixed effect, a problem exacerbated by unbalanced cluster sizes. As in most of the studies described above, covariance structures other than compound symmetry were not examined.

The primary aim of our study was to compare the performance of the Kenward–Roger and between-within methods in inference for fixed effects in negative binomial mixed models under a range of small sample sizes. A secondary aim was to develop a better sense of the minimum sample size requirements for applying these methods in analyses of clustered count data. Both cluster- and observation-level fixed effects were examined. In keeping with the statistics community's shift away from binary hypothesis testing, we did not directly examine Type I error rates or statistical power. Instead we compared the coverage and average length of confidence intervals, which subsume hypothesis tests and reflect precision of estimation (and, by extension, test size and statistical power).

2. Methods

To address these research aims, we carried out two simulation studies. In the first we examined inferences for a Level 1 and a Level 2 effect in a scenario resembling a cluster-randomized trial, where compound symmetry is assumed. In the second we examined inferences in a scenario involving temporal clustering and a cross-level interaction between time (Level 1) and a dichotomous Level 2 variable under both correct (autoregressive) and incorrect (compound symmetry) covariance structures.

2.1. Simulation study 1

In the first simulation study, we generated count data to resemble outcomes from a cluster randomized trial, where each cluster (e.g., study site) is randomly assigned to one of two treatment arms, and individuals are nested within clusters. We considered an extensive combination of study parameters, including two ICCs (0.1, 0.4), a varying number of clusters (5, 10, 15, 20), and a range of average cluster sizes (3, 6, 12) for a total of 24 conditions. For average cluster size *n*, we simulated data for an equal number of clusters of size $n-2$, $n\ -\ 1$, n, $n+1$, and $n+2$ to yield the specified number of clusters (5, 10, 15, or 20). One thousand datasets were simulated for each of the 24 unique conditions.

Each dataset was generated by simulating a negative binomial draw for each hypothetical individual. Unlike the Poisson distribution, the negative binomial is not restricted to have variance equal to its mean and can therefore be used to generate, and model, data that are over-dispersed relative to the Poisson distribution. Letting $\mathrm{y_{ij}}$ denote the outcome for the j^{th} individual in the i^{th} cluster, $\mathrm{y_{ij}}$ was drawn from a negative binomial distribution with dispersion (scale) 5 and mean λ_{ij} computed as follows:

$$\log(\lambda_{ij}) = \mu_{ij} = \beta_0 + \beta_1 * Arm_i + \beta_2 * x_{ij} + u_i$$

where μ_{ij} denotes the linear predictor, Arm_i is an indicator for the ith cluster's randomly assigned study arm (0 or 1), individual-level covariate value x_{ij} was drawn from distribution $\mathrm{N}(0, \sigma_{\mathrm{x}}{}^2)$, random cluster intercept u_i was drawn from distribution $\mathrm{N}(0, \sigma_{\mathrm{u}}{}^2)$, and fixed parameters β_0, β_1, and β_2 were set to 1. For conditions with (log-scale) ICC $= 0.1$, $\sigma_{\mathrm{x}}{}^2$ was set to 0.65 and $\sigma_{\mathrm{u}}{}^2$ to 0.10; for ICC $= 0.4$, $\sigma_{\mathrm{x}}{}^2$ was set to 0.35 and $\sigma_{\mathrm{u}}{}^2$ to 0.40.

After simulating the data, we fit a negative binomial mixed model to each dataset. Consistent with the generating model, each model included study arm and the covariate (x) as explanatory variables, as well as a random cluster intercept. Each dataset was analyzed by fitting two variations of this GLMM using residual pseudo-likelihood as implemented in the SAS GLIMMIX Procedure. First, we fit the model to each dataset with the Kenward–Roger method specified to obtain 95% confidence limits for the two fixed effects of interest. Then we re-fit the model to each dataset but with between-within degrees of freedom specified instead of the Kenward–Roger method. The KR and KR2 versions of the Kenward–Roger method were not compared because the covariance structure (compound symmetry) is linear. Coverage rates for the 95% Kenward–Roger and between-within intervals were computed as the percentage of simulated datasets for which the confidence interval contained the true parameter value. In addition, we computed the average length of the intervals produced under each method for comparison.

2.2. Simulation Study 2

In the second simulation study, we simulated longitudinal count data of the kind we might encounter in a study where patients are randomly assigned to one of two treatment arms and repeatedly assessed over time to compare the average rate at which symptoms change for patients in the two arms. We considered a range of participant counts (5, 10, 15, 20) and a varying number of time points per participant (3, 6, 9) for a total of 12 conditions. Again, 1,000 datasets were simulated for each unique condition.

Following the method of Kalema and Molenberghs (2016), we simulated correlated, negative binomial draws for each hypothetical participant using the Poisson-Gamma conceptualization of the negative binomial distribution. As draws from a Poisson distribution with a Gamma-distributed rate parameter follow a negative binomial distribution, negative binomial draws for participants $i = 1, 2, \ldots, m$ with time points $j = 1, 2, \ldots, n$ can be simulated by drawing from a Poisson(λ_{ij}) distribution with $\lambda_{\mathrm{ij}} = \theta_{ij}e^{\mu_{ij}}$, where Θ_{ij} is a draw from a multivariate Gamma distribution, and μ_{ij} is the linear predictor. Correlations between draws for the i^{th} participant are induced by drawing $\Theta_{\mathrm{i1}}, \Theta_{\mathrm{i2}}, \ldots$ from an n-dimensional Gamma distribution with a non-diagonal variance-covariance matrix.

For each hypothetical participant, the value of the linear predictor was computed as follows:

$$\mu_{ij} = \beta_1 * Time_i + \beta_2 * Arm_i * Time_i + u_i$$

where $Time_j$ is the j^{th} time point $(0, 1, 2, \ldots)$, Arm_i is an indicator for the ith participant's study arm (0 or 1), random participant intercept $\mathrm{u_i}$ was drawn from distribution N(0, 0.01), and fixed parameters β_1 and β_2 were set to 0.10 and 0.05, respectively. Thus, on average, values for the two arms are equal at baseline, increase by 0.10 per time point for Arm $= 0$, and increase by 0.15 per time point for Arm $= 1$.

Using the SimCorrMix package in R (Fialkowski and Tiwari 2019), we simulated a vector of n draws $(\Theta_{\mathrm{i1}}, \Theta_{\mathrm{i2}}, \ldots)$ for each hypothetical participant from a multivariate Gamma distribution with n x n variance-covariance matrix $\sum$, where $\sum$ had first-order autoregressive structure with $\rho = 0.3$. The shape and rate of the Gamma distribution were set to 12 and 6, respectively, implying a marginal mean of 2 and a marginal variance of 0.33. After multiplying each Gamma draw by the corresponding exponentiated value of the linear predictor to compute λ_{ij}, the ijth count outcome was drawn from the Poisson(λ_{ij}) distribution.

We analyzed each simulated dataset by fitting two models. First, we fit the correct model: a negative binomial mixed model with Time and Arm x Time as explanatory variables, a random participant intercept, and a first-order autoregressive structure. Confidence intervals for the explanatory variables were computed using the KR, KR2, and between-within methods. Then we fit the same negative binomial mixed model but with compound symmetric covariance structure specified, meaning outcomes for each pair of time points for a given participant were incorrectly assumed to have the same correlation, regardless of the temporal distance between them. Again, confidence intervals were computed using the KR, KR2, and between-within methods. After excluding simulated datasets for which the software did not provide confidence limits for both fixed effects (e.g., due to failure to converge after 100 iterations of the GLIMMIX algorithm), we computed coverage rates and average interval lengths as in Study 1.

Table 1. Study 1: coverage rates and average lengths of 95% confidence intervals for Level 1 covariate.

		Coverage rate					
ICC = 0.10	Average cluster size	3 observations		6 observations		12 observations	
	Interval	KR[a]	BW[b]	KR	BW	KR	BW
	5 clusters	95.5	95.1	96.2	94.8	93.9	93.0
	10 clusters	96.1	94.9	94.5	93.4	93.3	93.3
	15 clusters	95.2	93.8	94.9	94.6	94.7	94.4
	20 clusters	95.8	95.4	94.5	94.1	95.0	94.6
		Average length					
	Average cluster size	3 observations		6 observations		12 observations	
	Interval	KR	BW	KR	BW	KR	BW
	5 clusters	0.15	0.13	0.07	0.07	0.05	0.04
	10 clusters	0.08	0.08	0.05	0.05	0.03	0.03
	15 clusters	0.07	0.06	0.04	0.04	0.02	0.02
	20 clusters	0.05	0.05	0.03	0.03	0.02	0.02
		Coverage rate					
ICC = 0.40	Average cluster size	3 observations		6 observations		12 observations	
	Interval	KR	BW	KR	BW	KR	BW
	5 clusters	96.2	95.5	95.3	94.2	95.4	95.0
	10 clusters	96.0	95.2	95.7	95.0	95.8	95.6
	15 clusters	95.3	94.5	94.8	94.6	93.3	93.3
	20 clusters	96.5	95.9	94.3	94.0	94.5	94.4
		Average length					
	Average cluster size	3 observations		6 observations		12 observations	
	Interval	KR	BW	KR	BW	KR	BW
	5 clusters	0.20	0.18	0.10	0.09	0.06	0.06
	10 clusters	0.11	0.11	0.06	0.06	0.04	0.04
	15 clusters	0.09	0.08	0.05	0.05	0.03	0.03
	20 clusters	0.07	0.07	0.04	0.04	0.03	0.03

[a] KR = Kenward–Roger interval.
[b] BW = between-within interval.

3. Results

Looking first to the cluster-randomized trial data of Study 1, coverage rates for the Kenward–Roger and between-within intervals were generally adequate, ranging from 93.2% to 96.8% and from 93.0% to 96.7%, respectively (Tables 1 and 2). For the Level 1 covariate the between-within intervals generally had lower coverage than the corresponding Kenward–Roger intervals. In some conditions this meant being closer to the nominal 95% rate, but more often it meant falling below it (Table 1). Coverage did not consistently improve with more clusters or larger average cluster size. For the Level 2 effect, coverage rates for the Kenward–Roger and between-within intervals were nearly identical in most cases (Table 2). For the effects at both levels, intervals tended to be wider for the higher ICC (0.40). This was much more pronounced for the Level 2 effect, whose average interval widths for the higher ICC were almost twice those for the lower ICC (0.10) across most combinations of cluster size and count. Estimation problems were rare in Study 1, even for the smallest sample size conditions; models converged and provided confidence limits for both fixed effects for 99.7% of simulated data sets under both the Kenward–Roger and between-within methods.

Turning to the temporally clustered data in Study 2, all three confidence interval types tended to be conservative (Tables 3 and 4). There was little difference in coverage rates between the KR and KR2 intervals. The between-within intervals were generally more precise (shorter), and although in a few cases their coverage was too low (90.8%-93.2%), their over-coverage was less severe, making them more accurate overall. Importantly, when the correct, autoregressive

Table 2. Study 1: coverage rates and average lengths of 95% confidence intervals for Level 2 effect.

		Coverage rate					
ICC = 0.10	Average cluster size	3 observations		6 observations		12 observations	
	Interval	KR[a]	BW[b]	KR	BW	KR	BW
	5 clusters	95.1	94.7	96.8	96.7	95.2	95.2
	10 clusters	95.9	96.1	94.9	94.9	94.3	94.3
	15 clusters	95.7	95.7	95.7	95.7	94.6	94.6
	20 clusters	93.2	93.1	95.3	95.3	95.0	95.0
		Average length					
	Average cluster size	3 observations		6 observations		12 observations	
	Interval	KR	BW	KR	BW	KR	BW
	5 clusters	1.71	1.68	1.72	1.72	1.66	1.66
	10 clusters	0.91	0.91	0.89	0.89	0.89	0.89
	15 clusters	0.70	0.70	0.70	0.70	0.69	0.69
	20 clusters	0.59	0.59	0.59	0.59	0.58	0.58
		Coverage rate					
ICC = 0.40	Average cluster size	3 observations		6 observations		12 observations	
	Interval	KR	BW	KR	BW	KR	BW
	5 clusters	95.6	95.8	95.7	95.7	95.7	95.6
	10 clusters	94.2	94.1	94.0	94.1	95.1	95.1
	15 clusters	94.7	94.7	95.6	95.6	93.5	93.4
	20 clusters	94.1	94.2	94.7	94.7	94.8	94.8
		Average length					
	Average cluster size	3 observations		6 observations		12 observations	
	Interval	KR	BW	KR	BW	KR	BW
	5 clusters	3.41	3.40	3.47	3.46	3.42	3.42
	10 clusters	1.78	1.78	1.82	1.82	1.81	1.81
	15 clusters	1.40	1.40	1.40	1.40	1.38	1.38
	20 clusters	1.17	1.17	1.17	1.17	1.16	1.16

[a] KR = Kenward–Roger interval.
[b] BW = between-within interval.

structure was specified for datasets with 3 time points, the average length of at least one of the two Kenward–Roger intervals was over 3.4-8.1 times the average length of the between-within interval, despite the three interval types having comparable coverage rates. Both the KR and KR2 intervals displayed this erratic behavior.

Surprisingly, interval coverage rates in Study 2 did not consistently improve with the number of clusters, though coverage was generally closer to 95% with 20 clusters than with 5. Similarly, coverage tended to be closer to the nominal rate when there were 9 time points than when there were only 3. Overall, specifying the auto-regressive covariance structure seemed to provide little benefit over fitting the simpler random intercept model with compound symmetry.

The three methods were generally comparable in the percentage of simulated data sets for which the model converged and produced confidence limits for both fixed effects of interest (Table 5). Compared to the KR and KR2 methods, the between-within method yielded both confidence limits for an additional 2.5–4.8% of simulated datasets when the autoregressive model was fit to datasets with 3 time points and 10–20 clusters. For the autoregressive model, estimation problems became more frequent for all three methods as the number of clusters increased and as the number of time points increased.

4. Discussion

Accurate small-sample inference for fixed effects in analyses of clustered data is critical, and the impact of the Kenward–Roger method on mixed modeling has been enormous; the original paper

Table 3. Study 2: coverage rates and average lengths of 95% confidence intervals for time.

		3 time points			6 time points			3 time points		
		KR[a]	KR2[b]	BW[c]	KR	KR2	BW	KR	KR2	BW
AR(1) model: Coverage rate	5 clusters	99.2	99.2	99.1	94.8	94.6	91.1	96.9	97.0	96.4
	10 clusters	98.6	98.6	98.2	98.3	98.3	98.1	97.1	97.1	96.8
	15 clusters	97.0	96.9	96.2	96.7	96.7	95.6	96.7	96.8	96.5
	20 clusters	97.4	97.7	97.0	96.8	96.7	96.7	96.5	96.6	96.6
		KR	KR2	BW	KR	KR2	BW	KR	KR2	BW
AR(1) model: Mean interval length	5 clusters	4.91	3.77	1.42	0.61	0.60	0.37	0.21	0.21	0.17
	10 clusters	3.54	2.12	0.83	0.27	0.27	0.24	0.13	0.13	0.12
	15 clusters	1.34	5.45	0.67	0.20	0.20	0.19	0.10	0.10	0.10
	20 clusters	1.77	2.89	0.54	0.17	0.17	0.16	0.09	0.08	0.08
		KR	KR2	BW	KR	KR2	BW	KR	KR2	BW
Compound symmetry model: Coverage rate	5 clusters	99.5	99.5	99.2	94.7	94.7	90.8	96.9	96.9	96.0
	10 clusters	98.9	98.9	98.5	98.4	98.4	98.0	96.8	96.8	96.7
	15 clusters	96.7	96.7	96.5	96.7	96.7	95.9	97.0	97.0	97.0
	20 clusters	97.7	97.7	97.0	96.6	96.6	96.6	96.6	96.6	96.6
		KR	KR2	BW	KR	KR2	BW	KR	KR2	BW
Compound symmetry model: Mean interval length	5 clusters	2.39	2.39	1.41	0.73	0.73	0.36	0.23	0.23	0.17
	10 clusters	1.24	1.24	0.83	0.29	0.29	0.24	0.13	0.13	0.12
	15 clusters	1.03	1.03	0.67	0.24	0.24	0.19	0.10	0.10	0.10
	20 clusters	0.76	0.76	0.54	0.18	0.18	0.16	0.08	0.08	0.08

[a]KR = Kenward–Roger (1997) interval.
[b]KR2 = Improved Kenward–Roger (2009) interval.
[c]BW = between-within interval.

Table 4. Study 2: coverage rates and average lengths of 95% confidence intervals for Time x Arm.

		3 time points			6 time points			9 time points		
		KR[a]	KR2[b]	BW[c]	KR	KR2	BW	KR	KR2	BW
AR(1) model: Coverage rate	5 clusters	99.6	99.6	99.5	98.5	98.6	97.2	95.9	95.8	92.3
	10 clusters	97.9	98.0	97.2	99.1	99.1	97.9	97.4	97.3	96.7
	15 clusters	97.8	97.7	97.1	95.2	95.0	93.7	96.1	96.0	95.3
	20 clusters	92.6	93.0	92.4	94.2	94.2	93.7	97.2	97.1	96.9
		KR	KR2	BW	KR	KR2	BW	KR	KR2	BW
AR(1) model: Mean interval length	5 clusters	6.24	5.34	1.49	0.73	0.74	0.32	0.23	0.23	0.15
	10 clusters	4.00	2.28	0.88	0.30	0.30	0.25	0.13	0.13	0.12
	15 clusters	1.52	5.45	0.70	0.21	0.21	0.18	0.09	0.09	0.09
	20 clusters	1.96	2.11	0.59	0.17	0.17	0.16	0.08	0.08	0.08
		KR	KR2	BW	KR	KR2	BW	KR	KR2	BW
Compound symmetry model: Coverage rate	5 clusters	99.9	99.9	99.6	98.4	98.4	97.1	95.7	95.7	92.5
	10 clusters	98.2	98.2	97.6	98.9	98.9	97.9	97.2	97.2	96.6
	15 clusters	98.0	98.0	97.4	95.5	95.5	94.2	96.5	96.5	96.0
	20 clusters	94.3	94.3	93.2	94.4	94.4	93.8	97.2	97.1	96.6
		KR	KR2	BW	KR	KR2	BW	KR	KR2	BW
Compound symmetry model: Mean interval length	5 clusters	2.54	2.54	1.49	0.66	0.66	0.32	0.22	0.22	0.15
	10 clusters	1.32	1.32	0.88	0.30	0.30	0.24	0.13	0.13	0.12
	15 clusters	1.08	1.08	0.70	0.23	0.23	0.18	0.09	0.09	0.09
	20 clusters	0.83	0.83	0.60	0.18	0.18	0.16	0.08	0.08	0.08

[a]KR = Kenward–Roger (1997) interval.
[b]Improved Kenward–Roger (2009) interval.
[c]BW = between-within interval.

has been cited thousands of times. The method has been studied primarily in linear mixed models, where Level 1 errors are assumed to be Gaussian. Count data, however, are discrete, bounded by zero on the left, often severely skewed to the right, and prone to yielding non-Gaussian residuals when modeled. Because neither the Kenward–Roger method nor the use of between-within degrees of freedom has been thoroughly studied for modeling clustered count data, we compared

Table 5. Percentage of simulated datasets in Study 2 for which model converged and provided confidence limits for both fixed effects of interest.

		3 time points			6 time points			9 time points		
		KR[a]	KR2[b]	BW[c]	KR	KR2	BW	KR	KR2	BW
AR(1) model	5 clusters	98.9	99.2	99.8	98.2	98.1	98.4	96.7	96.7	96.8
	10 clusters	96.7	97.1	99.2	96.6	96.4	96.8	94.7	94.2	95.2
	15 clusters	94.6	95.3	98.6	93.6	94.2	94.2	91.5	91.9	91.3
	20 clusters	93.5	94.4	98.3	94.8	93.6	94.5	85.9	86.0	86.4
		KR	KR2	BW	KR	KR2	BW	KR	KR2	BW
Compound symmetry model	5 clusters	99.9	99.9	99.9	99.8	99.7	99.7	99.7	99.8	99.7
	10 clusters	99.1	99.0	99.1	99.7	99.7	99.7	98.6	98.6	98.6
	15 clusters	99.3	99.3	99.3	97.7	97.6	97.7	99.3	99.0	99.3
	20 clusters	99.3	99.3	99.3	94.1	94.0	93.9	98.3	98.2	98.2

[a] KR = Kenward–Roger (1997) interval.
[b] Improved Kenward–Roger (2009) interval.
[c] BW = between-within interval.

their performance in confidence intervals using simulated data across a variety of small-sample conditions reflective of situations encountered in health-related research.

In our analyses the between-within intervals were often more precise than those computed using the Kenward–Roger method and, overall, provided comparably accurate, if not more accurate, coverage rates. In Study 1 coverage rates for both interval types were adequate (93–97%) and often in the 94–96% range, even for the small-sample conditions. Differences were more striking for the auto-correlated data in Study 2, where coverage rates overall were less accurate for both interval types. Here the between-within intervals were often shorter than their overly conservative Kenward–Roger counterparts and provided coverage rates closer to the nominal 95% rate under most conditions. In addition, the between-within intervals were less susceptible to estimation problems in the autoregressive models when there were only 3 time points per cluster.

The autoregressive and compound symmetry models in Study 2 performed about equally well overall in terms of interval coverage rates. Although the correct covariance specification (first-order autoregressive) provided a slight advantage in coverage when there were at least 15 clusters, there was a substantially higher rate of estimation problems for the autoregressive models than for the compound symmetry models under many sample size conditions. Estimation problems became more frequent as sample size increased.

It should be noted that although non-Gaussian data can often be transformed to yield Level 1 residuals consistent with the linear mixed model assumption of Gaussian errors, this approach can be problematic for count data. In a study involving comparisons of different approaches to modeling clustered binomial and count data, Stroup observed that “transformations consistently were no help and often made matters worse.” (Bell and Grunwald 2011, p. 122). Moreover, count data are not always amenable to normalizing transformation. For example, zeros cannot be log-transformed without first shifting them, along with all the non-zero counts, upward by some constant, a workaround we find unsatisfying. Further, if zero is the most common observed value in a set of count data, no rank-preserving transformation can move the mode to the middle of the distribution. It seems better to assume clustered count data follow a distribution with support on the non-negative integers and fit a GLMM or another appropriate model.

Critically, the results of this study alone cannot be used for definitive guidance on sample size requirements for GLMM analysis of clustered count data, which will depend on study specifics, including anticipated effect sizes, covariance structure and parameters, and numbers of fixed and random effects in the model. A sample size sufficient for accurate confidence interval coverage may be insufficient for estimating effects with the desired precision or power. We do however offer two general observations.

First, where interval coverage rates deviated markedly from the nominal rate for conditions with a small number of clusters (5 or 10), the problem was usually over-coverage. For these cases, in other words, the Kenward–Roger and between-within intervals over-corrected for anticonservative bias associated with use of residual degrees of freedom and the naïve EBLUE standard error approximation. Conservative bias with the Kenward–Roger method has been observed in other GLMM contexts, as well (Bell and Grunwald 2011; Li and Redden 2015; Jackson et al. 2021). In using either of these methods in practice, therefore, we would tend to be more concerned with the minimum number of clusters required to achieve the desired level of precision in estimating effects (or in terms of the binary hypothesis testing framework, the desired level of statistical power) than with having enough clusters to guard against interval under-coverage and over-estimated precision. Analyses with very small samples and complex covariance structures would be an exception (Schaalje, McBride, and Fellingham 2002).

Second, we note that recommendations for the number of clusters needed in mixed modeling are often too high, at least if accurate interval coverage for fixed effects is the criterion. Mixed modeling is sometimes described as a large-sample method, requiring dozens if not scores of clusters, but this need not be the case if we are not relying on asymptotic results (e.g., to overcome the effects non-Gaussian errors). In cluster-randomized trials even ten clusters can be prohibitively expensive, and it is important for researchers to know there are reasonably accurate methods for analyzing clustered count data when the number of clusters is not large.

As institutional data sharing and pervasive monitoring of health-related data both within and outside the clinical setting become increasingly common, the ability to draw accurate inferences from complex clustered data is critical. Given the prevalence of clustered count data in medical and health-related research, more guidance is needed regarding sample size and the choice of methods for standard error approximation and denominator degrees of freedom. This study represents a step toward development of such guidance. Based on the scenarios examined here, use of between-within degrees of freedom may be a suitable or even preferable alternative to the Kenward–Roger method in some analyses of clustered count data with simple covariance structures. Opportunities for advancing this line of research include considering more complex covariance structures, models with random slopes, and GLMMs for clustered data with other distributions.

ORCID

Vincent S. Staggs http://orcid.org/0000-0002-6232-9149

Data availability statement

Code used to simulate and analyze data is available upon request.

References

Bell, M. L., and G. K. Grunwald. 2011. Small sample estimation properties of longitudinal count models. *Journal of Statistical Computation and Simulation.* 81 (9):1067–79. doi:10.1080/00949651003674144.

Demidenko, E. 2004. *Mixed models: Theory and applications*. Hoboken, NJ: John Wiley & Sons.

Durán, P. G., J. Hattendorf, J. M. Colford, Jr, D. Mäusezahl, and T. Smith. 2009. Performance of analytical methods for overdispersed counts in cluster randomized trials: Sample size, degree of clustering and imbalance. *Statistics in Medicine* 28 (24):2989–3011. doi:10.1002/sim.3681.

Fialkowski, A., and H. Tiwari. 2019. SimCorrMix: Simulation of correlated data with multiple variable types including continuous and count mixture distributions. *The R Journal* 11 (1):250–86. doi:10.32614/RJ-2019-022.

Francq, B. G., D. Lin, and W. Hoyer. 2019. Confidence, prediction, and tolerance in linear mixed models. *Statistics in Medicine* 38 (30):5603–22. doi:10.1002/sim.8386.

Jackson, C. L., K. Colborn, D. Gao, S. Rao, H. C. Slater, S. Parikh, B. D. Foy, and J. Kittelson. 2021. Design and analysis of a 2-year parallel follow-up of repeated ivermectin mass drug administrations for control of malaria: Small sample considerations for cluster-randomized trials with count data. *Clinical Trials (London, England)*. Advance online publication. doi:10.1177/17407745211028581.

Kalema, G., and G. Molenberghs. 2016. Generating correlated and/or overdispersed count data: A SAS implementation. *J Stat Softw* 70 (Code Snippet 1). Advance online publication. doi:10.18637/jss.v070.c01.

Kenward, M. G., and J. H. Roger. 1997. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 53 (3):983–97.

Kenward, M. G., and J. H. Roger. 2009. An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics and Data Analysis.* 53 (7):2583–95. doi:10.1016/j.csda.2008.12.013.

Li, P., and D. T. Redden. 2015. 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 (1):38. doi:10.1186/s12874-015-0026-x.

Littell, R. C. 2002. Analysis of unbalanced mixed model data: A case study comparison of ANOVA versus REML/GLS. *Journal of Agricultural, Biological, and Environmental Statistics* 7 (4):472–90. doi:10.1198/108571102816.

Luke, S. G. 2017. Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods* 49 (4):1494–502. doi:10.3758/s13428-016-0809-y.

McNeish, D. 2019. Poisson multilevel models with small samples. *Multivariate Behavioral Research* 54 (3):444–55. doi:10.1080/00273171.2018.1545630.

McNeish, D., and L. M. Stapleton. 2016. Modeling clustered data with very few clusters. *Multivariate Behavioral Research* 51 (4):495–518. doi:10.1080/00273171.2016.1167008.

Schaalje, G. B., J. B. McBride, and G. W. Fellingham. 2002. Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural, Biological, and Environmental Statistics* 7 (4):512–24. doi:10.1198/108571102726.

Schluchter, M. D., and J. T. Elashoff. 1990. Small-sample adjustments to tests with unbalanced repeated measures assuming several covariance structures. *Journal of Statistical Computation and Simulation* 37 (1-2):69–87. doi:10.1080/00949659008811295.

Staggs, V. S. 2017. Comparison of naïve, Kenward–Roger, and parametric bootstrap interval approaches to small-sample inference in linear mixed models. *Communications in Statistics - Simulation and Computation* 46 (3):1933–43. doi:10.1080/03610918.2015.1019002.

Stroup, W. W. 2015. Rethinking the analysis of non-normal data in plant and soil science. *Agronomy Journal.* 107 (2):811–27. doi:10.2134/agronj2013.0342.

Stroup, W. W. 2018. Non-normal data in agricultural experiments. Conf. Appl. Stat. Agric. doi:10.4148/2475-7772.1018.