

RESEARCH

Open Access



A comparative study of *R* functions for clustered data analysis

Wei Wang^{1*}  and Michael O. Harhay^{1,2}

Abstract

Background: Clustered or correlated outcome data is common in medical research studies, such as the analysis of national or international disease registries, or cluster-randomized trials, where groups of trial participants, instead of each trial participant, are randomized to interventions. Within-group correlation in studies with clustered data requires the use of specific statistical methods, such as generalized estimating equations and mixed-effects models, to account for this correlation and support unbiased statistical inference.

Methods: We compare different approaches to estimating generalized estimating equations and mixed effects models for a continuous outcome in *R* through a simulation study and a data example. The methods are implemented through four popular functions of the statistical software *R*, “geese”, “gls”, “lme”, and “lmer”. In the simulation study, we compare the mean squared error of estimating all the model parameters and compare the coverage proportion of the 95% confidence intervals. In the data analysis, we compare estimation of the intervention effect and the intra-class correlation.

Results: In the simulation study, the function “lme” takes the least computation time. There is no difference in the mean squared error of the four functions. The “lmer” function provides better coverage of the fixed effects when the number of clusters is small as 10. The function “gls” produces close to nominal scale confidence intervals of the intra-class correlation. In the data analysis and the “gls” function yields a positive estimate of the intra-class correlation while the “geese” function gives a negative estimate. Neither of the confidence intervals contains the value zero.

Conclusions: The “gls” function efficiently produces an estimate of the intra-class correlation with a confidence interval. When the within-group correlation is as high as 0.5, the confidence interval is not always obtainable.

Keywords: Clustered data analysis, Cluster randomized trials, Generalized estimating equations, Mixed effects models

Introduction

Clustered data

Clustered data arise when the study population can be classified into different groups (referred to as clusters), and the measurements of subjects, in particular the response, within the same cluster are more alike than those in other clusters. For instance, in cluster-randomized trials, entire groups of participants such as classrooms, clinics, communities, or hospitals, rather than

individuals, are randomly assigned to intervention arms [1, 2].

The key feature of clustered data is that the similarity (or homogeneity) of measurements within the same cluster induces a correlation. That is, measurements within a cluster are likely to be correlated, whereas those from separate clusters are regarded as independent. The intra-class correlation coefficient (ICC) measures this similarity of the responses within a cluster and can be defined as a function of the variance components in the model: variation between clusters and within clusters [3, 4]. Since the responses within a cluster do not contribute completely independent information, the “effective” sample size is less than the total number of subjects from all clusters [5–7].

*Correspondence: wei.wang@penncmedicine.upenn.edu

¹Clinical Trials Methods and Outcomes Lab, Palliative and Advanced Illness Research (PAIR) Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Statistical methods

Classical statistical methods such as ordinary least squares regression assume that each individuals' data is independent. The clustered data have a hierarchical structure where individuals are not likely independent within the same cluster (i.e., $ICC > 0$). Thus, methods taking the correlation into account, such as generalized estimating equation (GEE) and mixed-effects models, are well suited for the analysis of clustered data [8–10].

GEE models can be viewed as an extension of generalized linear models for correlated data where a within-cluster correlation structure is specified [11]. Parameter estimates are then obtained as solutions of the estimating equations [12]. In mixed-effects models, the cluster effect is a random variable representing a random deviation for a given cluster from the overall fixed effects [13–15]. Maximum likelihood estimation is often used to obtain estimates of parameters via iterative algorithms such as the expectation-maximization (EM) algorithm and the Newton-Raphson algorithm [16–21]. As shown in [22], for normal outcomes, GEE reduces to the score equation of the maximum likelihood estimation only when there are no missing observations and the correlation is unstructured. A further comparison shows that GEE and mixed-effects models produce the same generalized least squares estimator of the fixed effects [23]. We will review the derivation of the fixed effects estimator in the next section.

Simulation studies have been conducted to compare the two methods for analyzing continuous outcomes with an emphasis on the fixed effects components. In the comparison of the estimation and the coverage probability of the confidence intervals, Park [22] found that the GEE estimation was more sensitive to missing observations. In the study [24], the authors compared the estimation and the nominal level of hypothesis testing and made several recommendations. For instance, if knowledge is available to specify the covariance structure correctly, the maximum likelihood estimation is slightly more efficient for balanced or near balanced data. When there is concern about the misspecification of the covariance structure, GEE is

preferred when the number of clusters is larger than 20. For hypothesis testing, Kahan et al. [25] and Leyrat et al. [26] found that without an appropriate correction, both methods can lead to inflated type I error rates (finding a statistically significant treatment effect when it does not exist) when the number of clusters is smaller than 40. *R*, *SAS*, and *Stata* commands to correct the type I error rate are provided in [26].

R functions

In this study, we compared the performance of the GEE method and the linear mixed-effects model to analyze clustered data through the implementation of both popular and newer packages of the statistical software *R* [27]. Specifically, the “geese” function of the *geepack* package (1.3.2) fits a GEE model [28, 29]. The “gls” function of the *nlme* package (3.1.149) [30, 31] fits a linear model using generalized least squares where the errors are allowed to be correlated. Two frequently used functions for conducting linear mixed-effects model analysis are “lme” of the *nlme* package [30, 31] and “lmer” of the *lme4* package [32] (1.1.25). Detailed implementation of these functions is provided in the “Implementation” subsection of the next section and also summarized in Table 1.

Methods

We compared the performance of the four functions via a simulation study and through a real data example. In the simulation study, we compared the computation time; the mean squared error (MSE) of estimating all the model parameters, including the ICC; and the coverage proportion of the 95% confidence intervals. Parameter estimates in the linear mixed-effects models are found by maximizing the likelihood of the data. In the following, we review the model setup followed by the simulation study and the example dataset.

Model review

Let n_i be the number of subjects in the i -th cluster and let $\mathbf{1}_i$ be a $n_i \times 1$ vector of one's. \mathbf{I}_i is a square identity matrix of dimension n_i and let \mathbf{J}_i be a square matrix of one's of

Table 1 Summary of how to obtain the CI's of the fixed effects and the variance-covariance parameters

	95% confidence intervals of parameters		
	β	$\{\sigma^2, \rho\}$	$\{\sigma_u, \sigma_\epsilon\}$
geese	Estimate \pm 1.96 standard deviation		–
gls	confint intervals(,which="coef")	intervals(,which="var-cov")*	–
lme	intervals(,which="fixed")	–	intervals(,which="var-cov")
lmer	confint(,method="Wald") confint(,method="profile")	–	confint(,method="profile")

*Applying the “intervals” function to a “gls” fitted object returns the CI of σ instead of σ^2

dimension n_i . Let u_i be the random effect associated with the i -th cluster. The linear mixed-effects modeling of the i -th cluster's response \mathbf{y}_i is

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{1}_i u_i + \epsilon_i, \quad u_i \sim N(0, \sigma_u^2), \quad \epsilon_i \sim N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_i).$$

The matrix \mathbf{X}_i is the design matrix and β is the unknown fixed effects. The random vector ϵ_i represents the error and is independent of u_i . Independence is also assumed between u_i and u_j , and between ϵ_i and ϵ_j whenever $i \neq j$. It follows that $\mathbf{y}_i \sim N(\mathbf{X}_i\beta, \Sigma_i)$, where $\Sigma_i = \sigma_u^2 \mathbf{J}_i + \sigma_\epsilon^2 \mathbf{I}_i$. Let elements of \mathbf{y}_i be y_{ij} , $j = 1, \dots, n_i$. We obtain $\text{Var}(y_{ij}) = \sigma_u^2 + \sigma_\epsilon^2$ and $\text{Cov}(y_{ij}, y_{ik}) = \sigma_u^2$ where $j \neq k$. The intra-class correlation coefficient ρ is naturally defined by the variance components σ_u^2 and σ_ϵ^2 as

$$\rho \equiv \text{Corr}(y_{ij}, y_{ik}) = \frac{\sigma_u^2}{\sigma^2}, \quad \text{where } \sigma^2 = \sigma_u^2 + \sigma_\epsilon^2. \quad (1)$$

By definition, $0 < \rho < 1$ since both σ_u^2 and σ_ϵ^2 are positive.

The above modeling uses the random effect u_i explicitly to explain within cluster correlation. From a marginal model perspective, one can instead start with the modeling $\mathbf{y}_i \sim N(\mathbf{X}_i\beta, \Sigma_i)$ directly with a special structure for Σ_i . Let $\rho = \text{Corr}(y_{ij}, y_{ik})$ and $\sigma^2 = \text{Var}(y_{ij})$, and sequentially we get $\Sigma_i = \sigma^2[\mathbf{I}_i + \rho(\mathbf{J}_i - \mathbf{I}_i)]$. Using this marginal parametrization $\{\sigma^2, \rho\}$, the matrix Σ_i is positive definite if $-1/(n_i - 1) < \rho < 1$ [30, 33–35]. That is, ρ does not have to be positive as it was defined from a variance components perspective. At the same time, we also observe that when n_i is large, the boundary $-1/(n_i - 1)$ can be very close to 0. Starting with this parameterization, we derive the corresponding relationship of (1) as

$$\sigma_u^2 = \sigma^2 \rho, \quad \sigma_\epsilon^2 = \sigma^2(1 - \rho). \quad (2)$$

It is clear that $\Sigma_i^{-1} = \{\mathbf{I}_i - \rho \mathbf{J}_i / [1 + (n_i - 1)\rho]\} / \sigma^2(1 - \rho)$. Let $\mathbf{T}_i(n_i, \rho) \equiv \mathbf{T}_i(\rho) = \mathbf{I}_i - \rho \mathbf{J}_i / [1 + (n_i - 1)\rho] = \sigma^2(1 - \rho) \Sigma_i^{-1}$. The estimate of β follows as $\hat{\beta} = \left\{ \sum_{i=1}^{n_c} \mathbf{X}'_i \Sigma_i^{-1} \mathbf{X}_i \right\}^{-1} \left\{ \sum_{i=1}^{n_c} \mathbf{X}'_i \Sigma_i^{-1} \mathbf{y}_i \right\}$ which reduces to

$$\hat{\beta} = \left\{ \sum_{i=1}^{n_c} \mathbf{X}'_i \mathbf{T}_i(\rho) \mathbf{X}_i \right\}^{-1} \left\{ \sum_{i=1}^{n_c} \mathbf{X}'_i \mathbf{T}_i(\rho) \mathbf{y}_i \right\}.$$

The variance-covariance matrix of the estimate has the form

$$\text{Cov}(\hat{\beta}) = \left\{ \sum_{i=1}^{n_c} \mathbf{X}'_i \Sigma_i^{-1} \mathbf{X}_i \right\}^{-1} = \sigma^2(1 - \rho) \left\{ \sum_{i=1}^{n_c} \mathbf{X}'_i \mathbf{T}_i(\rho) \mathbf{X}_i \right\}^{-1}.$$

We notice that if there is no within cluster correlation, i.e. $\rho = 0$, then $\mathbf{T}_i(0) = \mathbf{I}_i$ and $\hat{\beta}$ is simply the ordinary least squares model. In the extreme case of perfect correlation, i.e., $\rho = 1$, then $\mathbf{T}_i(1) = \mathbf{I}_i - \mathbf{J}_i/n_i$ and we get

$$\hat{\beta}|_{\rho=1} = \left[\sum_{i=1}^{n_c} \mathbf{X}'_i \mathbf{T}_i(1) \mathbf{X}_i \right]^{-1} \left[\sum_{i=1}^{n_c} \mathbf{X}'_i (\mathbf{y}_i - \mathbf{1}_i \bar{y}_i) \right].$$

In this scenario, $y_{ij} = y_{ik}$ and $\mathbf{y}_i = \bar{y}_i \mathbf{1}_i$, so $\hat{\beta} = \mathbf{0}$.

Simulation

Our simulation setups are similar to those in the literature. In the simulation, Feng et al. [24] tried the number of clusters $n_c = 10, 20$, and 50 with cluster sizes 10, 30, and 100, and $\rho = 0.1, 0.5$. The simulation study [25] considered two scenarios of 5 patients per cluster with $\rho = 0.15$, and of 100 patients per cluster with $\rho = 0.01$. The number of clusters n_c varied from 6, 10, 20, ..., 90, and 100. The simulation scenarios in [26] include $\rho = 0.001, 0.01$, and 0.05, and $n_c = 4, 6, 8, 10, 20, 30, 40$, and 200. The average cluster size ranges from 7 to 300. A review of published cluster-randomized trials by Kahan et al. [25] shows that the median number of clusters was 25 with interquartile range 15 to 44, 14% of the trials had fewer than 10 clusters, and 9% of the trials had more than 100 clusters. The cluster size had a median of 31 and an interquartile range 14 to 94.

In our study, we tried $n_c = 10, 30, 50$, and 100. In each cluster, the number of subjects was simulated from a normal distribution after rounding with mean $m = 50$ and 100, and standard deviation 5. The sample size calculation in [36] preassumed an ICC of 0.05. Some studies also found large ICC values such as 0.47 with 95% confidence interval [0.29, 0.65] [6, Table I] and 0.60 [37, Table 4.4.2]. In our setup, we considered $\rho = 0.05, 0.1$, and 0.5. We simulated two covariates independently from a Bernoulli distribution with a probability of success of 0.5, and from a standard Normal distribution. The associated regression coefficients are respectively $\beta_1 = -2$, and $\beta_2 = 1.5$, and the intercept in the regression model is $\beta_0 = 1$. We simulated the outcome from a marginal model with the variance parameter $\sigma^2 = 0.6$. In each of the settings, we examined 2000 simulations.

Example dataset

We reanalyzed data from the cluster-randomized controlled trial in [36]. In this study, participants with hypertension from 15 clusters in rural India were recruited and randomized to the intervention or usual care in a 1:2 ratio. The study hypothesis was that a CHW (community health worker)-led group-based education and monitoring intervention would result in improved blood pressure control. Outcomes were assessed approximately two months after completion of the intervention.

One of the main outcomes in this trial was the change in diastolic blood pressure (DBP), defined as follows: the DBP at baseline minus the DBP at follow-up. Fixed effects in the analysis include the following variables: age, sex,

diastolic blood pressure at baseline (mm Hg), education, use of antihypertensive medications, change in BMI (body mass index) defined as BMI at follow-up minus the BMI at baseline, number of serves of fruit per week at follow-up, and self-reported drinking alcohol at least once in the 30 days prior to follow-up. The education variable has four categories: no formal schooling, class 1 to 6, class 7 to 11, and class 12 or more.

We analyzed the data from the 1428 participants with no missing values. The histogram of the outcome shows a bell shaped pattern (Fig. 1). The intervention group has a larger proportion of a positive difference than the usual care group suggesting more DBP decline at the follow-up. The normal quantile-quantile plot in Fig. 2 shows the normal distribution assumption of the outcome is plausible which provides a justification of the application of linear mixed-effects models.

Implementation

All the four *R* functions compute $\hat{\beta}$ and the corresponding confidence interval (CI), but they adopt different parameterizations for the variance-covariance matrix. The function “geese” uses $\{\sigma^2, \rho\}$ though a specification of an

“exchangeable” correlation structure. The function “glms” uses a compound symmetry structure of the parameters $\{\sigma, \rho\}$. Both “lme” and “lmer” find estimates and CI’s of $\{\sigma_u, \sigma_\epsilon\}$. It is straightforward to obtain estimates of σ^2 , or σ_u^2 and σ_ϵ^2 from their corresponding square root estimates. We can then find estimates of another parameterization using Eqs. (1) or (2). As the “geese” method does not fit in the framework of hierarchical modeling with random effects, it is not appropriate to find its estimates of $\{\sigma_u^2, \sigma_\epsilon^2\}$ using Eq. (2) due to a possible negative $\hat{\rho}$. Thus, we do not include it for the comparison of estimating $\{\sigma_u, \sigma_\epsilon\}$. Methods obtaining the CI’s of ρ have been discussed in [38, 39]. In the model-based setup, the CI of ρ is readily available from the output of the “geese” or “glms” fitted object. Below we explain how to get the CI’s of the model parameters with the summary presented in Table 1.

From the “geese” output, we apply the estimate ± 1.96 standard deviation rule to obtain the CI’s. We apply two generic functions “confint” and “intervals” to “glms”, “lme”, or “lmer” fitted objects. The “confint” function assumes normality and has two options. The option *method* = “Wald” returns approximate CI’s of the fixed effects based on the estimated local curvature of the likelihood surface.

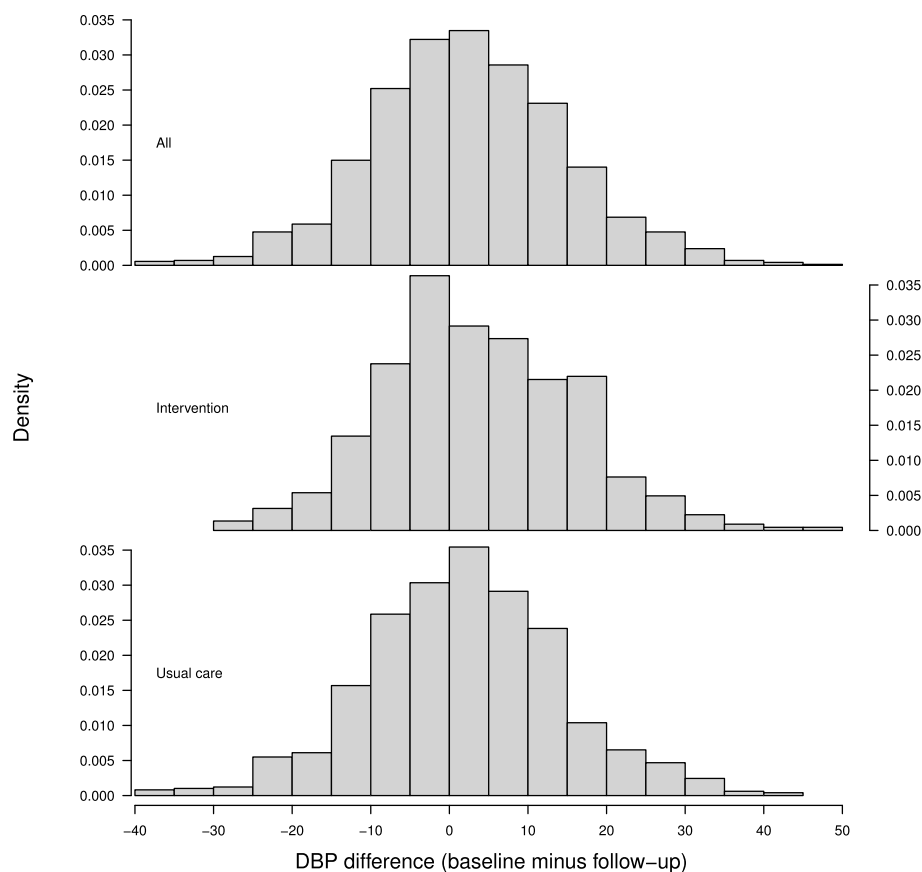
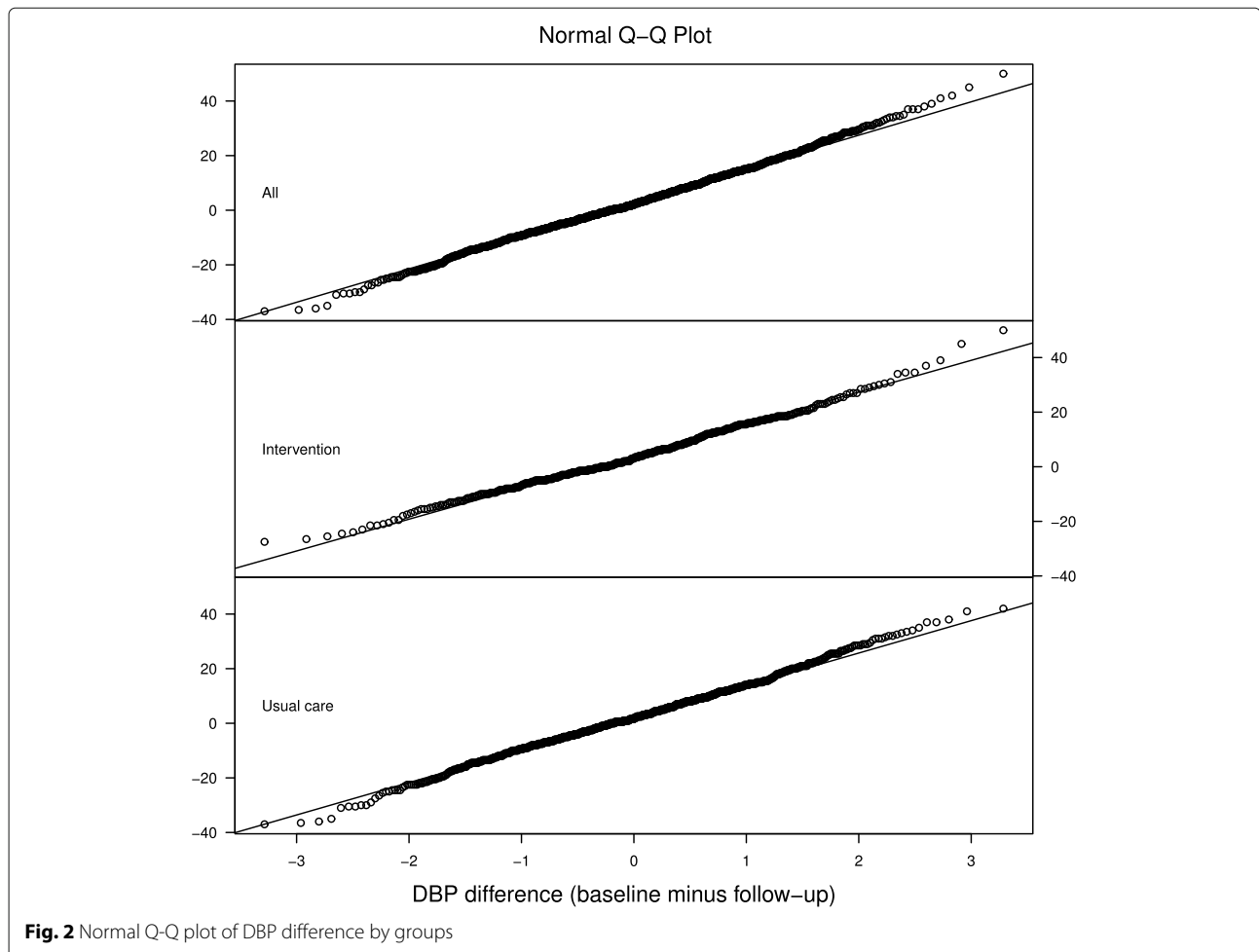


Fig. 1 Histogram of DBP difference by groups



The other option *method*="profile" computes a likelihood profile and find the appropriate cutoffs based on the likelihood ratio test [32]. The "intervals" function calculates approximate confidence intervals for the parameters in the linear model using a normal approximation to the distribution of the maximum likelihood estimators. The estimators are assumed to have a normal distribution centered at the true parameter values and with covariance matrix equal to the negative inverse Hessian matrix of the log-likelihood evaluated at the estimated parameters [30, 31].

Results

Simulation

With detailed comparison results presented in the supplementary material, we summarize our findings in the following text. The "lme" approach took the least computation time, 1.4 h, followed by the "gls" approach, 2.41 h (Table S1). Sometimes the "gls" approach failed to construct the confidence intervals for the variance-covariance parameters when $\rho=0.5$ (Table S2). The number of failure increases with the number of clusters n_c and also the

cluster size m . When $n_c = 100$ and $m = 100$, there were 528 failures out of 2000 simulations. Occasionally, the "lme" and "lmer" approaches also failed to construct the confidence intervals. The performance of the four approaches of estimating the model parameters is very similar with almost identical standard deviation and MSE (Tables S3-S6).

Next we summarize the performance of the different functions on the coverage proportion of the model parameters. First, in general, the coverage proportions of the fixed effects are very similar among "gls", "lme", and "lmer" approaches (Tables S7-S9). They are very close to the nominal 95% level except for β_0 when $n_c = 10$ (Table S7). In that case, confidence intervals obtained by specifying the "profile" option of the "confint" function to a "lmer" fitted object outperforms the others. Their coverage proportions are generally closer to the nominal 95% level than the "geese" approach when n_c is less than 100. Second, though the coverage proportions of $\{\sigma^2, \rho\}$ of "geese" or $\{\sigma, \rho\}$ of "gls" are usually below the 95% nominal level, the "gls" method generally provided better coverage (Table S10). Third, both the "lme" and "lmer" produced

Table 2 CI's of the fixed effects and the variance-covariance parameters

	Intervention effect	ρ	σ^2 or σ	σ_u	σ_ϵ
geese	2.161 (1.060, 3.263)	-0.076 (-0.125, -0.027)	101.768 (93.518, 110.019)		
gls	2.252 (0.597, 3.907)	0.012 (3.7×10^{-4} , 0.405)	10.093 (9.727, 10.473)	1.107	10.032
	2.252 (0.595, 3.908)				
lme	2.252 (0.602, 3.901)	0.012	10.093	1.107 (0.589, 2.081)	10.032 (9.669, 10.409)
lmer	2.252 (0.604, 3.900)	0.012	10.093	1.107 (0.234, 2.111)	10.032 (9.673, 10.414)
	2.252 (0.525, 4.086)				

The “geese” method finds the point estimate and CI of the parameter σ^2 instead of σ as the other methods. The order of the CI's of the intervention follows the order of the CI's of β in Table 1

coverage proportions about 95% for σ_ϵ , and for σ_u when $n_c \neq 10$. When $n_c = 10$, we observed unstable performance of over coverage or under coverage (Table S11).

Example dataset

All the four methods produced similar results that suggested more DBP reduction in the intervention group, 2.161 mm Hg by “geese” and 2.252 by the other three methods (Table 2). The 95% confidence interval bounds of the intervention effect are slightly different. However, our conclusion is consistent with the finding in [36] where the analysis was conducted in Stata (Stata IC/11.2, StataCorp, College Station, TX, USA). DBP declined 2.1 mm Hg more in the intervention group with a 95% confidence interval of (0.6, 3.6), and the estimation of ICC was 0.02. The “gls” method gives a positive $\hat{\rho}$ and a confidence interval does not contain 0. The “geese” method produces a negative $\hat{\rho}$ with negative confidence interval bounds.

Discussion and conclusion

Throughout our study, we compare the performance of the four R functions, “geese”, “gls”, “lme”, and “lmer”, of analyzing single level clustered data. The “exchangeable” correlation structure of the “geese” function and the compound symmetry structure of the function “gls” both provide a single-level cluster model. We note that the “lme” and the “lmer” function can model multi-level data, and the “lmer” function is capable of modeling crossed random effects. The *lme4* package also includes generalized linear mixed model capability via the “glmer” function. It does not currently implement *nlme*'s features for modeling heteroscedasticity of residuals or offer the same flexibility for composing complex variance-covariance structures.

Our simulation study found that all four methods perform equally well for model parameters estimation. This result is consistent with the study in [24]. It was found that the MSEs are very similar except when the number of clusters is 10 where the linear mixed-effects model method has slightly smaller MSE than the GEE method using SAS PROC MIXED. We observe similar coverage proportions of the fixed effects among “gls”, “lme,” and

“lmer” approaches. They are generally closer to the nominal 95% level than the “geese” approach when the number of clusters is less than 100.

The estimated ICC from the “geese” method can be negative, and the confidence interval of the ICC from the “gls” method provides better coverage. However, when the ICC is large as $\rho = 0.5$, confidence intervals is not always obtainable from the “gls” method. In our comparison of the coverage of the variance-covariance parameters in the model, the “lme” and the “lmer” methods have similar performance while the former is considerably faster. The latter provides better coverage of the intercept in the model when the number of clusters is 10.

In the simulation settings, we examined that the “gls” function is preferable to analyze single-level clustered data. The limitations of our simulation study include the lack of the scenario of the very large number of clusters (e.g., 200 as in [26]) or the scenario of small ICC values (e.g., 0.01 as in [25, 26]). It may also be of interest to further compare the performance of the four functions for complex trials such as stepped-wedge cluster-randomized trials. In a stepped-wedge cluster-randomized trial, all clusters begin in the control phase and then are randomized to interventions at different time points [40–43]. Simulations have been conducted to investigate the effect of varying degrees of imbalance in cluster size on the power [44].

Abbreviations

BMI: Body mass index; CHW: Community health worker; CI: Confidence interval; DBP: Diastolic blood pressure; EM: Expectation-maximization; GEE: Generalized estimating equation; ICC: Intra-class correlation coefficient; MSE: Mean squared error

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13063-021-05900-7>.

Additional file 1: The supplementary material in a pdf format contains the tables of the simulation study results.

Acknowledgements

The authors would like to acknowledge the availability of the data in the study of [36], which were made publicly accessible by the PLOS Medicine journal. PLOS journals require authors to make all data necessary to replicate their

study's findings publicly available without restriction at the time of publication (<https://journals.plos.org/plosmedicine/s/data-availability>).

Authors' contributions

Both authors made substantial contributions to the manuscript. WW conducted the data analysis, simulation studies, and drafted the manuscript. MOH reviewed and revised the article. Both authors approved the manuscript to be submitted.

Funding

Michael O. Harhay was supported by grant R00 HL141678 from the National Heart, Lung, and Blood Institute (NHLBI) of the US National Institutes of Health (NIH). The funding body had no role in the study design, data collection and analysis, or decision to publish.

Availability of data and materials

The data analyzed in our example were originally studied in [36]. The data file and the data dictionary are publicly available from the figshare database (accession number [dx.doi.org/10.26180/5cc80de987113](https://doi.org/10.26180/5cc80de987113)). The data were reanalyzed without collaboration with the original study authors.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Clinical Trials Methods and Outcomes Lab, Palliative and Advanced Illness Research (PAIR) Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. ²Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

Received: 27 November 2020 Accepted: 1 December 2021

Published online: 27 December 2021

References

- Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. *Am J Public Health*. 2004;94(3):423–32.
- Campbell M, Donner A, Klar N. Developments in cluster randomized trials and statistics in medicine. *Stat Med*. 2007;26(1):2–19.
- Fisher RA. *Statistical Methods for Research Workers*, 5th edn. Edinburgh: Oliver and Boyd Ltd.; 1934.
- Eldridge SM, Ukoumunne OC, Carlin JB. The intra-cluster correlation coefficient in cluster randomized trials: a review of definitions. *Int Stat Rev*. 2009;77(3):378–94.
- Baio G, Copas A, Ambler G, Hargreaves J, Beard E, Omar RZ. Sample size calculation for a stepped wedge trial. *Trials*. 2015;16(354). <https://pubmed.ncbi.nlm.nih.gov/26282553/>.
- Campbell MK, Mollison J, Grimshaw JM. Cluster trials in implementation research: estimation of intracluster correlation coefficients and sample size. *Stat Med*. 2001;20(3):391–9.
- Rutterford C, Copas A, Eldridge S. Methods for sample size determination in cluster randomized trials. *Int J Epidemiol*. 2015;44(3):1051–67.
- Lee KJ, Thompson SG. The use of random effects models to allow for clustering in individually randomized trials. *Clin Trials*. 2005;2(2):163–73.
- Barker D, McElduff P, D'Este C, Campbell M. Stepped wedge cluster randomised trials: a review of the statistical methodology used and available. *BMC Med Res Methodol*. 2016;16(69). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4895892/>.
- Turner EL, Prague M, Gallis JA, Li F, Murray DM. Review of recent methodological developments in group-randomized trials: part 2-analysis. *Am J Public Health*. 2017;107(7):1078–86.
- Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73(1):13–22.
- Zeger SL, Liang K-Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*. 1986;1:121–30.
- Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics*. 1982;38:963–74.
- Ware JH. Linear models for the analysis of longitudinal studies. *Am Stat*. 1985;39(2):95–101.
- Laird N, Lange N, Stram D. Maximum likelihood computations with repeated measures: application of the em algorithm. *J Am Stat Assoc*. 1987;82(397):97–105.
- Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B*. 1977;39:1–38.
- Wu JCF. On the convergence properties of the EM algorithm. *Ann Stat*. 1983;11(1):95–103.
- Jennrich RI, Schluchter MD. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*. 1986;42:805–20.
- Lindstrom M, Bates D. Newton-Raphson and EM algorithms for linear mixed effects models for repeated measures data. *J Am Stat Assoc*. 1988;83:1014–22.
- Meng XL, Rubin DB. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*. 1993;80:267–78.
- Liu CH, Rubin DB. The ECME algorithm - a simple extension of EM and ECM with faster monotone convergence. *Biometrika*. 1994;81(4):633–48.
- Park T. A comparison of the generalized estimating equation approach with the maximum likelihood approach for repeated measurements. *Stat Med*. 1993;12(18):1723–32.
- Wu C-T, Gumpertz ML, Boos DD. Comparison of GEE, MINQUE, ML, and REML estimating equations for normally distributed data. *Am Stat*. 2001;55(2):125–30.
- Feng Z, McLerran D, Grizzle J. A comparison of statistical methods for clustered data analysis with gaussian error. *Stat Med*. 1996;15(16):1793–806.
- Kahan BC, Forbes G, Ali Y, Jairath V, Bremner S, Harhay MO, Hooper R, Wright N, Eldridge SM, 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*. 2016;17. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5013635/>.
- Leyrat C, Morgan KE, Leurent B, Kahan BC. Cluster randomized trials with a small number of clusters: which analyses should be used?. *Int J Epidemiol*. 2018;47(1):321–31.
- R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2020. R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Yan J, Fine JP. Estimating equations for association structures. *Stat Med*. 2004;23:859–80.
- Halekoh U, Højsgaard S, Yan J. The R package geepack for generalized estimating equations. *J Stat Softw*. 2006;15(2):1–11.
- Pinheiro J, Bates D. *Mixed-Effects Models in S and S-PLUS*. New York: Springer; 2009.
- Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team. nlme: Linear and Nonlinear Mixed Effects Models. 2020. R package version 3.1-149. <https://cran.r-project.org/web/packages/nlme/ChangeLog>. Accessed Aug 2020.
- Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw*. 2015;67(1):1–48. [10.18637/jss.v067.i01](https://doi.org/10.18637/jss.v067.i01).
- Wang W. Identifiability of linear mixed effects models. *Electron J Stat*. 2013;7:244–63.
- Wang W. Identifiability of covariance parameters in linear mixed effects models. *Linear Algebra Appl*. 2016;506:603–13.
- Wang W. Checking identifiability of covariance parameters in linear mixed effects models. *J Appl Stat*. 2017;44(11):1938–46.
- Gamege DG, Riddell MA, Joshi R, Thankappan KR, Chow CK, Oldenburg B, Evans RG, Mahal AS, Kalyanram K, Kartik K, Suresh O, Thomas N, Mini GK, Maulik PK, Srikanth VK, Arabshahi S, Varma RP, Guggilla RK, D'Esposito F, Sathish T, Alim M, Thrift AG. Effectiveness of a scalable group-based education and monitoring program, delivered by health workers, to improve control of hypertension in rural india: A cluster randomised controlled trial. *PLoS Med*. 2020;17:1–22.
- Pal N, Lim WK. On intra-class correlation coefficient estimation. *Stat Pap*. 2004;45:369–92.
- Ukoumunne OC. A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials. *Stat Med*. 2002;21(24):3757–74.

39. Demetrashvili N, Wit EC, van den Heuvel ER. Confidence intervals for intraclass correlation coefficients in variance components models. *Stat Methods Med Res.* 2016;25(5):2359–76.
40. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials.* 2007;28(2):182–91.
41. Beard E, Lewis JJ, Copas A, Davey C, Osrin D, Baio G, Thompson JA, Fielding KL, Omar RZ, Ononge S, et al. Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014. *Trials.* 2015;16(353). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4538902/>.
42. Kristunas C, Morris T, Gray L. Unequal cluster sizes in stepped-wedge cluster randomised trials: a systematic review. *BMJ Open.* 2017;7(11):10.1136/bmjopen-2017-017151. <https://bmjopen.bmj.com/content/7/11/e017151.full.pdf>.
43. Li F, Hughes JP, Hemming K, Taljaard M, Melnick ER, Heagerty PJ. Mixed-effects models for the design and analysis of stepped wedge cluster randomized trials: An overview. *Stat Methods Med Res.* 2020;30(2):612–39. <https://pubmed.ncbi.nlm.nih.gov/32631142/>.
44. Kristunas CA, Smith KL, Gray LJ. An imbalance in cluster sizes does not lead to notable loss of power in cross-sectional, stepped-wedge cluster randomised trials with a continuous outcome. *Trials.* 2017;18(109). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5341460/>.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

