Simulation-based evaluation of the linear-mixed model in the presence of an increasing proportion of singletons

BRUYNDONCKX, Robin; HENS, Niel & AERTS, Marc (2018) Simulation-based evaluation of the linear-mixed model in the presence of an increasing proportion of singletons. In: BIOMETRICAL JOURNAL, 60(1), p. 49-65.

DOI: 10.1002/bimj.201700025
Handle: http://hdl.handle.net/1942/28744

Biometrical Journal **52** (2017) 61, zzz–zzz / DOI: 10.1002/bimj.DOIsuffix

Simulation-based evaluation of the linear mixed model in the presence of an increasing proportion of singletons

Robin Bruyndonckx[*,1,2], **Niel Hens** [1,3], and **Marc Aerts** [2]

[1] Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BIOSTAT), Hasselt University, Diepenbeek, Belgium

[2] Laboratory of Medical Microbiology, Vaccine & Infectious Disease Institute (VAXINFECTIO), University of Antwerp, Antwerp, Belgium

[3] Centre for Health Economic Research and Modelling of Infectious Diseases (CHERMID), Vaccine & Infectious Disease Institute (VAXINFECTIO), University of Antwerp, Antwerp, Belgium

Received zzz, revised zzz, accepted zzz

Data in medical sciences often have a hierarchical structure with lower-level units (e.g. children) nested in higher-level units (e.g. departments). Several specific but frequently studied settings, mainly in longitudinal and family research, involve a large number of units that tend to be quite small, with units containing only one element referred to as singletons. Regardless of sparseness, hierarchical data ~~are generally~~ should be analysed with appropriate methodology such as e.g. linear mixed models.
Using a simulation study, based on the structure of a data example on ceftriaxone consumption in hospitalized children, we assess the impact of an increasing proportion of singletons ($\mathbf{0} - \mathbf{95}\%$), in data with a low, medium or high intracluster correlation, on the stability of linear mixed models parameter estimates, confidence interval coverage and F test performance. Some techniques that are frequently used in the presence of singletons include ignoring clustering, dropping the singletons from the analysis and grouping the singletons into an artificial unit. We show that both the fixed and random effects estimates and their standard errors are stable in the presence of an increasing proportion of singletons. We demonstrate that ignoring clustering and dropping singletons should be avoided as they come with biased standard error estimates. Grouping the singletons into an artificial unit might be considered, although the linear mixed model performs better even when the proportion of singletons is high.
~~We conclude that in the presence of a high proportion of singletons, the linear mixed model is stable. Ignoring clustering, grouping or removing singletons should be avoided at all times.~~ We conclude that the linear mixed model is stable in the presence of singletons when both lower- and higher level sample sizes are fixed. In this setting, the use of remedial measures, such as ignoring clustering and grouping or removing singletons, should be dissuaded.

Key words: F test; Hierarchical data; Intracluster correlation; Performance characteristics; Sparseness;
Supporting Information for this article is available from the author or on the WWW under http://dx.doi.org/10.1022/bimj.XXXXXXX

1 Introduction

Data that are collected in e.g. medical sciences often have a hierarchical structure. This means that units at a lower level (secondary units) are nested within units at a higher level (primary units) (Snijders and Bosker, 1999). Some well-known examples of such hierarchies include patients nested within hospitals, workers nested within factories and animals nested within litters. Multi-level hierarchies also occur frequently (e.g. students nested within classes within schools within cities within countries). As subjects that are nested within one unit tend to be more alike than subjects from different units, the observations are typically no

*Corresponding author: e-mail: robin.bruyndonckx@uhasselt.be, Phone: +32-11-268246, Fax: +32-11-268298

longer independent. Ignoring the correlation within clusters will usually cause a downward bias in the standard errors, resulting in possible misinterpretation of the effect of predictor variables (Garson, 2013; Hox, 1998; Kreft and De Leeuw, 1998; Moulton, 1986). To account for the hierarchical nature of the data, linear mixed models are often used (Goldstein, 2003; Verbeke and Molenberghs, 2009).
Fitting such models can be done with a statistical software package such as SAS. A description on the use of the SAS `PROC MIXED` procedure to fit a linear mixed model is given by Littell *et al.* and Singer (Littell et al., 2006; Singer, 1998). For a comprehensive elaboration on linear mixed models, we refer to the books by Snijders and Bosker, Goldstein, Raudenbush and Bryk, Hox and Wang *et al.* (Goldstein, 2003; Hox, 2010; Raudenbush and Bryk, 2002; Snijders and Bosker, 1999; Wang et al., 2012). For some illustrations on the application of linear mixed models to hierarchical data we refer to Goldstein *et al.*, Renard *et al.* and Lee (Goldstein et al., 1993; Lee, 2000; Renard et al., 1998).

The study presented in this paper was motivated by the Ceftriaxone data, which contain information on doses (expressed in mg/kg/day) of ceftriaxone prescribed to hospitalized children (one day surveillance in September 2011). These data were collected within work package 5 (European Neonatal and Paediatric Antimicrobial Point Prevalence Survey) of the ARPEC project (Antibiotic Resistance and Prescribing in European Children), which was set up to improve the quality of antibiotic prescribing and reduce the prevalence of antimicrobial resistance in European children, and is described in detail elsewhere (Versporten et al., 2013). The Ceftriaxone data include 329 children hospitalized in 124 departments. Figure 1 shows that in most departments, only few children were prescribed with ceftriaxone, which resulted in 47.6% singleton departments.

[Figure 1 about here.]

Most multi-level settings consist of a small number of units at the primary level that tend to be quite large. When these units contain only a small number of secondary units, this is referred to as primary unit sparseness. Examples are macro-geographical regions containing only a few countries (Australia) or schools in which only a few classes decide to participate in a study. Regardless of primary unit sparseness, such data are generally analysed with linear mixed models in which F tests are used to evaluate the significance of the included explanatory variables. We have however demonstrated that the F test becomes unstable in the presence of primary unit sparseness and the permutation test is a more trustworthy alternative (Bruyndonckx et al., 2016).

Another specific but frequently studied setting (e.g. family research) involves a large number of units that tend to be quite small. In the Ceftriaxone data, 47% of included departments contain only one child (referred to as singleton). Regardless of sparseness caused by the high proportion of singletons, such data are generally analysed with a linear mixed model. Several studies to determine the impact of small cluster sizes on different aspects of the linear mixed model showed that both residual and random effects variance were biased when the number of subjects within the units is small. The impact on fixed effects appeared to be smaller, as both fixed effects estimates and their standard error were unbiased in the presence of small clusters (Bell et al., 2014; Clarke, 2008; Maas and Hox, 2005). Although the small sample setting has been extensively studied and renders promising results, we are specifically interested in the setting with different proportions of singletons. To our knowledge, only a few studies assessed the impact of singletons on the linear mixed model. Pickering and Weatherall investigated a setting with 15% of singletons and found that fixed effects estimates and standard errors were unbiased (Pickering and Weatherall, 2007). Sauzet *et al.* studied a setting with $80 - 99\%$ of singletons and found that parameter estimates for fixed effects were biased when the proportion of singletons became extreme (Sauzet et al., 2012). While these studies already give an idea about the impact of the presence of singletons, they focus on specific singleton proportions (either very high or fairly low) and use rather simple models only containing explanatory variables at the lowest level of the hierarchy.

In this paper, we will use a simulation study, keeping lower- and higher-level sample sizes fixed, to assess the impact of an increasing proportion of singletons ($0-95\%$) on different aspects of the linear mixed model fitted using restricted maximum likelihood. Here, we will focus ~~focussing~~ on a two-level setting with a low, medium or high intracluster correlation, and includ~~ing~~e explanatory variables both at the primary and at the secondary level. We will assess whether, when high proportions of singletons are present, the model's performance improves by applying some frequently used techniques to cope with singletons: ignoring the dependency within units, removing singletons and grouping singletons in an artificial unit.

2 Methods

To reflect a realistic setting, we set up a simulation study based on the Ceftriaxone data, which contain information on the prescribed ceftriaxone dose for 329 children in 124 departments. Included explanatory variables are the size of the department (large, medium, small), the reason for treatment of the child (mild, moderate, severe, different) and the age of the child. Parameter estimates and standard errors obtained by fitting a linear mixed model, that accounts for the correlation within departments by including a random intercept (unmodified model), are given in Table 1.

The unmodified model can be presented as follows:

$$Y_{ij} = \beta_0 + b_{0j} + \beta_1 Size_{1j} + \beta_2 Size_{2j} + \beta_3 Age_{ij} + \beta_4 Reason_{1ij} + \beta_5 Reason_{2ij} + \beta_6 Reason_{3ij} + \epsilon_{ij}, \quad (1)$$

where Y_{ij} represents the ceftriaxone dose prescribed to child i ($i = 1, ..., n_j$) in department j ($j = 1, ..., J$), n_j is the number of children in department j, J is the number of included departments, β_0 is the general intercept, b_{0j} is the department-specific intercept, $Size_{1j}$ is 1 if department j is large and 0 otherwise, $Size_{2j}$ is 1 if department j is medium and 0 otherwise, Age_{ij} is the age of child i in years, $Reason_{1ij}$ is 1 if the reason for treatment is different and 0 otherwise, $Reason_{2ij}$ is 1 if the reason for treatment is mild and 0 otherwise, $Reason_{3ij}$ is 1 if the reason for treatment is moderate and 0 otherwise, β_1 up to β_6 are the respective coefficients for the listed parameters and ϵ_{ij} is the residual error term. We assume that the ~~random effect~~ department-specific intercept follows a normal distribution with mean zero and variance σ_B^2 and that the error terms are independent and follow a normal distribution with mean zero and variance σ_W^2.

[Table 1 about here.]

For computational feasibility and to be able to vary the intracluster correlation and the percentage of singletons included, we considered 350 children divided equally over 50 departments, rather than 329 children divided over 124 departments according to Figure 1, in the first setting of our simulation study. Throughout the simulation study, the lower- and higher-level sample sizes were kept constant at 350 and 50, respectively, in order to eliminate the known impact of changing sample sizes and depict the impact of the proportion of singletons as purely as possible. The proportion of singletons ranged from 0 to 95% (in steps of 5%, Figure 2). The number of singleton departments was rounded upwards (e.g. 5% singletons implies 2.5 departments containing only one child. Hence, for 5% singletons, we included 3 departments with one child). The remaining children were divided equally amongst the remaining departments. For the intracluster correlation we used a low, realistic and high intracluster correlation coefficient (ICC) which is defined in terms of the variance between departments (σ_B^2) and the variance within departments (σ_W^2) as:

$$ICC = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}$$

The realistic ICC (0.27) was obtained using the Ceftriaxone data. The low (0.15) and high (0.64) ICC were obtained after adjusting σ_W^2 to 1000 and 100, respectively.

2.1 Simulation procedure

For each of the 60 scenarios, 1000 datasets were simulated according to the following procedure:

1. Sample a random intercept from a normal distribution with mean zero and standard deviation σ_B for each of the 50 included departments.
   Note that we assume that the random intercepts follow a normal distribution while for real data they ~~are often~~ might be non-Gaussian. This however has been shown ~~not to impact~~ to have little or no impact on the stability of the linear mixed model (Verbeke and Lesaffre, 1997)

2. Assign a department size to the simulated departments according to the distribution of department sizes in the Ceftriaxone data (16 small, 18 medium and 16 large departments).
   Note that department size is unrelated to the number of children included in the department and is merely a characteristic of the department (referring to the total number of beds on the department).

3. Group the combination of age and reason for treatment for the 329 children in the Ceftriaxone Data based on the size of the department they are treated in. Then, conditional on the size of the simulated department, sample a combination of age and reason for treatment for 350 children.

4. Sample a residual error term from a normal distribution with mean zero and standard deviation σ_W for each of the 350 included children.
   Note that we assume that the residuals follow a normal distribution while for real data they ~~are often~~ might be non-Gaussian. This however has been shown not to impact the stability of the linear mixed model (Jacqmin-Gadda et al., 2007)

5. Simulate the prescribed dose for each child using Equation 1 and parameter estimates reported in Table 1.

2.2 Models fitted

All simulated datasets were analysed with the unmodified model. Because some of the simulated scenarios contain a fairly high proportion of singletons, one might doubt the need to correct for clustering. Therefore, we studied three different methods that are frequently used to handle singletons in the data. The first method is to simply ignore the dependence within units (i.e. ignoring clustering). This is done by fitting a model containing fixed effects for reason for treatment, age and department size, but no random effect. Other options are to discard the singletons from the data (i.e. dropping singletons) or to group the singletons into an artificial unit treated as one additional department (i.e. grouping singletons). Both approaches were evaluated by fitting the unmodified model to all simulated datasets either after dropping or after grouping the included singletons.

2.3 Analyses of simulated datasets

For each scenario, we assessed the performance of the fitted model using three performance characteristics. The first is the relative difference between the mean of the parameter estimates and the true parameters (mean estimated minus true parameter over true parameter, with true parameters reported in Table 1; RDM). The second characteristic is the relative difference between the mean estimated standard error and the empirical standard error (mean estimated minus empirical standard error over empirical standard error; RDE). Here, the estimated standard error reflects the uncertainty within the simulations while the empirical standard error (SES) reflects the variability between simulations. The first is calculated as the mean of the ~~obtained~~ estimated standard errors while the latter is calculated as the standard deviation of ~~obtained~~ parameter estimates when fitting the unmodified model. The last performance characteristic is the coverage of the confidence interval, calculated as the percentage of times the true parameter (reported in Table 1) falls within the estimated 95% Wald confidence interval. The stability of the F test was assessed by

comparing the number of times the null hypothesis was rejected (rejection rate) in the presence of different proportions of singletons to the rejection rate when no singletons were present.

3 Results

3.1 The unmodified model

All simulated datasets were analysed with the unmodified model (Equation 1). We report three performance characteristics for the multi-level model that were introduced in Section 2.3. These characteristics are shown for one fixed effect at the level of the child (Figure 3, solid lines) and one fixed effect at the level of the department (Figure 4, solid lines). Performance characteristics for the other fixed effects can be consulted in Figures A1 up to A5 (in Appendix, solid lines). Stability of the F test for parameters at the level of the child (*Age* and *Reason*) and at the level of the department (*Size*) is reported in Figure A6 (solid lines). Accuracy of the random effects variance and the residual variance is illustrated in Figure A7 (solid lines).

The differences between the estimated and the true parameter (RDM) for the fixed effects at the level of the child were not affected by the proportion of singletons or the intracluster correlation and were consistently small (Figures 3 and A1-A3: first row, solid lines). This indicates that the parameter is estimated well regardless of the proportion of singletons present in the data. The differences between the estimated and true standard error (RDE) were small throughout the simulation study, indicating that the standard error accurately estimated the true standard error for the unmodified model (Figures 3 and A1-A3: second row, solid lines). Coverage of the confidence intervals was around 95% throughout the simulation study (Figures 3 and A1-A3:third row, solid lines).

The RDM for the fixed effects at the level of the department were slightly higher than for a covariate at the level of the child, but fluctuated regardless of the proportion of singletons (Figures 4 and A4-A5: first row, solid lines). The RDE were small throughout the simulation study, indicating that the standard error accurately estimated the true standard error for the unmodified (Figures 4 and A4-A5: second row, solid lines). Coverage of the confidence intervals remained around 95% throughout the simulation study (Figures 4 and A4-A5: third row, solid lines).

The rejection rate for the F test for the effects at the level of the child increased slightly with an increasing proportion of singletons while the rejection rate for the effect at the level of the department decreased slightly with an increasing proportion of singletons (Figure A6: solid lines).

The RDM for both the random effects variance and the residual variance was small throughout the simulation study (Figure A7: solid lines, left and right, respectively). This indicates that generally, in the presence of singletons, the estimated variances approach the true variances quite well. The RDM was slightly higher when the intraclass correlation decreased.

3.2 Ignoring clustering and dropping or grouping singletons

Three different methods that are currently used to handle singletons in the data are compared to the unmodified model. To ignore clustering, all simulated datasets were analysed with a model containing a fixed effect for age, reason for treatment and department size. Additionally, the unmodified model was fitted to the datasets where singletons were removed (dropping singletons) or grouped into an artificial department (grouping singletons).
Obtained performance characteristics for one fixed effect at the level of the child and one fixed effect at the level of the department are visualized in Figures 3 and 4, respectively. Performance characteristics for

the other fixed effects can be consulted in Figures A1 up to A5. Stability of the F test for parameters at the level of the child (*Age* and *Reason*) and at the level of the department (*Size*) is reported in Figure A6. RDM for both residual and random effects variance are shown in Figure A7.

[Figure 2 about here.]

Figure 3 shows that the RDM and confidence interval coverage for the fixed effect at the level of the child ($Reason_{1ij}$) were comparable for the unmodified model fitted to the original data and the three options to handle the singletons (ignoring clustering and dropping or grouping singletons) regardless of the intracluster correlation. When clustering was ignored (dotted lines), the RDE was higher compared to the unmodified model fitted to the original data (full lines) or when dropping and grouping the singletons (dashed and dot-dashed lines, respectively). This seems to be counter-intuitive, but can be explained by the design of the simulation study where a combination of age and reason for treatment for a child were sampled conditional on the size of the simulated department (in step 4). Therefore, the observed doses were not only clustered within departments, but also stratified according to *Size*. When ignoring clustering, but accounting for stratification, the standard error estimates tend to be overestimated (Stepleton and Kang, 2016). The overestimation increased with increasing intracluster correlation, indicating that the importance of accounting for clustering increases with the homogeneity of observations within clusters.

[Figure 3 about here.]

Figure 4 shows that the RDM for the fixed effects at the level of the department were comparable for the unmodified model fitted to the original data and the three options to handle the singletons. Ignoring clustering (dotted lines) resulted in a decreased RDE and an unacceptably low confidence interval coverage for all proportions of singletons and all intracluster correlations under study. When dropping the singletons (dashed lines), the RDE increased with an increasing proportion of singletons. This increase was steeper for the scenario with a high intracluster correlation. The confidence interval coverage remained stable throughout the simulation study. For the scenario with 95% of singletons, the plots show a severe drop in both RDE and confidence interval coverage. When grouping the singletons into an artificial department (dot-dashed lines), RDE and confidence interval coverage decreased slightly with an increasing proportion of singletons, with this increase being again steeper for the scenario with a high intracluster correlation.
Figure A6 shows that dropping and grouping the singletons (dashed and dot-dashed lines, respectively) does not influence the performance of the F test for an effect at the level of the child. Ignoring the dependency within clusters (dotted lines) causes the rejection rate to be slightly lower compared to the rejection rate for the unmodified model fitted to the original data (solid lines).
Dropping and grouping the singletons (dashed and dot-dashed lines, respectively) cause the rejection rate for the fixed effect at the level of the department to be respectively lower and higher compared to the rejection rate for the unmodified model fitted to the original data (solid lines). Ignoring the dependency within clusters (dotted lines) causes the rejection rate to be a lot higher than the rejection rate for the unmodified model fitted to the original data (solid lines). The rejection rates increased with an increasing ICC, which can be explained by the decrease in variance within departments, which increases rejection rates.
Figure A7 shows that when dropping the singletons (dashed lines), the residual variance stayed close to the true residual variance regardless of the intracluster correlation. The random effects variance was close to the true random effects variance throughout the simulation study, but decreased steeply at the end (for the scenario with 95% singletons).
When grouping the singletons (dot-dashed lines), the residual variance was slightly overestimated while the random effects variance was slightly underestimated, with the difference between estimated and true variance getting bigger with an increasing proportion of singletons and with an increase in intracluster correlation.

4 Discussion

We conducted a simulation study, inspired by the structure of the Ceftriaxone data, to investigate the impact of an increasing proportion of singletons ($0 - 95\%$) combined with different intracluster correlations (low-medium-high) on different aspects of the linear multi-level model. Note that this simulation study includes rather extreme situations (up to 95% of singletons), ~~and that although these situations rarely occur in practice, they were~~ included to demonstrate the effect of increasing the proportion of singletons as clearly as possible. Note also that throughout this simulation study both the lower- and the higher-level sample sizes were kept constant, in order to illustrate the pure impact of the proportion of singletons, which was the focus of this study ~~, from the impact of changing sample sizes~~. The impact of the presence of singletons was assessed through three performance characteristics, which revealed that neither the RDM nor the RDE were affected by the proportion of singletons in the data or by the increase in intracluster correlation. They were consistently low, with RDM and RDE for an effect at the level of the child being slightly lower than RDM and RDE for an effect at the level of the department. This might be explained by the number of independent observations that are available to estimate both effects, with this number being considerably lower for the effect at the level of the department. The confidence interval coverage and rejection rates fluctuated slightly while the proportion of singletons changed, with this fluctuation being unrelated to the proportion of singletons in the data. Confidence interval coverage stayed stable while increasing the intracluster correlation, while the F test rejection rates increased with an increasing intracluster correlation. This can be explained by the variance within departments which is decreased in order to increase the intracluster correlation and increased the rejection rates. The rejection rate for the effect at the level of the department decreased slightly with an increasing proportion of singletons, which can be explained by the more stable estimation of the average dose for a department when the number of children in that department is larger.
These findings verify the conclusions reached by (Pickering and Weatherall, 2007), and hence increases confidence to use a linear mixed model, even when the proportion of singletons increases above 15%. While Sauzet *et al.* mention biased parameter estimates for fixed effects in the presence of extreme proportions of singletons, which were not found in the present study. One possible explanation is that Sauzet *et al.* considered only singletons and clusters of size two, while we considered singletons together with clusters of medium to large size ($7-151$). Another possible explanation is that Sauzet *et al.* varied both the sample size ($152-1200$) and the number of clusters ($150-1000$), while in the simulation study presented here, both were kept constant in order to filter out the impact of an increasing proportion of singletons as clearly as possible.
In an additional exploratory simulation study, we varied the total sample size ($62 - 350$) according to the proportion of singletons, while keeping the number of departments (50) and department size (1 or 7) fixed, and varied the number of departments ($50 - 290$) according to the proportion of singletons while keeping the total sample size (350) and department size (1 or 7) fixed. Results from this study (not shown here) verify that a decrease in total sample size goes together with a decreased rejection rate (mainly at the level of the child), and that a decrease in the number of departments goes together with a decreased rejection rate (mainly at the level of the department). Therefore, the simulation setting presented in this paper, which varies department size according to the proportion of singletons but keeps both the number of departments and total sample size constant, is the setting which presents the impact of a changing proportion of singletons most accurately.

Three different methods that are currently used to handle singletons in the data were compared to the unmodified model (ignoring clustering, dropping singletons and grouping singletons). A simulation study was conducted to investigate the consequences of these three options on different aspects of the multi-level model. Impact on the level of the child was minor, while impact on the level of the department was more clear. As mentioned before, this can be explained by the number of independent observations available.
When ignoring clustering, the RDM did not change notably. The RDE at the level of the child increased, while the rejection rate for the F test decreased in comparison to the unmodified model fitted to the original

data. This can be explained by the design of the simulation study where the observed doses were not only clustered within departments, but also stratified according to *Size*. When ignoring clustering, but accounting for stratification, the standard error estimates tend to be overestimated (Stepleton and Kang, 2016). The overestimation increased with increasing intracluster correlation, indicating that the importance of accounting for clustering increases with the homogeneity of observations within clusters. Confidence interval coverage remained stable. The RDE at the level of the department decreased, while the rejection rate for the F test increased in comparison to the unmodified model fitted to the original data. This can be explained by the consistent underestimation of the standard error when ignoring clustering (Verbeke and Molenberghs, 2009). The rejection rates increased with an increasing intracluster correlation, which can be explained by the decrease in variance within departments, which increases rejection rates. Confidence interval coverage was unacceptably low for all proportions of singletons and ICCs, indicating that ignoring the dependency within the clusters is never a good idea.
When dropping the singletons, the RDE was higher while the rejection rate for the F test was lower compared to the unmodified model fitted to the original data, with this difference increasing when the proportion of singletons increases or when the intracluster correlation increases. This can be explained by the increase in standard error due to the decrease in number of remaining departments, which is steeper when intracluster correlation is high. For the scenario with 95% singletons (Figure 4), there is a severe drop in RDE and confidence interval coverage that goes together with a steep increase in F test rejection rate, which is explained by the presence of only one large department in this scenario. The low confidence interval coverage resulting from dropping the singletons forces us to conclude that it worsens the performance of the multi-level model.
When grouping the singletons into an artificial department, the RDE was lower while the rejection rate was higher compared to the unmodified model fitted to the original data, with this difference again enlarged by an increase in intracluster correlation. The residual variance was slightly overestimated while the random effects variance was slightly underestimated, with the difference between estimated and true variance increasing with an increasing proportion of singletons and an increasing intracluster correlation. All these findings can be explained by the grouping of singletons that are not actually related, which decreases the variance between included departments and causes a slight underestimation of the true standard error for the effect at the level of the department. Although grouping is an option that might be considered when the data at hand contain a high proportion of singletons, the regular multi-level model performs better even when the proportion of singletons is large.
An alternative that could be used in the presence of sparseness at the lowest level of multi-level hierarchy is to select a more convenient clustering level (e.g. country or hospital in which the department is situated) (Cortiñas Abrahantes et al., 2004). Although this strategy would improve the model's stability, it was not considered here because we focussed on a two-level setting, where clustering by department is the only option.

5 Conclusion

The linear mixed model appears to be stable enough to handle high proportions of singletons when both lower- and higher level sample sizes remain fixed, also when the intracluster correlation is high. Alternatives which are frequently used, such as ignoring clustering or removing the singletons, should be avoided as they provide biased standard error estimates for the fixed effects. Although grouping the singletons is an option, the regular multi-level model performs better. ~~Therefore, we can be confident in using the linear mixed model, even when the data at hand contain high proportions of singletons.~~

Acknowledgements

We would like to thank Ann Versporten for providing the Ceftriaxone data. Support from the Methusalem financement program of the Flemish Government is gratefully acknowledged. NH acknowledges support from the University of Antwerp scientific chair in Evidence-Based Vaccinology, financed in $2009 - 2016$ by a gift from Pfizer and in 2016 from GSK.

Conflicts of interest

The authors have no conflicts of interest to declare.

Appendix

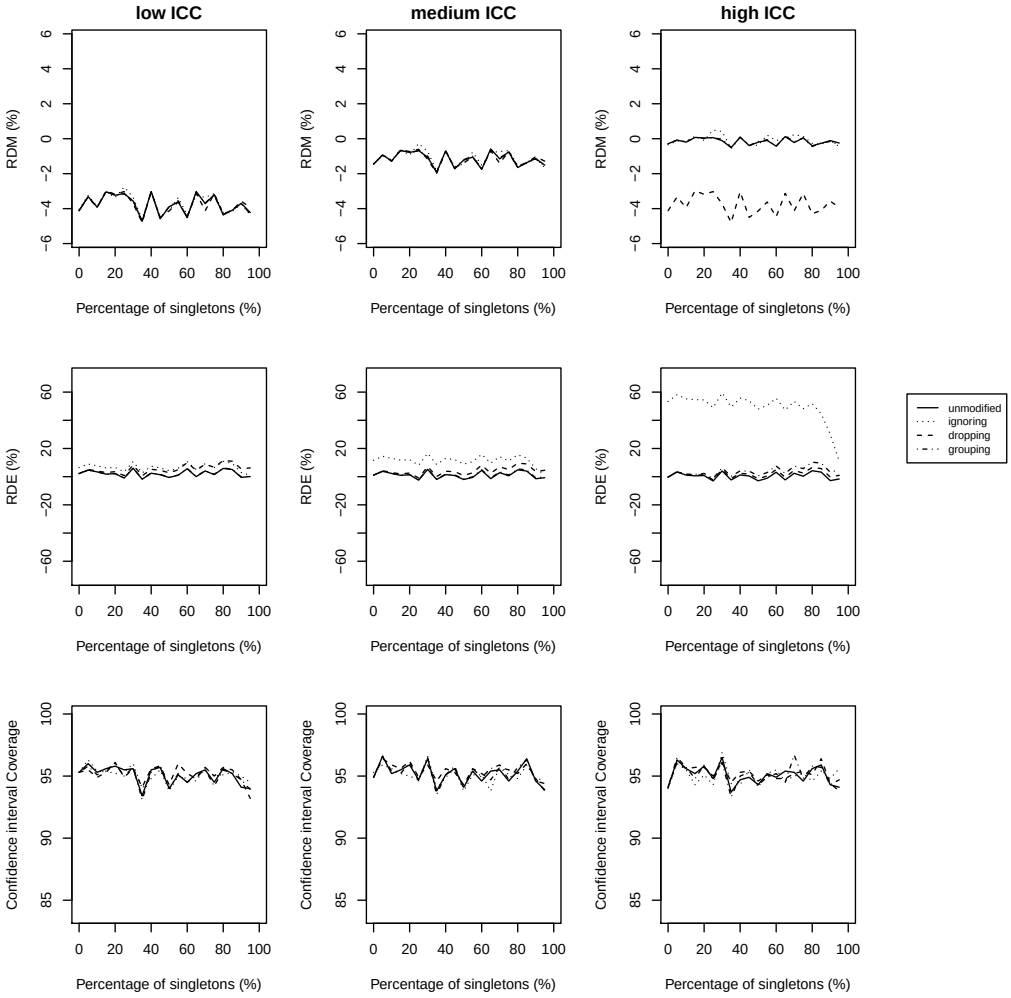

Figure A1 Performance characteristics for the fixed effect Age_{ij} in the unmodified model, when ignoring clustering and when dropping and grouping the singletons with an increasing intracluster correlation (ICC) and an increasing proportion of singletons; RDM: relative difference between estimated and true mean; RDE: relative difference between estimated and true standard error.

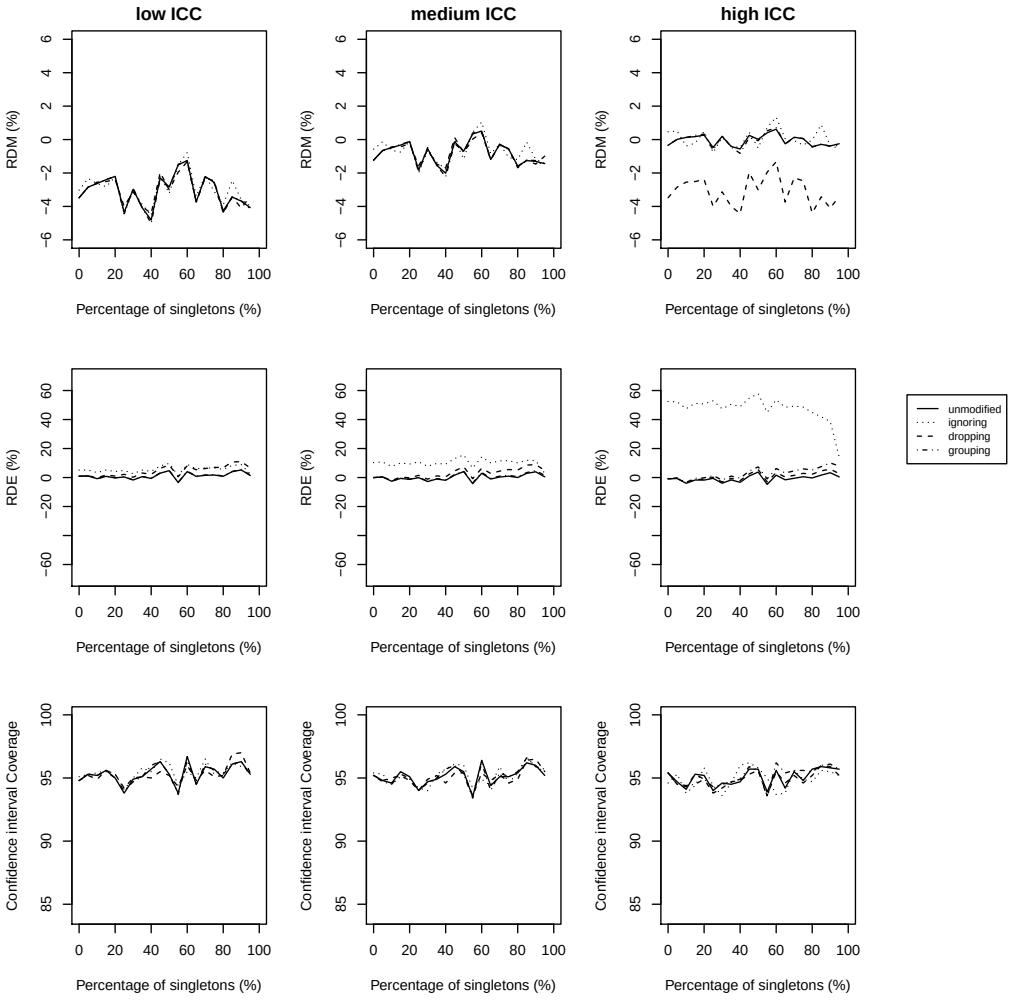

Figure A2 Performance characteristics for the fixed effect $Reason_{2ij}$ in the unmodified model, when ignoring clustering and when dropping and grouping the singletons with an increasing intracluster correlation (ICC) and an increasing proportion of singletons; RDM: relative difference between estimated and true mean; RDE: relative difference between estimated and true standard error.

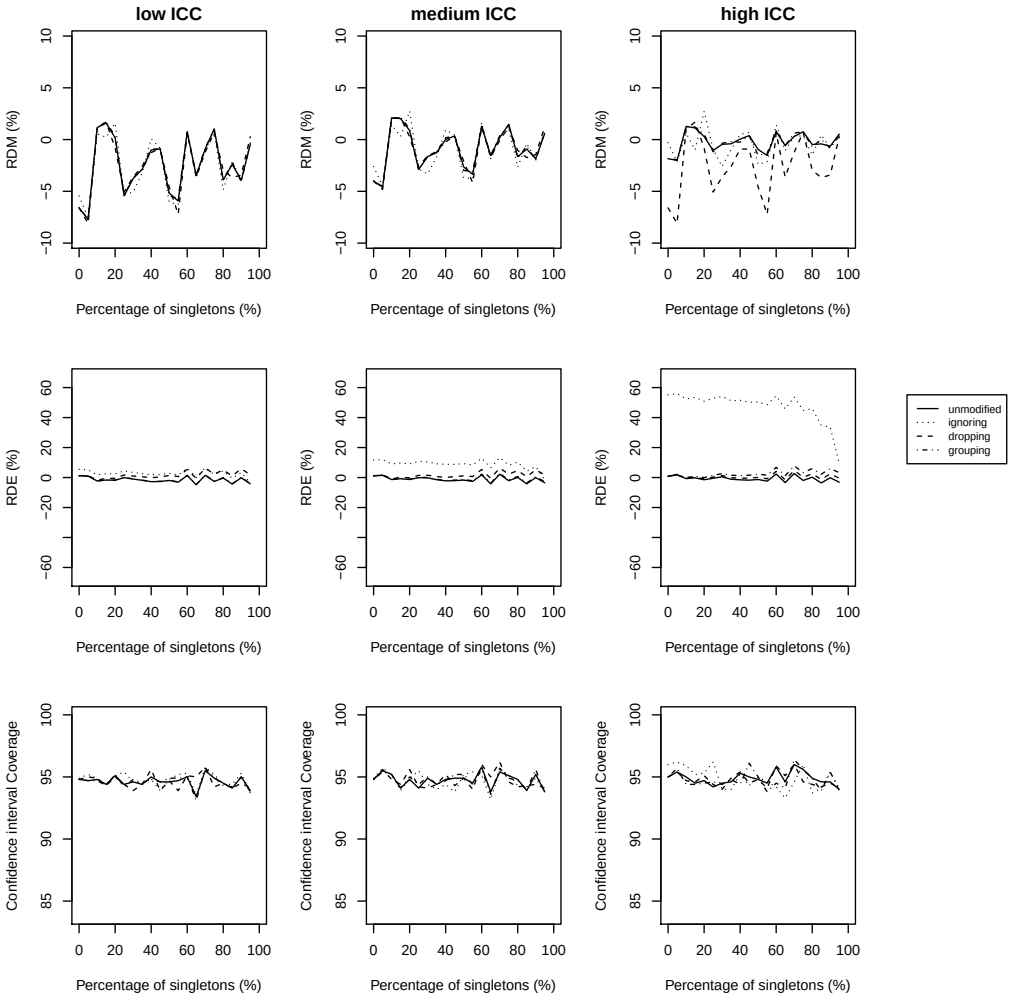

Figure A3 Performance characteristics for the fixed effect $Reason_{3ij}$ in the unmodified model, when ignoring clustering and when dropping and grouping the singletons with an increasing intracluster correlation (ICC) and an increasing proportion of singletons; RDM: relative difference between estimated and true mean; RDE: relative difference between estimated and true standard error.

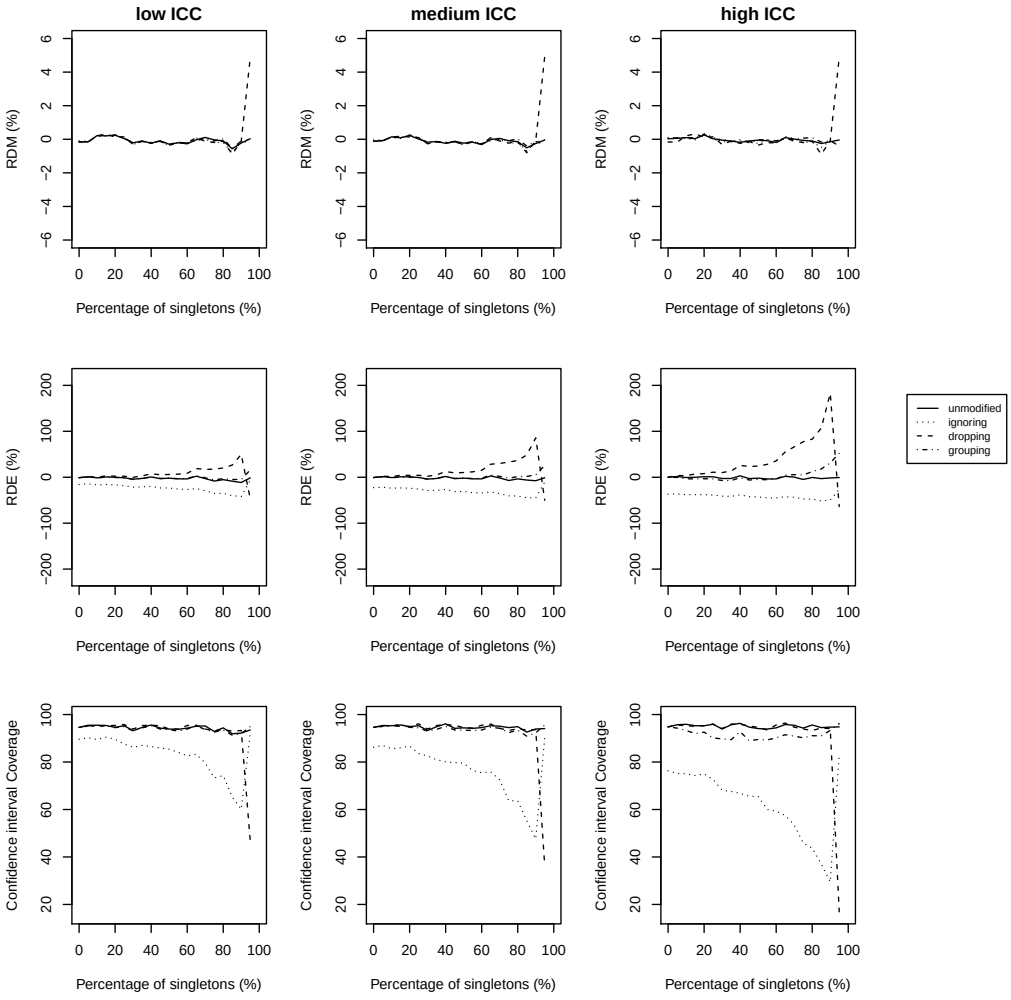

Figure A4 Performance characteristics for the fixed intercept in the unmodified model, when ignoring clustering and when dropping and grouping the singletons, with an increasing intracluster correlation (ICC) and an increasing proportion of singletons; RDM: relative difference between estimated and true mean; RDE: relative difference between estimated and true standard error.

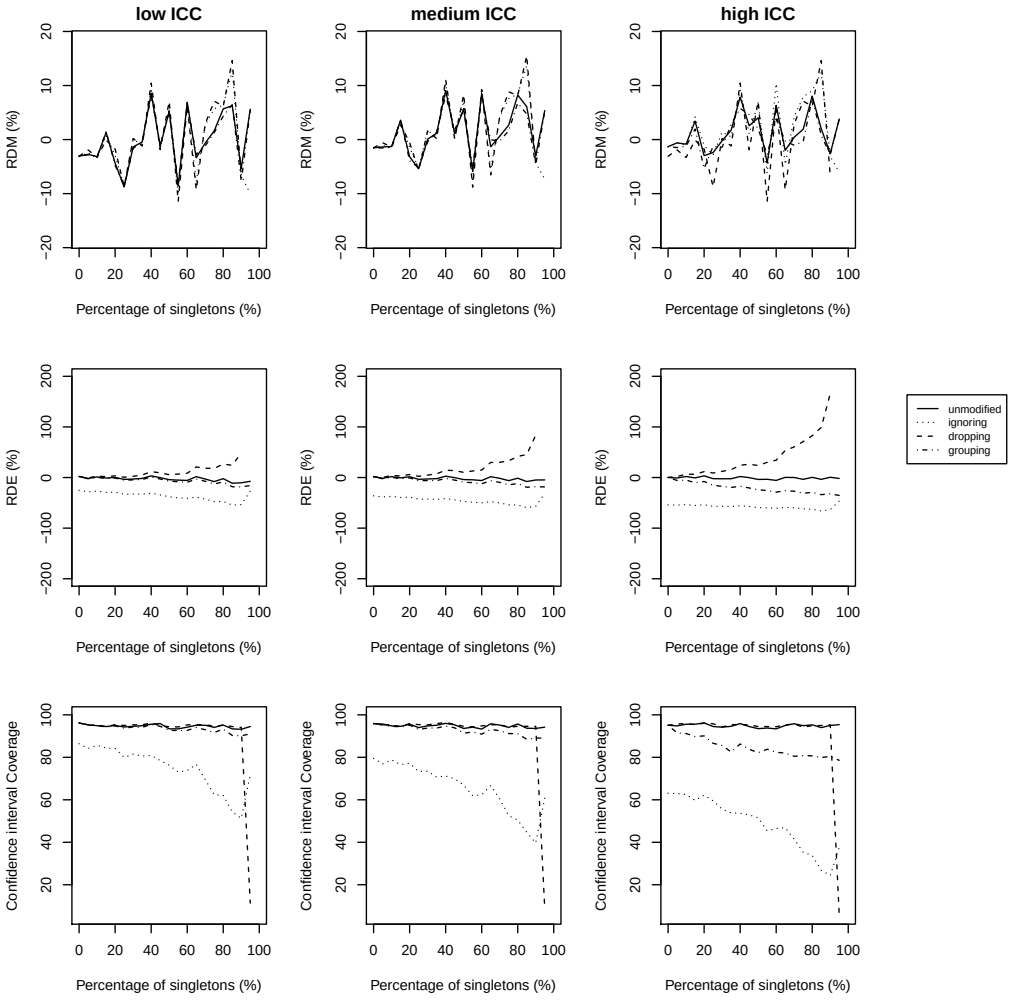

Figure A5 Performance characteristics for the fixed effect $Size_{2j}$ in the unmodified model, when ignoring clustering and when dropping and grouping the singletons, with an increasing intracluster correlation (ICC) and an increasing proportion of singletons; RDM: relative difference between estimated and true mean; RDE: relative difference between estimated and true standard error.

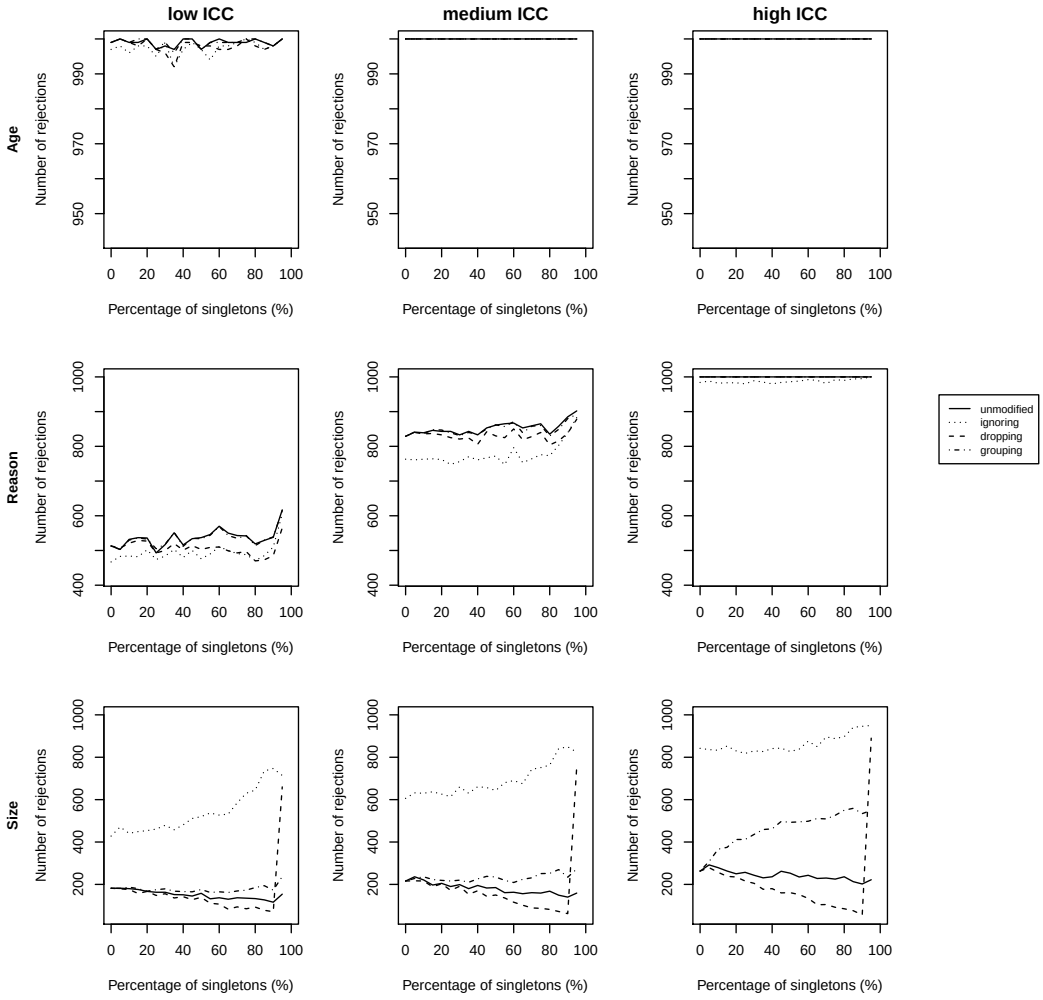

Figure A6 Rejection rate of the F test for the fixed effects in the unmodified model, when ignoring clustering and when dropping and grouping the singletons, with an increasing intracluster correlation coefficient (ICC) and an increasing proportion of singletons; RDM: relative difference between estimated and true mean; RDE: relative difference between estimated and true standard error.

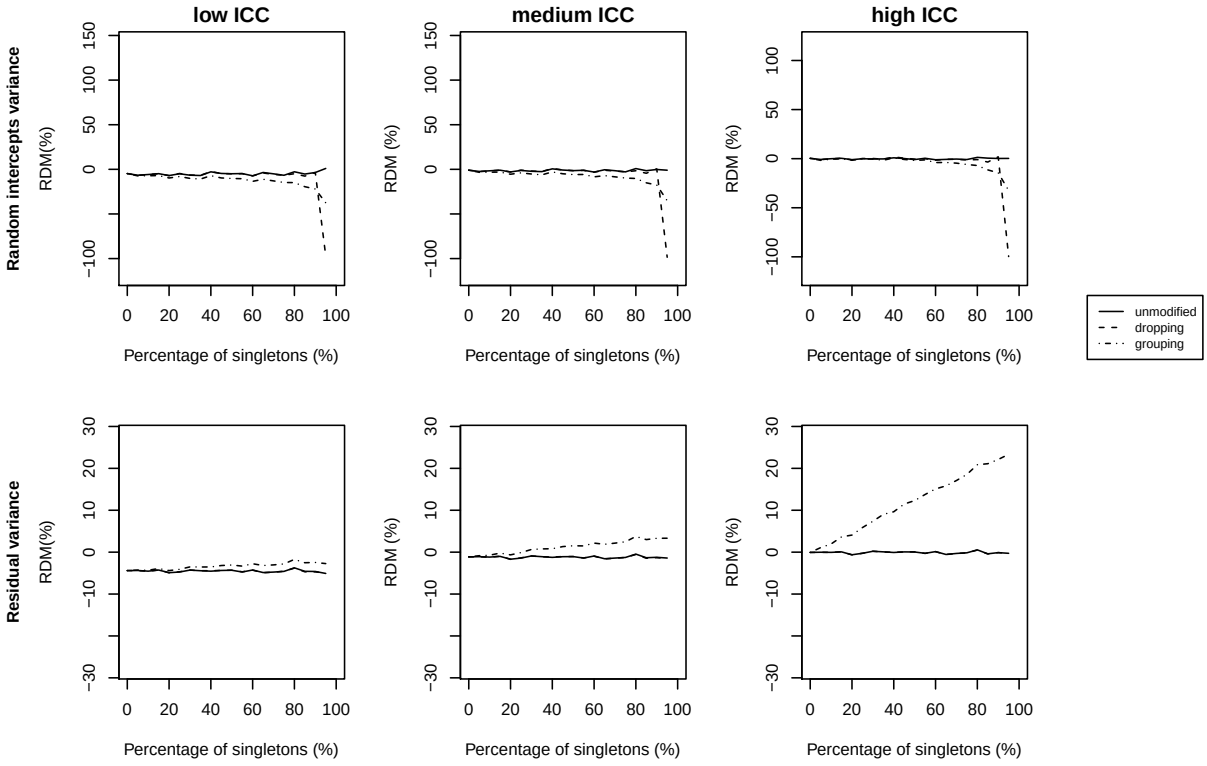

Figure A7 Relative difference between estimated and true mean (RDM) for the random effects variance (top) and residual variance (bottom) in the unmodified model, when ignoring clustering and when dropping and grouping the singletons, with an increasing intracluster correlation coefficient (ICC) and an increasing proportion of singletons.

References

Bell, B. A., Morgan, G. B., Schoeneberger, J. A., Kromrey, J. D., and Ferron, J. M. (2014). How low can you go? an investigation of the influence of sample size and model complexity on point and interval estimates in two-level linear models. *Methodology*, 10:1–11.

Bruyndonckx, R., Hens, N., and Aerts, M. (2016). Simulation-based evaluation of the performance of the f test in a linear multilevel model setting with sparseness at the level of the primary unit. *Biometrical journal*, 58(5):1054–70.

Clarke, P. (2008). When can group level clustering be ignored? multilevel models versus single-level models with sparse data. *Journal of epidemiology & community health*, 62:752–758.

Cortiñas Abrahantes, J., Molenberghs, G., Burzykowski, T., Shkedy, Z., and Renard, D. (2004). Choice of units of analysis and modeling strategies in multilevel hierarchical models. *Computational statistics and data analysis*, 47:537–563.

Garson, D. G. (2013). *Hierarchical linear modeling: Guide and appliations*. Sage publications, London.

Goldstein, H. (2003). *Multilevel statistical models*. John Wiley & Sons, Chichester, 4th edition.

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D., and Thomas, S. (1993). A multilevel analysis of school examination results. *Oxford review of education*, 19(4):425–433.

Hox, J. (1998). Multilevel modeling: when and why. In Balderjahn, I., Mathar, R., and Schader, M., editors, *In: Classification, data analysis and data highways*, pages 147–154. Springer Verlag, New York.

Hox, J. (2010). *Multilevel analysis: techniques and applications*. Routledge, New York, 2nd edition.

Jacqmin-Gadda, H., Sbillot, S., Proust, C., Molina, J. M., and Thiébaut, R. (2007). Robustness of the linear mixed model to misspecified error distributions. *Computational statistics and data analysis*, 51:5142–5154.

Kreft, I. and De Leeuw, J. (1998). *Introducing multilevel modeling*. Sage publications, London.

Lee, V. E. (2000). Using hierarchical linear modeling to study social contexts: the case of school effects. *Educational psychologist*, 35(2):125–141.

Littell, R., Milliken, G., Stroup, W., Wolfinger, R., and Schabenberger, O. (2006). *SAS ® for mixed models*. SAS Institute Inc., Cary, 2nd edition.

Maas, C. and Hox, J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, 1:86–92.

Moulton, B. R. (1986). Random group effects and the precision of regression estimates. *Journal of econometrics*, 32(3):385–397.

Pickering, R. and Weatherall, M. (2007). The analysis of continuous outcomes in multi-centre trials with small centre sizes. *Statistics in medicine*, 26:5445–5456.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical linear models: applications and data anlysis methods*. Sage publications, Thousand Oaks, 2nd edition.

Renard, D., Molenberghs, G., Van Oyen, H., and Tafforeau, J. (1998). Investigation of the clustering effect in the belgian health interview survey 1997. *Archives of public health*, 56:345–361.

Sauzet, O., Wright, K. C., Marston, L., Brocklehurst, P., and Peacock, J. L. (2012). Modelling the hierarchical structure in datasets with very small clusters: a simulation study to explore the effect of the proportion of clusters when the outcome is continuous. *Statistics in medicine*, 32:1429–1438.

Singer, J. (1998). Using sas proc mixed to fit multilevel models, hierarchical models, and individual growth models. *Journal of educational and behavioral statistics*, 24(4):323–355.

Snijders, T. A. B. and Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage Publications, London.

Stepleton, L. M. and Kang, Y. (2016). Design effects of multilevel estimates from national probability samples. *Sociological methods & research*, pages published online before print, doi: doi: 10.1177/0049124116630563.

Verbeke, G. and Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational statistics and data analysis*, 23:541–556.

Verbeke, G. and Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. Springer Verlag, New York.

Versporten, A., Sherland, M., Bielicki, J., Drapier, N., Vankerckhoven, V., and Goossens, H. (2013). The antibiotic resistance and prescribing in european children project: a neonatal and pediatric antimicrobial web-based point prevalence survey in 73 hospitals worldwide. *Pediatric infectious disease journal*, 32:e242–53.

Wang, J., Xie, H., and Fisher, J. H. (2012). *Multilevel models: applications using SAS ®*. De Gruyter, Berlin.

Figures

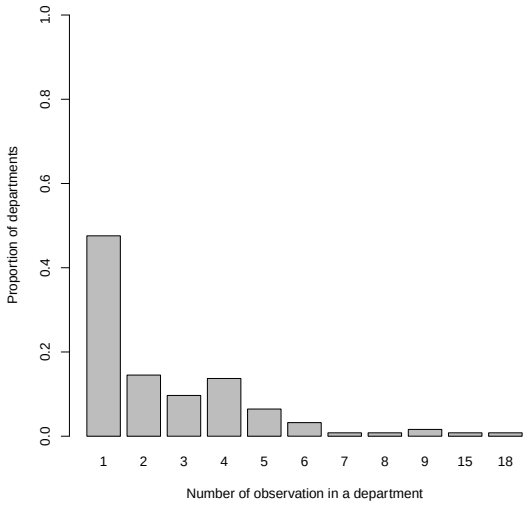

Figure 1 Barplot representing the distribution of the number of children within the 124 departments included in the Ceftriaxone data.

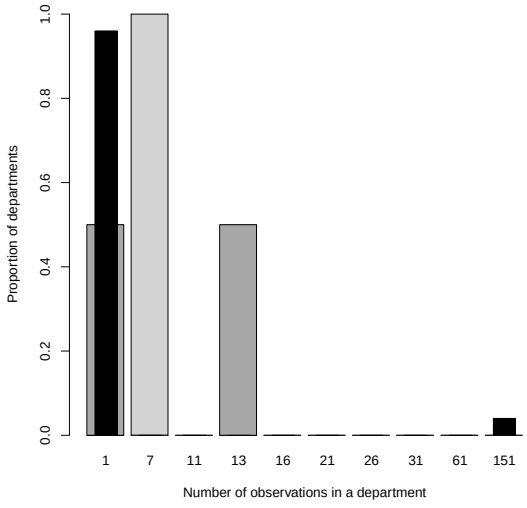

Figure 2 Barplot representing the distribution department sizes for the 50 simulated departments in the presence of 0% (light grey), 50% (dark grey) and 95% (black) singletons.

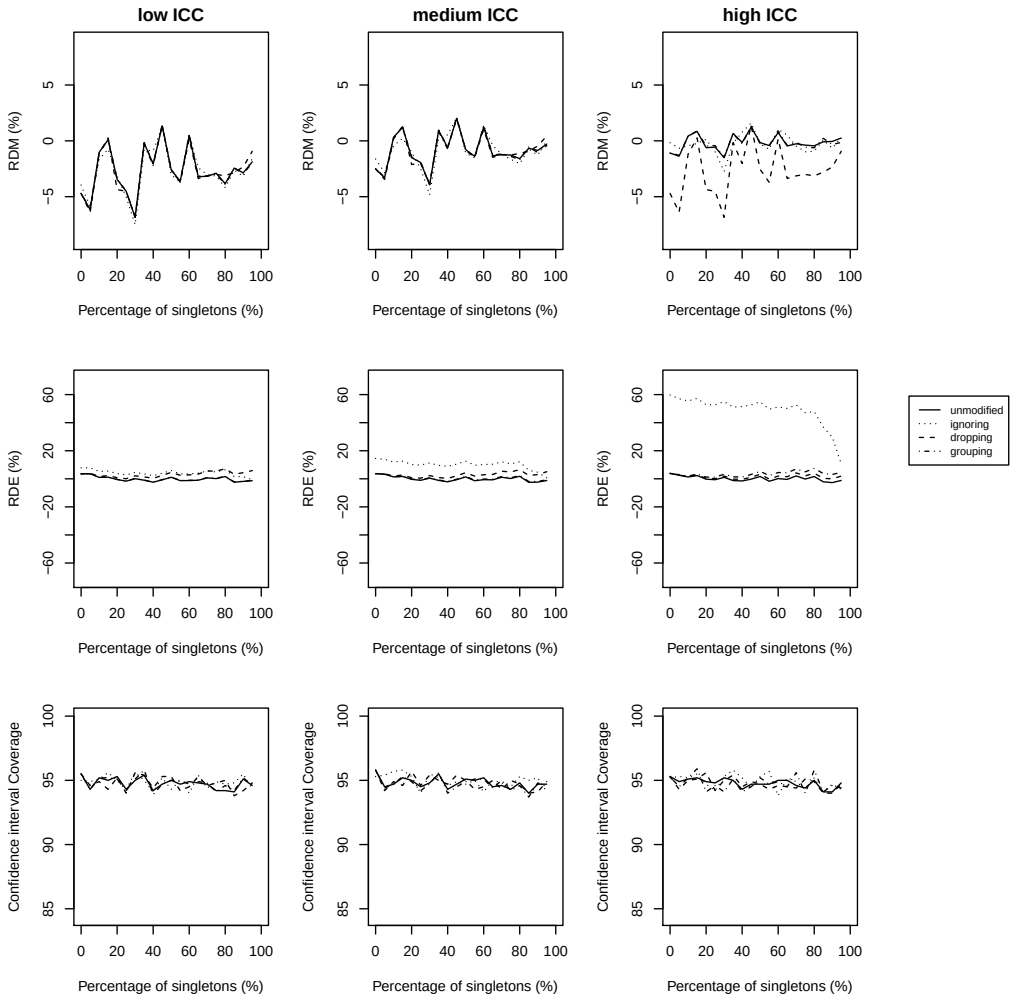

Figure 3 Performance characteristics for the fixed effect $Reason_{1ij}$ in the unmodified model, when ignoring clustering and when dropping and grouping the singletons with an increasing intracluster correlation coefficient (ICC) and an increasing proportion of singletons; RDM: relative difference between estimated and true mean; RDE: relative difference between estimated and true standard error.

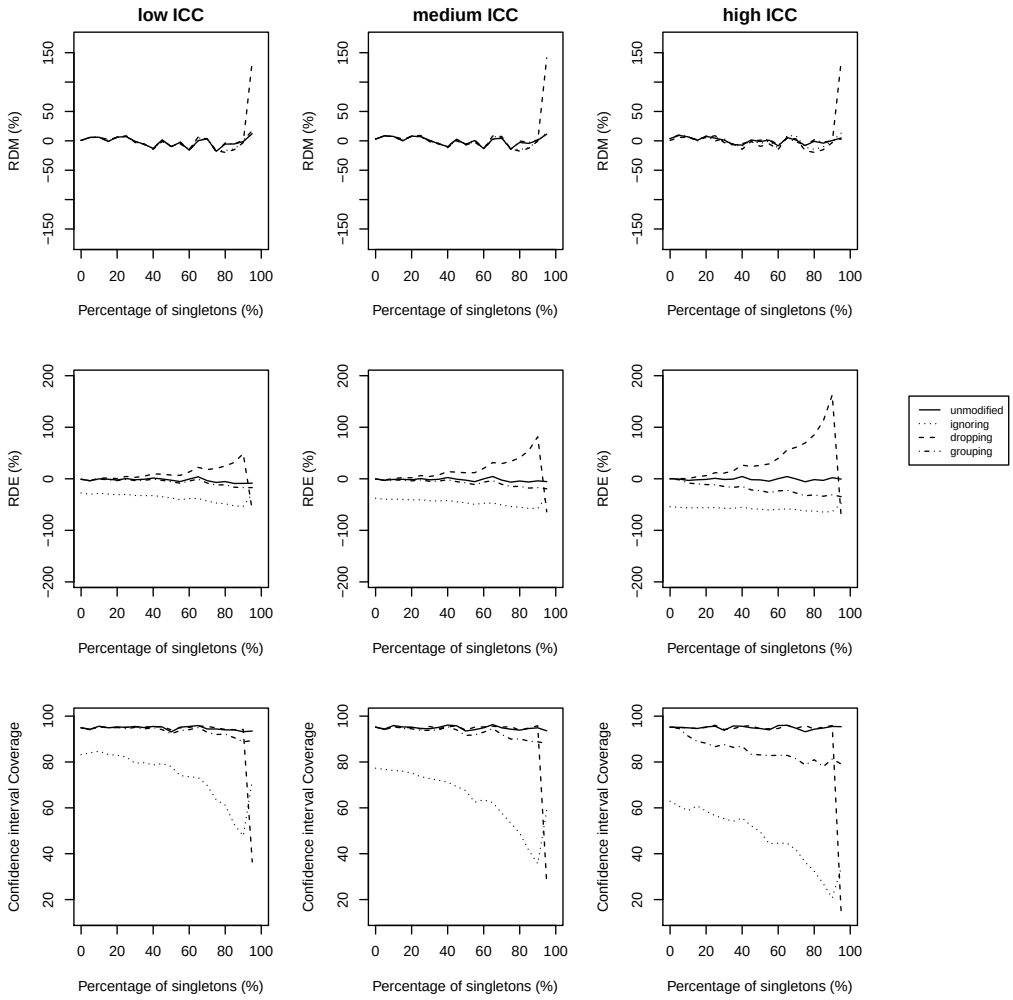

Figure 4 Performance characteristics for the fixed effect $Size_{1j}$ in the unmodified model, when ignoring clustering and when dropping and grouping the singletons, with an increasing intracluster correlation coefficient (ICC) and an increasing proportion of singletons; RDM: relative difference between estimated and true mean; RDE: relative difference between estimated and true standard error.

Tables

Table 1 Parameter estimates and standard errors obtained by fitting the unmodified model to the Ceftriaxone data.

Parameter	Estimate	Std. error
Intercept	82.951	4.513
Size_{1j}	-3.224	4.771
Size_{2j}	4.469	4.747
Age_{ij}	-2.060	0.338
Reason_{1ij}	-8.254	4.319
Reason_{2ij}	-21.995	7.432
Reason_{3ij}	-5.586	3.574
σ_B^2	180.370	47.139
σ_W^2	489.510	44.756

