Analysis of Randomised Trials Including Multiple Births when Birth Size is Informative

Lisa N Yelland[a,b], Thomas R Sullivan[b], Menelaos Pavlou[c], Shaun R Seaman[d]

[a]Women's and Children's Health Research Institute, North Adelaide, South Australia, Australia

[b]School of Population Health, The University of Adelaide, Adelaide, South Australia, Australia

[c]Department of Statistical Science, University College London, London, United Kingdom

[d]MRC Biostatistics Unit, Cambridge, United Kingdom

Correspondence: Dr Lisa Yelland, School of Population Health, Mail Drop DX650 511, The University of Adelaide, SA 5005, Australia, Email: lisa.yelland@adelaide.edu.au, Phone: +61 8 8313 3215, Fax: +61 8 8223 4075.

Word count: 3496

Abstract

Background: Informative birth size occurs when the average outcome depends on the number of infants per birth. Although analysis methods have been proposed for handling informative birth size, their performance is not well understood. Our aim was to evaluate the performance of these methods and to provide recommendations for their application in randomised trials including infants from single and multiple births.

Methods: Three generalised estimating equation (GEE) approaches were considered for estimating the effect of treatment on a continuous or binary outcome: cluster weighted GEEs, which produce treatment effects with a mother-level interpretation when birth size is informative; standard GEEs with an independence working correlation structure, which produce treatment effects with an infant-level interpretation when birth size is informative; and standard GEEs with an exchangeable working correlation structure, which do not account for informative birth size. The methods were compared through simulation and analysis of an example dataset.

Results: Treatment effect estimates were affected by informative birth size in the simulation study when the effect of treatment in singletons differed from that in multiples (i.e. in the presence of a treatment group by multiple birth interaction). The strength of evidence supporting the effectiveness of treatment varied between methods in the example dataset.

Conclusions: Informative birth size is always a possibility in randomised trials including infants from both single and multiple births, and analysis methods should be pre-specified with this in mind. We recommend estimating treatment effects using standard GEEs with an independence working correlation structure to give an infant-level interpretation.

Keywords: informative cluster size, multiple births, statistical methodology, clustering, generalised estimating equations.

Many neonatal and perinatal trials include infants from both single and multiple births,[1,2] which makes the statistical analysis challenging. Whereas outcomes of infants born to different mothers can usually be considered independent, outcomes of infants from the same birth are likely to be similar due to shared genetic and environmental factors.[3,4] Multiple births therefore create clustering in the data, where the mother is the cluster and her infant(s) are the cluster member(s).[3]

Methods for analysing clustered data are widely available and their performance has been investigated in studies including infants from both single and multiple births.[1,3-8] It is now well established that clustering due to multiple births should be taken into account in the analysis,[1,3,7-9] especially when the multiple birth rate is not low.[4,5] Failure to account for clustering due to multiple births can increase the chance that an ineffective treatment is found to be effective,[7] which could lead to inappropriate recommendations for clinical practice. Generalised estimating equations (GEEs)[10] are the most popular analysis approach for handling clustering due to multiple births.[1,2]

Informative cluster size (ICS) is a common problem in clustered data. It occurs when the outcome of interest is related to the size of the cluster, conditional on the covariates in the analysis model.[11] For randomised trials including infants from both single and multiple births, the cluster size is the birth size (i.e. the number of infants per birth) and ICS is likely to arise in two main ways. Firstly, the average outcome may differ between singletons and multiples. For example, multiples have lower average birthweights[12] and increased risk of mortality and cerebral palsy.[13] Secondly, the average effect of the intervention may differ between singletons and multiples. For instance, antenatal corticosteroid therapy for preventing respiratory distress syndrome in preterm infants may be more effective in singletons than in twins.[12]

When ICS is present, GEEs do not necessarily estimate the treatment effect of interest.[14] Failure to account for ICS could therefore lead to biased treatment effect estimates and incorrect conclusions regarding the effectiveness of treatment. Analysis methods based on GEEs have been suggested for handling ICS,[15,16] and these have been used to account for informative birth size.[1,17] However, their performance has not been formally investigated in this setting and it is unclear when these methods should be applied. The aims of this article are to (1) study the performance of GEE methods for handling ICS through simulation and analysis of an example dataset; and (2) provide recommendations on the application of these methods in randomised trials including infants from both single and multiple births.

Methods

Statistical Methods

Three GEE methods were considered for estimating the marginal effect of treatment (i.e. the average treatment effect across all infants) on a continuous or binary outcome. Firstly, the cluster weighted GEE (CWGEE) approach is a GEE with an independence working correlation structure and weights equal to the inverse of the birth size (i.e. weight 1 for singletons, 1/2 for twins and so on).[15,16] This method gives equal weight to each mother in the analysis, irrespective of the birth size. When ICS is present, CWGEE estimates treatment effects for a randomly selected infant from a randomly selected mother, thus providing a mother-level interpretation.[14,15,18] Secondly, the GEE independence (GEE$_{\text{ind}}$) approach is a standard (unweighted) GEE with an independence working correlation structure. This method gives each infant equal weight in the analysis. When ICS is present, GEE$_{\text{ind}}$ estimates treatment effects for a randomly selected infant, irrespective of which mother they belong to, and thus provides an infant-level interpretation.[14,15,18] Thirdly, the GEE exchangeable (GEE$_{\text{exch}}$) approach is a standard GEE with an exchangeable working correlation structure, which assumes that the outcomes of all infants from the same birth are equally correlated. This method weights infants in a way that minimises the variance of the treatment effect estimate.[19] Although GEE$_{\text{exch}}$ is not recommended in the presence of ICS as it does not necessarily estimate a treatment parameter of interest,[14] we include it here to investigate what can go wrong with it when ICS is present. All methods were implemented using the GENMOD procedure with empirical sandwich variance estimation in SAS version 9.3 (Cary, NC, USA).

Simulation Study

A simulation study was conducted to evaluate the performance of CWGEE, GEE$_{\text{ind}}$ and GEE$_{\text{exch}}$ when ICS is present. Simulation scenarios were chosen based on an example dataset (described below) and 10,000 datasets were generated for analysis in each scenario. Mothers were randomised to the intervention or control group (300 per group), and independently assigned to have a single birth with 80% probability, or a twin birth with 20% probability; higher order multiples are rare in practice and were not considered. This produced an expected total sample size of 720 infants with 33.3% from a multiple birth, which is typical of preterm populations.[2] Outcomes of infants from the same birth were positively correlated, with an intracluster correlation coefficient (ICC) of 0.1, 0.5 or 0.9. Additional simulations were performed for an ICC of 0.5 while varying the probability of a twin birth from 5% to 95% by 5%.

Continuous outcomes were randomly generated from the model

$$Y_{ij} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + a_i + e_{ij}\,, \tag{1}$$

where Y_{ij} is the outcome for the j th infant from the i th mother, X_{1i} is the randomised treatment group (1=intervention, 0=control), X_{2i} is the multiple birth status (1=multiple birth, 0=single birth), a_i is a random mother effect drawn from an $N\left(0,\sigma_a^2\right)$ distribution, and e_{ij} is a random error drawn from an $N\left(0,\sigma_e^2\right)$ distribution. Under model (1), ICS occurs whenever the outcome and/or the effect of treatment on the outcome depends on birth size (i.e. whenever β_2 and/or β_3 are nonzero); more general definitions of ICS are discussed elsewhere.[14,18,20] Variances were chosen to give a total variance of $\sigma_a^2 + \sigma_e^2 = 15^2$ and produce the desired ICC according to the equation $\text{ICC} = \sigma_a^2 \big/ \left(\sigma_a^2 + \sigma_e^2\right)$. The mean outcome for singletons in the control group was set to 100 ($\beta_0 = 100$), since outcomes of many developmental assessments follow an $N\left(100,15^2\right)$ distribution in the population, such as the Mental Development Index (MDI) standardised score from the Bayley Scales of Infant Development.[21] The mean outcome for singletons in the intervention group was chosen to be 104 to produce a treatment effect of 4 among singletons ($\beta_1 = 4$), since the example trial was designed to detect a 4-point improvement in the MDI. The mean outcome for twins in the control group was set to 97 ($\beta_2 = -3$), since a 3 point reduction in the mean developmental outcome for twins compared with singletons is plausible based on the example dataset. The mean outcome for twins in the intervention group was chosen to be either 101, 99 or 103 ($\beta_3 = 0$, -2 or 2) to produce a treatment effect among twins of 4 (to match the singletons), 2 or 6, respectively, in order to explore the effect of ICS in the absence or presence of a treatment group by multiple birth interaction. Simulation methods for binary outcomes are described in the Online Supplement.

Treatment effects were estimated for each simulated dataset based on the unadjusted model $\mu_{ij} = \beta_0 + \beta_1 X_{1i}$, and the model adjusting for multiple birth status as a main effect $\mu_{ij} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$, where $\mu_{ij} = E\left[Y_{ij}\right]$ is the mean outcome, since both unadjusted and adjusted estimates are commonly presented. An interaction model $\mu_{ij} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i}$ was used to test for evidence of a treatment group by multiple birth interaction but not to estimate treatment effects, since these should primarily be estimated from main effects models.[22] Simulation results were summarised for each scenario by averaging the treatment effect (β_1) estimates and their estimated standard errors. Monte Carlo (simulation) standard deviations were very similar to the average estimated standard errors and are not reported. The power to detect a treatment group by multiple birth interaction was calculated as the percentage of simulated datasets where the interaction term was statistically significant (p<0.05).

Example Dataset

To illustrate the impact of choosing different GEE methods in a real trial where ICS may be present, we consider a trial of high-dose versus standard-dose docosahexaenoic acid (DHA), a source of omega-3 fatty acids, for preterm infants.[23] Consenting mothers of infants born less than 33 weeks' gestation were randomly assigned to receive capsules rich in DHA or placebo capsules and infants received the treatment through breast milk. There were 545 mothers and 657 infants included in the trial, of whom 33.6% were from a multiple birth (30.4% in the high-DHA group, 36.7% in the standard-DHA group). The primary outcome was neurodevelopment of the infant at 18 months, as measured by the MDI, while significant mental delay (MDI<70) was a key secondary outcome. These outcomes were reanalysed for the 584 infants from a single or twin birth who remained after excluding infants with missing outcomes, along with nine sets of triplets for comparison with the simulation results. Treatment effects were estimated based on a linear model for the MDI, and both a logistic and log binomial model (see Online Supplement) for significant mental delay.

Results

Simulation Study

The simulation results for a continuous outcome are given in Table 1 (see also Online Supplement). Average unadjusted treatment effect estimates were very similar for all methods and ICCs when the treatment effect was the same for singletons and twins, but varied otherwise. When the true treatment effect among twins was 2, average unadjusted treatment effect estimates were close to the true overall mother-level treatment effect of 3.60 for CWGEE, which is the average of the true treatment effects for singletons and twins, weighted by the expected proportion of mothers with single and twin births (i.e. $(0.8\times4)+(0.2\times2)=3.60$). For GEE$_{\text{ind}}$, estimates were close to the true overall infant-level treatment effect of 3.33, which is the average of the true treatment effects for singletons and twins, weighted by the expected proportion of singleton and twin infants (i.e. $(0.667\times4)+(0.333\times2)=3.33$). Likewise, when the true treatment effect was 6 among twins, average unadjusted treatment effect estimates were around 4.40 and 4.67 for CWGEE and GEE$_{\text{ind}}$ respectively (see Online Supplement). As the ICC increased, average unadjusted treatment effect estimates remained stable for CWGEE and GEE$_{\text{ind}}$. For GEE$_{\text{exch}}$, estimates were similar to GEE$_{\text{ind}}$ when the ICC was low but similar to CWGEE when the ICC was high. Independently of whether the treatment effect differed between singletons and twins, average unadjusted standard errors increased with the ICC for all methods. CWGEE produced the largest standard errors when the ICC was low, whereas GEE$_{\text{ind}}$ produced the largest standard errors when the ICC was high, although differences between methods were fairly small. Adjusting for multiple birth status only slightly reduced the standard errors and made little difference to the treatment effect estimates on average. All methods produced identical results for each simulated dataset when the correct interaction model (1) was fitted to the data, although the power to detect an interaction was very low, ranging from 10.3% to 13.6%. Similar results were obtained for binary outcomes (see Online Supplement).

The impact of varying the multiple birth rate on the average unadjusted treatment effect estimate for a continuous outcome with different treatment effects for singletons and twins is shown in Figure 1. When the multiple birth rate was low, estimates were similar between methods and close to the true treatment effect of 4 for singletons. As the multiple birth rate increased, estimates approached the true treatment effect of 2 (Figure 1A) or 6 (Figure 1B) for twins. The relationship was approximately linear for CWGEE but exponential for GEE_{ind}, with GEE_{exch} estimates falling in between. The difference between methods increased as the percentage of twins moved away from 0% or 100%. A similar pattern was seen for adjusted treatment effect estimates (data not shown) and binary outcomes (see Online Supplement).

Example Dataset

Treatment effect estimates for the DHA trial are given in Table 2. For the MDI, treatment effect estimates were somewhat different between analysis methods but none produced sufficient evidence to support the hypothesis that high-DHA increases the mean MDI. For significant mental delay, unadjusted odds ratios ranged from 0.47 for GEE_{ind} to 0.52 for CWGEE. This relatively small difference between methods is consistent with the results from the simulations where a treatment group by multiple birth interaction was present. Subgroup analyses produced an estimated odds ratio of 0.60 among singletons and 0.31 among twins, but there was little evidence to suggest that the effect of treatment varied by birth size ($p>0.4$ for all GEE methods). If an infant-level interpretation of the odds ratio is of interest, the GEE_{ind} results suggest that treatment reduces the odds of significant mental delay by 53% for a randomly selected high-DHA infant compared with a randomly selected standard-DHA infant, irrespective of which mother they belong to. If a mother-level interpretation is desired, the CWGEE results suggest that treatment reduces the odds of significant mental delay by 48% comparing a randomly selected infant from a randomly selected high-DHA mother with a randomly selected infant from a randomly selected standard-DHA mother. Estimated relative risks can be interpreted similarly.

High-DHA was shown to be effective for reducing both the odds and risk of significant mental delay using $\mathrm{GEE}_{\mathrm{ind}}$ (p=0.03) and $\mathrm{GEE}_{\mathrm{exch}}$ (p=0.04), while the evidence in favour of the intervention was less convincing using CWGEE (p=0.06).

Comments

We have explored the problem of ICS in randomised trials including infants from both single and multiple births. We considered scenarios where ICS arises due to differences in average outcomes between singletons and multiples (no interaction), and differences in average treatment effects between singletons and multiples (interaction). Treatment effect estimates were obtained from main effects models only, as recommended for randomised trials.[22] Our simulation results indicate that treatment effect estimates are only influenced by ICS in the latter scenario (Figure 2), in which case different GEE methods of analysis are expected to produce different treatment effect estimates, although the differences we found were relatively small. Whether these differences are of practical importance will depend on the context. Our example dataset illustrates the potential for the strength of evidence supporting the effectiveness of treatment to vary according to the GEE method chosen.

When treatment effects estimates differ between methods, $\mathrm{GEE}_{\mathrm{exch}}$ produces estimates similar to $\mathrm{GEE}_{\mathrm{ind}}$ when the ICC is low and similar to CWGEE when the ICC is high. This makes sense intuitively, since $\mathrm{GEE}_{\mathrm{exch}}$ weights infants in a way that minimises the variance of the treatment effect estimate.[19] When the ICC is low, a set of twins provides almost as much information as two singletons and the variance is minimised by giving each twin a weight close to one, similar to GEE_{ind}. In contrast, when the ICC is high, a set of twins provides little more information than a singleton and the variance is minimised by giving each mother a weight close to one, similar to CWGEE.

Since CWGEE, GEE_{ind} and GEE_{exch} are expected to produce different treatment effect estimates when a treatment group by multiple birth interaction is present, a method of analysis could be chosen after interactions have been investigated, according to Figure 2. If there is insufficient evidence of an interaction, treatment effects could be estimated using any GEE method. GEE_{exch} may be preferred for maximising efficiency, although efficiency gains associated with this method were minimal in our simulation study due to treatment assignment at the mother level.[24,25] If evidence of an interaction is found, treatment effects could be estimated using CWGEE or GEE_{ind}, depending on whether a mother-level or an infant-level interpretation is preferred. GEE_{exch} should not be used in this case, since it fails to estimate a treatment parameter of interest.[14] The alternative to this data-driven approach is to pre-specify a method of analysis that remains appropriate across a range of scenarios. Since interactions are often plausible, CWGEE or GEE_{ind} would be chosen for analysis, depending on the desired interpretation. We prefer this approach in the randomised trial setting, since statistical methods should be pre-specified before issues such as ICS can be investigated in the data and interaction tests are typically underpowered.[26]

The choice between GEE_{ind} and CWGEE will depend on the context. GEE_{ind} estimates the effect of treatment for a typical infant, while CWGEE estimates the effect of treatment for a typical infant from a typical mother.[14] Since the former interpretation is most relevant for describing the impact of treatment on the total burden of disease and the demand for specialised child health or education services, we recommend using GEE_{ind} to estimate treatment effects in general. This method has appropriate type I error and coverage rates when birth size is uninformative,[7] and has the advantage of producing unadjusted treatment effect estimates that are consistent with the raw means or percentages for each treatment group. If the mother's perspective is actually of primary interest, this should be justified in the trial protocol and CWGEE can then be used for analysis.

It may be argued that any GEE method can be chosen when the multiple birth rate is low, since differences in treatment effect estimates between methods are small in this case. Figure 1 suggests this may be a reasonable strategy when the multiple birth rate is 5% or less, as would be expected in trials recruiting from the general population of pregnant women. However, this strategy may be problematic for treatments that have very different effects in singletons and multiples, where larger differences between methods are expected, or for outcomes where small changes would be considered clinically important. The safest approach is to pre-specify a method of analysis that acknowledges the possibility of ICS and produces treatment effect estimates with the desired interpretation, irrespective of the multiple birth rate.

Adjusting for multiple birth status as a main effect had little impact on average treatment effect estimates but led to small gains in efficiency. Whether adjustment should be made for multiple birth status in practice depends on the context. Adjustment is problematic when there are few multiples in a trial, since the outcome may be the same for all multiples, and when multiple birth status is determined after treatment commences, since fetal resorption may be influenced by treatment group. However, adjustment can be useful in other settings. It is recommended when multiple birth status is used as a balancing factor in the randomisation,[27] and corrects for chance imbalance in the multiple birth rate between treatment groups otherwise. Adjustment can also increase efficiency if multiple birth status is associated with the outcome.[28] Our results indicate that adjustment doesn't eliminate ICS when treatment effect estimation is of interest and the effect of treatment varies according to cluster size.

Randomised trials including multiple births differ from most settings where GEE methods for handling ICS have been investigated previously,[11,15,16,29-32] due to the small cluster sizes and focus on treatment effect estimation. Small clusters were considered in a recent simulation study comparing GEE_{ind} and GEE_{exch} to within cluster resampling,[11] which is asymptotically equivalent to CWGEE,[15] in a non-randomised setting. The authors concluded that all methods performed well for estimating covariate effects, but that within cluster resampling should be preferred if intercept estimation is of interest.[33] Others have noted that when the covariate effects are the same regardless of cluster size, ICS often has little impact on parameter estimates aside from the intercept,[34] which is of limited interest in randomised trials. Our findings indicate that treatment effect estimates are also influenced by ICS when the effect of treatment varies according to cluster size. As such interactions are often plausible, ICS is a serious concern for trials including multiple births.

ICS can arise in neonatal and perinatal trials whenever clustering is present and cluster sizes vary, which may occur for reasons other than multiple births. Our findings can reasonably be extended to settings where siblings from different births are present, while further research is needed to understand how methods for handling ICS perform in longitudinal settings. ICS is rarely a concern when analysing outcomes that are measured on the mother, since each mother can usually be considered independent.

A limitation of this study is that only randomisation at the mother level was considered. This approach is necessary for interventions given to the mother, and is often preferred by parents for interventions given to the infant, making it the most common choice in practice.[1,2] If randomisation is performed at the infant level, choosing CWGEE or GEE_{ind} over GEE_{exch} in the absence of a treatment group by multiple birth interaction is expected to result in greater efficiency losses than those observed in our study.[24,25] A further limitation is that only GEE methods for addressing ICS were examined. Clustered data are also commonly analysed using mixed-effects models, and methods for handling ICS in this context have been discussed previously; see[14] and references therein. Such approaches may be of limited use in randomised trials including multiple births, since GEEs are more popular in practice[1,2] and perform well in this setting.[7]

In conclusion, informative birth size is always a possibility in randomised trials including infants from both single and multiple births, and analysis methods should be pre-specified with this in mind. If a treatment group by multiple birth interaction is present, different GEEs are expected to produce different treatment effect estimates with different interpretations. We recommend estimating treatment effects using standard GEEs with an independence working correlation structure to give an infant-level interpretation.

Acknowledgements

The authors would like to thank Professor Maria Makrides (Chair) on behalf of the DINO Steering Committee (Maria Makrides, Robert Gibson, Andrew McPhee, Carmel Collins, Peter Davis, Lex Doyle, Karen Simmer, Paul Colditz, Scott Morris and Philip Ryan) for granting permission to use data from the high-dose versus standard-dose DHA trial, and Professor John Carlin, Dr Beverly Muhlhausler and Dr Jacqueline Gould for providing valuable comments on earlier drafts of this paper. Lisa Yelland is the recipient of an Australian National Health and Medical Research Council Early Career Fellowship (ID 1052388). Shaun Seaman is funded by a United Kingdom Medical Research Council Programme Grant (ID U1052 60558).

References

1. Hibbs AM, Black D, Palermo L, Cnaan A, Luan XQ, Truog WE, et al. Accounting for multiple births in neonatal and perinatal trials: systematic review and case study. *Journal of Pediatrics* 2010; 156:202-208.
2. Yelland LN, Sullivan TR, Makrides M. Accounting for multiple births in randomised trials: a systematic review. *Archives of Disease in Childhood-Fetal and Neonatal Edition* 2015; 100:F116-120.
3. Gates S, Brocklehurst P. How should randomised trials including multiple pregnancies be analysed? *BJOG: an International Journal of Obstetrics and Gynaecology* 2004; 111:213-219.

4. Marston L, Peacock JL, Yu KM, Brocklehurst P, Calvert SA, Greenough A, et al. Comparing methods of analysing datasets with small clusters: case studies using four paediatric datasets. *Paediatric and Perinatal Epidemiology* 2009; 23:380-392.

5. Shaffer ML, Kunselman AR, Watterberg KL. Analysis of neonatal clinical trials with twin births. *BMC Medical Research Methodology* 2009; 9:12.

6. Shaffer ML, Hiriote S. Analysis of time-to-event and duration outcomes in neonatal clinical trials with twin births. *Contemporary Clinical Trials* 2009; 30:150-154.

7. Yelland LN, Salter AB, Ryan P, Makrides M. Analysis of binary outcomes from randomised trials including multiple births: when should clustering be taken into account? *Paediatric and Perinatal Epidemiology* 2011; 25:283-297.

8. Sauzet O, Wright KC, Marston L, Brocklehurst P, Peacock JL. 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* 2013; 32:1429-1438.

9. Ananth CV, Platt RW, Savitz DA. Regression models for clustered binary responses: implications of ignoring the intracluster correlation in an analysis of perinatal mortality in twin gestations. *Annals of Epidemiology* 2005; 15:293-301.

10. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; 73:13-22.

11. Hoffman EB, Sen PK, Weinberg CR. Within-cluster resampling. *Biometrika* 2001; 88:1121-1134.

12. Choi SJ, Song SE, Seo ES, Oh SY, Kim JH, Roh CR. The effect of single or multiple courses of antenatal corticosteroid therapy on neonatal respiratory distress syndrome in

singleton versus twin pregnancies. *Australian & New Zealand Journal of Obstetrics & Gynaecology* 2009; 49:173-179.

13. Shinwell ES, Haklai T, Eventov-Friedman S. Outcomes of multiplets. *Neonatology* 2009; 95:6-14.

14. Seaman S, Pavlou M, Copas A. Review of methods for handling confounding by cluster and informative cluster size in clustered data. *Statistics in Medicine* 2014; 33:5371-5387.

15. Williamson JM, Datta S, Satten GA. Marginal analyses of clustered data when cluster size is informative. *Biometrics* 2003; 59:36-42.

16. Benhin E, Rao JNK, Scott AJ. Mean estimating equation approach to analysing cluster-correlated data with nonignorable cluster sizes. *Biometrika* 2005; 92:435-450.

17. Ballard RA, Truog WE, Cnaan A, Martin RJ, Ballard PL, Merrill JD, et al. Inhaled nitric oxide in preterm infants undergoing mechanical ventilation. *New England Journal of Medicine* 2006; 355:343-353.

18. Seaman SR, Pavlou M, Copas AJ. Methods for observed-cluster inference when cluster size is informative: A review and clarifications. *Biometrics* 2014; 70:449-456.

19. Hanley JA, Negassa A, Edwardes MDD, Forrester JE. Statistical analysis of correlated data using generalized estimating equations: An orientation. *American Journal of Epidemiology* 2003; 157:364-375.

20. Nevalainen J, Datta S, Oja H. Inference on the marginal distribution of clustered data with informative cluster size. *Statistical Papers* 2014; 55:71-92.

21. Bayley N. *Manual for the Bayley Scales of Infant Development, Second Edition (BSID-II).* San Antonio, TX: Psychological Corp; 1993.

22. Committee for Proprietary Medicinal Products. Points to consider on adjustment for baseline covariates. *Statistics in Medicine* 2004; 23:701-709.

23. Makrides M, Gibson RA, McPhee AJ, Collins CT, Davis PG, Doyle LW, et al. Neurodevelopmental outcomes of preterm infants fed high-dose docosahexaenoic acid: a randomized controlled trial. *Journal of the American Medical Association* 2009; 301:175-182.

24. Fitzmaurice GM. A caveat concerning independence estimating equations with multivariate binary data. *Biometrics* 1995; 51:309-317.

25. Mancl LA, Leroux BG. Efficiency of regression estimates for clustered data. *Biometrics* 1996; 52:500-511.

26. ICH E9 Expert Working Group. Statistical principles for clinical trials. *Statistics in Medicine* 1999; 18:1905-1942.

27. Kahan BC, Morris TP. Improper analysis of trials randomised using stratified blocks or minimisation. *Statistics in Medicine* 2012; 31:328-340.

28. Neuhaus JM. Estimation efficiency with omitted covariates in generalized linear models. *Journal of the American Statistical Association* 1998; 93:1124-1129.

29. Williamson JM, Kim HY, Warner L. Weighting condom use data to account for nonignorable cluster size. *Annals of Epidemiology* 2007; 17:603-607.

30. Panageas KS, Schrag D, Localio AR, Venkatraman ES, Begg CB. Properties of analysis methods that account for clustering in volume-outcome studies when the primary predictor is cluster size. *Statistics in Medicine* 2007; 26:2017-2035.

31. Huang Y, Leroux B. Informative cluster sizes for subcluster-level covariates and weighted generalized estimating equations. *Biometrics* 2011; 67:843-851.

32. Pavlou M, Seaman SR, Copas AJ. An examination of a method for marginal inference when the cluster size is informative. *Statistica Sinica* 2013; 23:791-808.

33. Xu Y, Lee CF, Cheung YB. Analyzing binary outcome data with small clusters: a simulation study. *Communications in Statistics-Simulation and Computation* 2014; 43:1771-1782.

34. Neuhaus JM, McCulloch CE. Estimation of covariate effects in generalized linear mixed models with informative cluster sizes. *Biometrika* 2011; 98:147-162.

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web-site:

Additional Simulation Study Details

Figure 1: Average unadjusted treatment effect estimate for a continuous outcome with an ICC of 0.5 by varying percentage of mothers with a twin birth when the treatment effect is 4 for singletons and (A) 2 or (B) 6 for twins.

Figure 2: Flowchart summarising simulation results and analysis recommendations.