Accounting for Twin Births in Sample Size Calculations for Randomised Trials

Lisa N Yelland[a,b], Thomas R Sullivan[b], Carmel T Collins[a,c], David J Price[d,e], Andrew J McPhee[a,c,f], Katherine J Lee[g,h]

[a]South Australian Health and Medical Research Institute, Adelaide, South Australia, Australia

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

[c] The University of Adelaide, Adelaide Medical School, Adelaide, South Australia, Australia

[d]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria, Australia

[e]Victorian Infectious Diseases Reference Laboratory Epidemiology Unit at The Peter Doherty Institute for Infection and Immunity, The University of Melbourne and Royal Melbourne Hospital, Melbourne, Victoria, Australia

[f]Department of Neonatal Medicine, Women's and Children's Hospital, Adelaide, South Australia, Australia

[g]Melbourne Children's Trials Centre, Murdoch Children's Research Institute, Victoria, Australia

[h]Department of Paediatrics, University of Melbourne, Melbourne, Victoria, Australia

Correspondence: Dr Lisa Yelland, School of Public 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: 3388

Abstract

Background: Including twins in randomised trials leads to non-independence or clustering in the data. Clustering has important implications for sample size calculations, yet few trials take this into account. Estimates of the intracluster correlation coefficient (ICC), or the correlation between outcomes of twins, are needed to assist with sample size planning. Our aims were to provide ICC estimates for infant outcomes, describe the information that must be specified in order to account for clustering due to twins in sample size calculations, and develop a simple tool for performing sample size calculations for trials including twins.

Methods: ICCs were estimated for infant outcomes collected in four randomised trials that included twins. The information required to account for clustering due to twins in sample size calculations is described. A tool that calculates the sample size based on this information was developed in Microsoft Excel and in R as a Shiny web app.

Results: ICC estimates ranged between -0.12, indicating a weak negative relationship, and 0.98, indicating a strong positive relationship between outcomes of twins. Example calculations illustrate how the ICC estimates and sample size calculator can be used to determine the target sample size for trials including twins.

Conclusions: Clustering among outcomes measured on twins should be taken into account in sample size calculations to obtain the desired power. Our ICC estimates and sample size calculator

will be useful for designing future trials that include twins. Publication of additional ICCs is needed to further assist with sample size planning for future trials.

Keywords: multiple birth, statistical methodology, generalised estimating equations, power

Introduction

Perinatal trials often include infants from both single and multiple births.[1, 2] While outcomes measured on singletons can generally be considered independent, outcomes measured on infants from a multiple birth tend to be similar due to the shared environment and genetic factors.[3, 4] Data from these trials are therefore clustered, where each mother or family is a cluster and the infants are the members of the cluster.[3] The implications of this clustering for the statistical analysis of infant outcomes have been widely discussed, with many researchers recommending that clustering be taken into account in the analysis whenever multiple births are present.[1, 3, 5-7] In contrast, the impact of clustering due to multiple births on the sample size calculation has received little attention. It is important to account for clustering in sample size calculations to ensure trials are appropriately powered, since clustering can reduce the effective sample size.[8, 9] However, clustering due to multiple births is typically ignored when determining the target sample size for perinatal trials.[2]

Two main challenges arise when attempting to account for multiple births in sample size calculations. Firstly, an estimate of the strength of the dependence between outcomes of infants from the same birth is required. This is measured by the intracluster correlation coefficient (ICC), which ranges from -1 to 1.[9] An ICC of 0 indicates that outcomes of multiples are independent, and the ICC approaches 1 or -1 as their outcomes become more positively or negatively correlated, respectively. The ICC is rarely reported for trials including multiple births[2] and depends on a number of factors including the target population, study design, type of intervention, estimation method and outcome of interest.[10] ICC estimates for a range of infant outcomes from various trials are therefore needed to assist with sample size calculations for future trials. Secondly, standard

sample size calculation methods assume that the outcomes of all infants are independent. Methods have recently been developed that could be used to calculate the sample size for trials including multiple births but currently only apply to twins. These methods have been described in the statistical literature and involve complex equations.[11] To facilitate their use in practice, a more user-friendly presentation of the information required to calculate the sample size in the context of perinatal trials, and a simple tool for calculating the sample size, are needed.

In this article we aim to (1) provide ICC estimates for a range of infant outcomes that are commonly collected in perinatal trials; (2) describe the information that must be specified in order to account for clustering due to twin births in sample size calculations; and (3) provide a simple tool for calculating the target sample size for trials including twin births. Consistent with currently available sample size calculation methods,[11] we restrict our attention in aims 2 and 3 to trials involving singletons and twins, where either mothers or infants are randomised to an intervention or control treatment in a 1:1 ratio, and with a continuous or binary primary outcome that will be analysed using regression models with generalised estimating equation (GEE) estimation to account for clustering due to twins. Other scenarios, including higher order multiples, are considered in the Comment section.

Methods

Estimating intracluster correlation coefficients

ICCs were estimated using data from four randomised trials that included infants from a multiple birth: the Docosahexaenoic Acid for the Improvement of Neurodevelopmental Outcomes

in Preterm Infants (DINO) Trial,[12] the Bottles, Cups and Dummies (BCD) Trial,[13] the Twins Timing of Birth (TTB) Trial[14] and the n-3 Fatty Acids for Improvement in Respiratory Outcomes (N3RO) Trial.[15] The characteristics of these trials are summarised in Table 1 and a more detailed description is provided in the Supplementary Material. ICCs were estimated for a range of infant outcomes (Tables 2 and 3). Birthweight z-score and small for gestational age were derived using Australian standards for singletons and twins,[16, 17] and follow-up weight z-score was derived using World Health Organization child growth standards.[18]

ICCs were estimated in SAS v9.4 using a linear model for continuous variables, and both a logistic model and a log binomial model for binary variables. A treatment group effect was included in the model for variables measured after randomisation, since ignoring the treatment group can lead to biased ICC estimates,[19] and a single ICC was estimated across both treatment groups. Analyses were performed with and without adjustment for multiple birth status as a fixed effect for the three trials that included infants from both single and multiple births, since multiple birth status is sometimes used as an adjustment variable in the analysis[2] and adjustment can reduce the ICC.[20] Triplets were included in the analysis for the DINO and N3RO trials but were excluded in a sensitivity analysis. Clustering due to multiple births was taken into account using GEEs, as this is the most popular estimation method for accounting for this type of clustering [1, 2] and has been shown to perform well in this setting.[6] ICCs were estimated based on an exchangeable working correlation structure (no ICC is estimated if an independence structure is specified) and confidence intervals were obtained via bootstrapping using the bias corrected and accelerated method[21] with resampling of clusters.

Sample size calculations for trials including twins

A general approach for determining the target sample size for any study involving clustered data begins with calculating the sample size using standard methods that assume all outcomes are independent (see e.g.[22-24]). The resulting sample size is then multiplied by the design effect (DEFF), which measures the degree of inflation required to account for clustering in the data. The DEFF can be any positive number, with values above and below 1 indicating that a larger or smaller sample size is required to account for clustering in the data, respectively. The formula for the DEFF varies depending on the setting. For randomised trials including twins, DEFF equations have been published elsewhere[11] and several design features and parameter estimates must be specified in order to calculate the relevant DEFF. First, the method for randomising twins needs to be decided. If the mother is to receive the treatment then both twins will necessarily be randomised to the same group. If the infant is to receive the treatment then twins may be randomised to the same group (cluster randomisation), independently of each other (individual randomisation) or to opposite treatment groups. Second, the primary outcome needs to be defined, including whether it will be a binary (e.g. neonatal morbidity) or continuous measure (e.g. weight). For binary outcomes, the expected outcome prevalence in each treatment group and the effect measure of interest (odds ratio or relative risk, estimated by the logistic or log binomial model, respectively) must also be specified. Third, the approach that will be used to account for clustering in the statistical analysis needs to be chosen. We have focused on GEEs and a simple introduction to this estimation method can be found elsewhere.[25] It requires specification of an assumed pattern of correlation among outcomes for each cluster known as a working correlation structure, and an independence or exchangeable working correlation structure may be chosen for analysing trials including twins. The latter estimates the correlation between outcomes of pairs of members from

the same cluster, which is assumed to be constant, while the former assumes this correlation is 0 and accounts for clustering using robust variance estimation.[26] Fourth, an estimate of the twin birth rate for the trial is required. This will vary depending on the population of interest and estimates may be obtained from published trials in similar populations or national birth statistics. Care must be taken to distinguish between the percentage of mothers in a trial who give birth to twins, and the percentage of infants in a trial who are from a twin birth. For example, in a trial recruiting 100 pregnant women of whom 10 give birth to twins, 10% of women have a twin birth whereas 18% of infants (20 out of 110) are from a twin birth. Either percentage can be used to calculate the DEFF using our sample size tool. Finally, an estimate of the ICC among twins is needed for the primary outcome.

Once these factors have been specified, the DEFF can be calculated by hand using the appropriate equation[11] and used to multiply the sample size obtained assuming independence to produce the target sample size accounting for clustering due to twins. Alternatively, we have developed a sample size calculator that allows researchers to enter the necessary information into a Microsoft Excel worksheet or R Shiny web app that calculates the DEFF and sample size automatically. Example sample size calculations were performed using Stata v14.2 (assuming all infants are independent) and this calculator.

Results

Intracluster correlation coefficients

ICC estimates obtained from the four example trials are presented in Table 2 for continuous outcomes and Table 3 for binary outcomes based on a logistic model. Very similar results were obtained using a log binomial model (data not shown). ICC estimates were mostly positive and ranged between -0.12, indicating a weak negative relationship between outcomes of infants from the same birth, and 0.98, indicating a strong positive relationship. Most ICCs were estimated imprecisely, as reflected in the relatively wide confidence intervals. Differences in ICC estimates for the same outcome across trials were often substantial. For example, ICCs for birthweight ranged between 0.47 and 0.82. Means and prevalences of the outcomes also varied between trials, highlighting differences in the populations studied. Adjusting for multiple birth status generally made little difference to the ICC estimates, while excluding triplets typically resulted in slightly larger ICC estimates (see Supplementary Material).

Sample size calculator

A sample size tool that calculates the DEFF and target sample size for randomised trials including twins in Microsoft Excel is provided in the Supplementary Material. The calculator is also available as an R Shiny web app at https://djprice.shinyapps.io/gee_calculator/ and R code for local implementation of the calculator is available at https://github.com/DJPrice10/GEE-Sample-Size-Calculator (see Supplementary Material for further details). The user must input the information required to calculate the DEFF, including the twin birth rate and the ICC. To illustrate how this calculator and the ICC estimates can be used to determine the target sample size for a future trial, consider the following hypothetical examples. Suppose a trial is planned to assess the effect of a new infant formula designed to improve growth in infants born preterm (<37 weeks'

gestation). The primary outcome is weight z-score at 18 months of age and the new formula is expected to increase the mean z-score by 0.3 compared with standard infant formula. Assuming all infants in the trial will be independent, 235 infants are required per group to have 90% power to detect this increase, based on a two-sample t-test (alpha=0.05, standard deviation=1). Suppose, however, that 70% of infants recruited are expected to be from a single birth, while 30% are expected to be from a twin birth. As the intervention will be given to the infants, all methods of randomising twins are possible in this trial. The DEFF was calculated using the sample size calculator with an ICC of 0.58 based on the DINO Trial (Table 2), since both this trial and the hypothetical trial focus on a preterm population with weight measurements at 18 months. Assuming the treatment effect will be estimated using GEEs with an independence working correlation structure, the target sample size varies from 195 infants per group if twins are assigned to opposite treatment groups, to 276 infants per group if cluster randomisation is chosen (see Supplementary Material).

As a second example, suppose another trial is planned to assess the effect of a dietary intervention given to pregnant women on the percentage of infants born small for gestational age, where the intervention is expected to reduce this from 10% to 5%. Assuming all infants in the trial will be independent, 621 infants are required per group based on a two-sided continuity-corrected chi-square test (90% power, alpha=0.05). However, 90% of women recruited are expected to have a single birth, while 10% are expected to have a twin birth. As the intervention will be given to pregnant women, the trial will necessarily use cluster randomisation for twins. The target population for the hypothetical trial is somewhat different to the population included in the four example trials and hence the largest ICC estimate for being small for gestational age across the four trials of 0.36 (Table 3) was chosen to produce the largest sample size, which is most likely to

provide adequate power. Using the sample size calculator, the target sample size increases to 662 or 653 infants per group if the analysis will be performed assuming an independence or exchangeable working correlation structure respectively, independent of whether the odds ratio or relative risk is the effect measure of interest (see Supplementary Material).

To understand the impact of twins on sample size calculations more generally, the relationship between the ICC and the DEFF is shown in Figure 1, and the relationship between the ICC and the sample size is shown in the Supplementary Material. When the twin birth rate is 3%, which may be expected in the general population of pregnant women,[27] the DEFF is close to 1 and hence twins have little impact on the sample size unless the ICC is relatively large (Figure 1A). For a twin birth rate of 20%, which is typical in trials targeting preterm infants,[2] even a small ICC can lead to a DEFF that is substantially different from 1 and hence twins can have a large impact on the target sample size (Figure 1B). Assuming the ICC is positive, including twins in a trial results in a larger sample size for cluster randomisation, and the same or a smaller sample size for individual randomisation or randomisation to opposite treatment groups, compared with a trial involving singletons only. A larger sample size is required if the independence rather than the exchangeable working correlation structure is planned for the analysis.

Comment

Principal findings

We provide ICC estimates for a range of routinely collected infant outcomes from four randomised trials that included infants from a multiple birth. ICC estimates were generally positive and as high as 0.98. We also describe the information that must be specified in order to calculate

the DEFF and hence the sample size for randomised trials including twins, and provide a simple tool to perform the calculations automatically. Our example sample size calculations demonstrate that the inclusion of twins can have a substantial impact on the target sample size.

Interpretation

We found substantial variation in ICC estimates between outcomes and across trials, which may be due to differences in the outcome measure, study setting, and inclusion/exclusion criteria. We also found that confidence intervals for ICC estimates were relatively wide, even for the N3RO trial with over 1200 participants, suggesting that quite large sample sizes are required to accurately estimate this parameter. This highlights the need for estimates to be published for a range of outcomes collected in different trials and larger epidemiological studies performed in varying populations, thus allowing the researcher to use the most relevant ICC for sample size planning. Importantly, ICC estimates based on twins born at term may not be relevant to twins born preterm, and different outcomes will be of greater clinical interest depending on the gestational age range and other characteristics of the target population.

We presented ICCs estimated with and without adjustment for multiple birth status. While adjustment made little difference to ICC estimates for most outcomes, it did have a substantial impact in some cases. When planning a trial involving both singletons and twins, the sample size should be determined using adjusted ICC estimates if the primary analysis will be adjusted for twin birth status, and unadjusted ICC estimates otherwise. For trials that will only include twins, we suggest basing the sample size on adjusted ICC estimates, or estimates of the ICC obtained from datasets that only include twin births.

The target sample size can vary considerably depending on how twins will be randomised. Assuming the ICC is positive, assigning twins to different treatment groups requires the smallest sample size but may make recruitment challenging, since it is the least favoured method among parents of twins.[28] Cluster randomisation is the most common randomisation approach used in perinatal trials,[1, 2] is preferred by parents[28] and is the only option for treatments given to the mother, but it also requires the largest sample size. Where feasible, individual randomisation may provide a good compromise by requiring a smaller sample size than cluster randomisation and being more acceptable to parents than randomisation to different groups.[28]

To calculate the sample size for a trial that will be analysed using regression models with GEE estimation, a decision must be made between assuming an independence or exchangeable working correlation structure. Although the exchangeable working correlation structure is more realistic, requires a smaller sample size[11] and may result in a more efficient analysis,[29, 30] it is prone to convergence problems[11] and may not produce treatment effect estimates with a meaningful interpretation.[31] We therefore recommend assuming an independence working correlation structure for sample size planning and analysis.

When the outcome of interest is binary, the target sample size can vary depending on whether the treatment effect of interest is the odds ratio or the relative risk, although differences will generally be small. The relative risk is popular due to the ease of interpretation but convergence problems can occur when fitting the log binomial model. These can easily be overcome using a modified Poisson approach[32] that remains valid for clustered data.[33] As the log binomial and modified Poisson regression models have similar performance when both converge,[33] we expect sample size calculations based on the log binomial model will provide the desired power if the final analysis is performed using modified Poisson regression.

Although this study focused on randomised trials including twins, our findings will be helpful for sample size planning in other settings. For trials involving higher order multiples, ignoring them and using the sample size calculator to account for twins may give a reasonable approximation, provided higher order multiples are expected to be rare. Alternatively, new DEFF equations could be derived that allow for higher order multiples and this is an important area for future research. For observational studies involving twins, the ICC estimates remain relevant but the sample size calculator is currently limited to scenarios where the exposure of interest is binary with 50% prevalence (as is the case in randomised trials with two treatment groups and a 1:1 allocation ratio).

Strengths of the study

This study provides ICC estimates for a range of outcomes from multiple trials conducted using different methods of randomisation in different populations, thus allowing researchers to choose the most relevant ICC estimate for planning future trials. It also provides the first sample size tool that can be used to account for twins in sample size calculations automatically.

Limitations of the data

Our study has several limitations. First, zygosity was not considered in the ICC calculations, since this information was not collected in the example trials. ICCs may be higher for monozygous compared to dizygous twins[34] and future studies collecting zygosity could report ICCs separately for these subgroups. Second, the sample size calculator assumes the analysis will

be performed using regression models with GEE estimation. Further research is needed to determine whether the calculator is suitable to use when other analysis approaches are planned, such as mixed effects models, due to potential differences in power. Third, we only present ICC estimates from four trials and many outcomes were not available for all trials or were based on a neurodevelopmental assessment tool that has since been updated. Additional published ICCs are needed. Finally, sample size methods for survival and count outcomes were not considered and this is an interesting area for future work.

Conclusions

Clustering due to twins should be taken into account in sample size calculations for perinatal trials to obtain the desired power. We have shown how to achieve this by multiplying the sample size obtained assuming independence by an appropriate DEFF using our calculator. We have also reported ICC estimates that can be used to inform the sample size for future trials. Publication of ICCs for additional infant outcomes and from a broader range of studies is needed to further assist with sample size planning for future trials including twins.

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 DINO Trial; Professor Jodie Dodd on behalf of the TTB Steering Group (Jodie Dodd, Caroline Crowther, Ross Haslam and Jeffrey Robinson) for granting permission to use data from the TTB Trial; members of the N3RO Steering Committee (Carmel Collins, Robert Gibson, Andrew McPhee, Thomas Sullivan and Maria Makrides) for granting permission to use data from the N3RO Trial; Mr Vincent Ball for providing programming assistance for the Microsoft Excel sample size calculator; and Ms Louise Linsell for providing feedback on a draft version of the Microsoft Excel calculator. Ethical approval was obtained for this study from the Women's & Children's Health Network Human Research Ethics Committee (HREC/14/WCHN/165, December 2014). Lisa Yelland, Carmel Collins and Katherine Lee were supported by Australian National Health and Medical Research Council Fellowships [Early Career Fellowship ID 1052388 for L. Y., Translating Research Into Practice Fellowship ID 1132596 for C. C., and Career Development Fellowship ID 1053609 for K. L.]. This research was supported by a Women's and Children's Hospital Foundation Research Project Grant.

Supporting Information

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

Appendix S1. Microsoft Excel sample size calculator

Appendix S2. Supplementary material including: description of four randomised trials used to estimate intracluster correlation coefficients; adjusted intracluster correlation coefficient estimates; Excel sample size calculator output for example sample size calculations; R Shiny app sample size calculator instructions and output for example sample size calculations; and relationship between intracluster correlation coefficient and sample size.

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. 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.
6. 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.
7. 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.

8. Vierron E, Giraudeau B. Design effect in multicenter studies: gain or loss of power? *Bmc Medical Research Methodology*. 2009; 9:39.

9. Killip S, Mahfoud Z, Pearce K. What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. *Annals of Family Medicine*. 2004; 2:204-208.

10. Taljaard M, Donner A, Villar J, Wojdyla D, Velazco A, Bataglia V, et al. Intracluster correlation coefficients from the 2005 WHO Global Survey on Maternal and Perinatal Health: implications for implementation research. *Paediatric and Perinatal Epidemiology*. 2008; 22:117-125.

11. Yelland LN, Sullivan TR, Price DJ, Lee KJ. Sample size calculations for randomised trials including both independent and paired data. *Statistics in Medicine*. 2017; 36:1227-1239.

12. 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.

13. Collins CT, Ryan P, Crowther CA, McPhee AJ, Paterson S, Hiller JE. Effect of bottles, cups, and dummies on breast feeding in preterm infants: randomised controlled trial. *British Medical Journal*. 2004; 329:193-196.

14. Dodd JM, Crowther CA, Haslam RR, Robinson JS, Twins Timing of Birth Trial Group. Elective birth at 37 weeks of gestation versus standard care for women with an uncomplicated twin pregnancy at term: the Twins Timing of Birth Randomised Trial. *BJOG: an International Journal of Obstetrics and Gynaecology*. 2012; 119:964-974.

15. Collins CT, Makrides M, McPhee AJ, Sullivan TR, Davis PG, Thio M, et al. Docosahexaenoic Acid and Bronchopulmonary Dysplasia in Preterm Infants. *New England Journal of Medicine*. 2017; 376:1244-1254.

16. Roberts CL, Lancaster PA. Australian national birthweight percentiles by gestational age. *Medical Journal of Australia*. 1999; 170:114-118.

17. Roberts CL, Lancaster PA. National birthweight percentiles by gestational age for twins born in Australia. *Journal of Paediatrics and Child Health*. 1999; 35:278-282.

18. WHO Multicentre Growth Reference Study Group. WHO Child Growth Standards based on length/height, weight and age. *Acta Paediatrica Supplement*. 2006; 450:76-85.

19. Giraudeau B. Model mis-specification and overestimation of the intraclass correlation coefficient in cluster randomized trials. *Statistics in Medicine*. 2006; 25:957-964.

20. Murray DM, Blitstein JL. Methods to reduce the impact of intraclass correlation in group-randomized trials. *Evaluation Review*. 2003; 27:79-103.

21. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine*. 2000; 19:1141-1164.

22. Carlin JB, Doyle LW. Statistics for clinicians: 7: sample size. *Journal of Paediatrics and Child Health*. 2002; 38:300-304.

23. Kirby A, Gebski V, Keech AC. Determining the sample size in a clinical trial. *Medical Journal of Australia*. 2002; 177:256-257.

24. Rohrig B, du Prel JB, Wachtlin D, Kwiecien R, Blettner M. Sample size calculation in clinical trials part 13 of a series on evaluation of scientific publications. *Deutsches Arzteblatt International*. 2010; 107:552-556.

25. 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.

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

27. Ananth CV, Chauhan SP. Epidemiology of twinning in developed countries. *Seminars in Perinatology*. 2012; 36:156-161.

28. Bernardo J, Nowacki A, Martin R, Fanaroff JM, Hibbs AM. Multiples and parents of multiples prefer same arm randomization of siblings in neonatal trials (vol 35, pg 208, 2015). *Journal of Perinatology*. 2015; 35:310-310.

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

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

31. Yelland LN, Sullivan TR, Pavlou M, Seaman SR. Analysis of randomised trials including multiple births when birth size is informative. *Paediatric and Perinatal Epidemiology*. 2015; 29:567-575.

32. Zou GY. A modified Poisson regression approach to prospective studies with binary data. *American Journal of Epidemiology*. 2004; 159:702-706.

33. Yelland LN, Salter AB, Ryan P. Performance of the modified Poisson regression approach for estimating relative risks from clustered prospective data. *American Journal of Epidemiology*. 2011; 174:984-992.

34. Baird J, Osmond C, MacGregor A, Snieder H, Hales CN, Phillips DI. Testing the fetal origins hypothesis in twins: the Birmingham twin study. *Diabetologia*. 2001; 44:33-39.

Table 1. Characteristics of four randomised trials that included infants from multiple births

	DINO Trial[12]	BCD Trial[13]	TTB Trial[14]	N3RO Trial[15]
Recruitment period	2001-2005	1996-1999	2003-2010	2012-2015
Location(s)	Australia	Australia	Australia, New Zealand, Italy	Australia, New Zealand, Singapore
Method of randomising twins	Cluster	Cluster	Cluster	Individual
Gestational age	<33 weeks'	<34 weeks'	≥37 weeks'	<29 weeks'
Total number of infants	657	319	470	1273
Singletons, n (%)	436 (66)	237 (74)	0 (0)	9112)
Twins, n (%)	194 (30)	82 (26)	470 (100)	3306)
Triplets, n (%)	27 (4)	0 (0)	0 (0)	32 (3)

DINO, Docosahexaenoic Acid for the Improvement of Neurodevelopmental Outcomes in Preterm Infants; BCD, Bottles, Cups and Dummies; TTB, Twins Timing of Birth; N3RO, n-3 Fatty Acids for Improvement in Respiratory Outcomes.

Table 2. Intracluster correlation coefficients for continuous outcomes measured in four randomised trials

Outcome	DINO Trial (n=657)	BCD Trial (n=319)	TTB Trial (n=470)	N3RO Trial (n=1273)
Birthweight (g)				
Mean (SD), N	1309 (420), 657	1383 (466), 318	2786 (342), 470	919 (238), 1273
ICC (95% CI)	0.60 (0.45, 0.77)	0.82 (0.59, 1.00)	0.47 (0.35, 0.57)	0.52 (0.40, 0.67)
Birthweight z-score				
Mean (SD), N	-0.06 (0.99), 657	-0.10 (0.97), 318	0.23 (0.92), 470	0.03 (0.97), 1273
ICC (95% CI)	0.16 (-0.08, 0.39)	0.44 (0.04, 0.84)	0.43 (0.32, 0.52)	0.26 (0.13, 0.42)
Follow up weight z-score[a]				
Mean (SD), N	0.11 (1.24), 598	-0.73 (1.45), 174	-0.24 (1.10), 319	NA
ICC (95% CI)	0.58 (0.41, 0.79)	0.46 (0.12, 1.00)	0.74 (0.65, 0.82)	NA
Mental development index standardized score[a,b]				
Mean (SD), N	94.76 (16.27), 611	NA	NA	NA
ICC (95% CI)	0.70 (0.48, 0.96)	NA	NA	NA

Psychomotor development index standardized score[a,b]				
Mean (SD), N	93.35 (15.78), 609	NA	NA	NA
ICC (95% CI)	0.30 (0.08, 0.53)	NA	NA	NA

NA, not applicable due to data not being collected; SD, standard deviation; N, number of infants with available data; ICC, intracluster correlation coefficient; CI, confidence interval; DINO, Docosahexaenoic Acid for the Improvement of Neurodevelopmental Outcomes in Preterm Infants; BCD, Bottles, Cups and Dummies; TTB, Twins Timing of Birth; N3RO, n-3 Fatty Acids for Improvement in Respiratory Outcomes.

[a] Measured at 18 months corrected age for the DINO Trial, 6 months post discharge for the BCD Trial (parent report) and 4 months of age for the TTB Trial.

[b] Measured using the Bayley Scales of Infant Development, Second Edition.

Table 3. Intracluster correlation coefficients for binary outcomes measured in four randomised trials

Outcome	DINO Trial (n=657)	BCD Trial (n=319)	TTB Trial (n=470)	N3RO Trial (n=1273)
Low birthweight (<2500g)				
Prevalence (%), N	99.9, 657	99.1, 318	20.4, 470	100, 1273
ICC (95% CI)	NA	NA	0.23 (0.09, 0.38)	NA
Very low birthweight (<1500g)				
Prevalence (%), N	63.6, 657	60.1, 318	0.0, 470	98.6, 1273
ICC (95% CI)	0.39 (0.19, 0.58)	0.74 (0.48, 0.94)	NA	NA
Small for gestational age (birthweight <10th percentile)				
Prevalence (%), N	15.6, 630	14.7, 314	5.3, 470	19.6, 1200
ICC (95% CI)	-0.01 (-0.14, 0.19)	0.36 (0.01, 0.93)	0.06 (-0.06, 0.34)	0.03 (-0.06, 0.11)
Predischarge death				
Prevalence (%), N	2.3, 656	3.8, 319	0.2, 470	6.9, 1273
ICC (95% CI)	0.00 (-0.04, 0.02)	NA	NA	0.32 (0.06, 0.74)

Admitted to neonatal intensive care unit				
Prevalence (%), N	88.4, 655	NA	3.0, 470	NA
ICC (95% CI)	0.85 (0.50, 1.00)	NA	0.26 (-0.02, 0.76)	NA
Any intraventricular haemorrhage				
Prevalence (%), N	13.6, 656	12.9, 318	NA	31.9, 1210
ICC (95% CI)	0.24 (0.03, 0.52)	0.10 (-0.09, 0.59)	NA	0.17 (0.03, 0.33)
Intraventricular haemorrhage grade 3 or 4				
Prevalence (%), N	2.3, 656	2.5, 318	NA	6.2, 1210
ICC (95% CI)	-0.01 (-0.05, 0.01)	NA	NA	0.07 (-0.04, 0.42)
Any retinopathy of prematurity				
Prevalence (%), N	22.5, 654	NA	NA	44.7, 1203
ICC (95% CI)	0.68 (0.43, 0.95)	NA	NA	0.51 (0.35, 0.64)

Retinopathy of prematurity grade 3 or higher				
Prevalence (%), N	4.7, 654	NA	NA	10.4, 1201
ICC (95% CI)	0.15 (-0.05, 0.80)	NA	NA	0.28 (0.09, 0.58)
Any sepsis				
Prevalence (%), N	15.4, 655	NA	NA	27.7, 1234
ICC (95% CI)	0.36 (0.13, 0.63)	NA	NA	0.25 (0.10, 0.42)
Proven necrotising enterocolitis				
Prevalence (%), N	3.2, 656	2.8, 318	0.0, 470	7.7, 1220
ICC (95% CI)	0.17 (-0.04, 0.95)	NA	NA	0.08 (-0.07, 0.42)
Required continuous positive airway pressure				
Prevalence (%), N	75.6, 655	55.7, 316	0.6, 470	95.4, 1272
ICC (95% CI)	0.49 (0.29, 0.71)	0.56 (0.22, 0.83)	NA	0.21 (-0.02, 0.70)

Required intermittent positive pressure ventilation				
Prevalence (%), N	52.8, 655	65.5, 316	1.1, 470	78.9, 1272
ICC (95% CI)	0.68 (0.51, 0.83)	0.98 (0.69, 1.00)	NA	0.55 (0.36, 0.79)
Bronchopulmonary dysplasia[a]				
Prevalence (%), N	22.1, 653	NA	NA	47.5, 1149
ICC (95% CI)	0.46 (0.26, 0.72)	NA	NA	0.33 (0.17, 0.49)
Discharged home on oxygen				
Prevalence (%), N	10.1, 652	9.7, 319	0.0, 470	NA
ICC (95% CI)	0.21 (0.02, 0.59)	-0.12 (-0.28, 0.00)	NA	NA

NA, not applicable due to insufficient number of cases/non-cases or data not being collected; N, number of infants with available data; ICC, intracluster correlation coefficient; CI, confidence interval; DINO, Docosahexaenoic Acid for the Improvement of Neurodevelopmental Outcomes in Preterm Infants; BCD, Bottles, Cups and Dummies; TTB, Twins Timing of Birth; N3RO, n-3 Fatty Acids for Improvement in Respiratory Outcomes.

[a] Defined as supplemental oxygen required at 36 weeks' postmenstrual age for the DINO Trial, and supplemental oxygen and/or respiratory support required with an assessment of oxygen saturation at 36 weeks' postmenstrual age or discharge home, whichever occurred first, for the N3RO Trial.

Figure 1. Relationship between the intracluster correlation coefficient (ICC) and the design effect for a continuous outcome by method of randomising twins (cluster randomisation, individual randomisation or randomisation to opposite treatment groups) when (A) 3% or (B) 20% of mothers have a twin birth. Black and grey lines indicate scenarios where the independence and exchangeable working correlation structures will be assumed in the analysis, respectively.