STATISTICS IN MEDICINE
Statist. Med. 2001; **20**:3729–3740 (DOI: 10.1002/sim.1115)

Current and future challenges in the design and analysis of cluster randomization trials‡

Neil Klar[1,*,†] and Allan Donner[2]

[1]*Division of Preventive Oncology, Cancer Care Ontario, Toronto, Ontario, M5G 2L7, Canada*
[2]*Department of Epidemiology and Biostatistics, The University of Western Ontario, London, Ontario, N6A 5C1, Canada*

SUMMARY

Randomized trials in which the unit of randomization is a community, worksite, school or family are becoming widely used in the evaluation of life-style interventions for the prevention of disease. The increasing interest in adopting a cluster randomization design is being matched by rapid methodological developments. In this paper we describe several of these developments. Brief mention is also made of issues related to economic analysis and to the planning and conduct of meta-analyses for cluster randomization trials. Recommendations for reporting are also discussed. Copyright © 2001 John Wiley & Sons, Ltd.

1. INTRODUCTION

Randomized trials in which the unit of randomization is a community, worksite, school or family are becoming increasingly common for the evaluation of life-style interventions for the prevention of disease. Reasons for adopting cluster randomization are diverse, but include administrative convenience, a desire to reduce the effect of treatment contamination and the need to avoid ethical issues which might otherwise arise.

Dependencies among cluster members typical of such designs must be considered when determining sample size and in the subsequent data analyses. Failure to adjust standard statistical methods for within-cluster dependencies will result in underpowered studies with spuriously elevated type I errors.

These statistical features of cluster randomization were not brought to wide attention in the health research community until the now famous article by Cornfield [1]. However, the 1980s saw a dramatic increase in the development of methods for analysing correlated outcome data (for example, Ashby *et al.* [2]) in general and methods for the design and analysis of cluster

*Correspondence to: Neil Klar, Division of Preventive Oncology, 12th Floor, Cancer Care Ontario, Toronto, Ontario, M5G 2L7, Canada.
† E-mail: neil.klar@cancercare.on.ca
‡Presented at the International Society for Clinical Biostatistics, Twenty-first International Meeting, Trento, Italy, September 2000.

Contract/grant sponsor: Natural Sciences and Engineering Council of Canada

Received January 2001
Accepted August 2001

randomized trials in particular (for example, Donner *et al.* and Gillum *et al.* [3, 4]). Books summarizing this research have also recently appeared [5, 6] and new statistical methods are in constant development. There is also growing evidence that there is a shorter lag time between the development of new methodology and its appearance in medical journals [7].

In this paper we consider several new methodological developments likely to be of particular importance to the design and analysis of cluster randomization trials. Some of the ethical challenges posed by cluster randomization are described in Section 2 while new developments in sample size estimation and data analysis are described in Sections 3 and 4, respectively. Sections 5 and 6 of the paper consider issues in economic analysis and in the meta-analysis of cluster randomization trials, while guidelines for trial reporting are reviewed in Section 7. The last section also summarizes many of the key issues raised in the paper.

2. ETHICAL ISSUES IN CLUSTER RANDOMIZATION

According to the World Medical Association Declaration of Helsinki [8], consent must be obtained from each patient prior to random assignment. The situation is more complicated for cluster randomization trials, particularly when larger units such as schools, communities or worksites are randomized. In that case school principals, community leaders or other decision makers will usually provide permission for both the random assignment and implementation of the intervention. Individual study subjects must still be free to withhold their participation, although even then they may not be able to completely avoid the inherent risks imposed by an intervention that is applied on a cluster-wide level.

The relative absence of ethical guidelines for cluster randomized trials [9, 10] appears to have created a research environment in which the choice of randomization unit may determine whether or not informed consent is deemed necessary prior to random assignment. This phenomenon can be seen, for example, in the several published trials of vitamin A supplementation on childhood mortality. Informed consent was obtained from mothers prior to assigning children to either vitamin A or placebo in the household randomization trial reported by Herrera *et al.* [11]. This was not the case in the community intervention trial of vitamin A reported by the Ghana VAST Study Team [12], where consent to participate was likely obtained from children's parents only after random assignment of clusters to one of two intervention groups. In this trial clusters were defined to be geographic areas which included approximately 51 family compounds each. It seems questionable, on both an ethical and methodological level, whether the unit of randomization should play such a critical role in deciding whether or not informed consent is required prior to randomization.

Further debate is needed in order for the research community to be able to develop reasonably uniform standards regarding guidelines for informed consent. We encourage investigators to report the methods used to obtain informed consent in their own trials as a first step towards engaging this debate.

3. RECENT DEVELOPMENTS IN SAMPLE SIZE ESTIMATION

We consider a cluster randomization trial in which the primary aim is to compare two groups with respect to their mean values on a normally distributed response variable Y having a common but unknown variance σ^2. Suppose k clusters of m individuals are assigned to each

of an experimental group and a control group. Estimates of the population means μ_1 and μ_2 are given by the usual sample means $\bar{Y}_1$ and $\bar{Y}_2$ for the experimental and control groups, respectively. From a well-known result in cluster sampling (for example, Kish, reference [13], Chapter 5), the variance of each of these means is given by

$$\text{var}(\bar{Y}_i) = \frac{\sigma^2}{km}[1 + (m-1)\rho], \quad i = 1, 2 \tag{1}$$

where ρ is the intracluster correlation coefficient measuring the degree of similarity among responses within a cluster. If σ^2 is replaced by $P(1-P)$, where P denotes the probability of a success, equation (1) also provides an expression for the variance of a sample proportion under clustering. For sample size determination equation (1) implies that the usual estimate of the required number of individuals in each group should be multiplied by the variance inflation factor or design effect $\text{IF} = 1 + (m-1)\rho$ to provide the same statistical power as would be obtained by randomizing km individuals to each group when there is no clustering effect.

The expression for IF shows that the prior assessment of ρ has a unique role to play in estimating sample size for cluster randomization trials. Difficulties in obtaining accurate estimates of intracluster correlation are complicated in practice by the relatively small number of publications which present these values when reporting trial results (see, for example, Donner and Klar, reference [5], Table 5.1). However several new strategies have been suggested to improve sample size planning. For example, Spiegelhalter [14] suggests using Bayesian methods which explicitly allow the use of prior opinion in trial design. An alternative approach is to recalculate sample size using an internal pilot study, as has been suggested by Shih [15] in the context of a periodontal study in which clusters are composed of multiple sites in a subject's mouth. However, this approach can only be used when clusters are recruited over time and follow-up time is short enough that outcome data may be used in a reasonably timely fashion to update the estimate of ρ. Existing research as applied to clinical trials involving individual randomization demonstrates that sample size re-estimation has little effect on type I error rates for trials which are sufficiently large [16]. It is unclear what effect, if any, sample size re-estimation has on type I error rates for cluster randomization trials.

4. RECENT DEVELOPMENTS IN DATA ANALYSIS

4.1. Population-averaged versus cluster-specific models

A difficult issue in the analysis of categorical, count and time to event outcome data involves the decision to select either a 'population-averaged' model or a 'cluster-specific' model. Consider, for example, models for the analysis of correlated binary outcome data. The generalized estimating equations (GEE) extension of logistic regression may be characterized as population-averaged in the sense that it measures the expected (marginal) change in a response as the value of the covariate increases by one unit. This is in contrast to approaches based on a logistic-normal model, which may be characterized as cluster specific. The latter models are constructed so as to measure the expected change in response within a cluster as the value of a covariate increases by one unit, that is, they provide conditional measures of covariate effects.

Although both approaches estimate the same population parameters when the outcome variable is normally distributed, this equivalence disappears in the case of a binary outcome variable, in which case

$$\alpha_{PA} \approx \beta_{CS}(1 - \rho_0)$$

where α_{PA} and β_{CS} are the regression coefficients from a population-averaged and cluster-specific extension of logistic regression, respectively, while ρ_0 is the intracluster correlation coefficient at $\alpha_{PA} = \beta_{CS} = 0$. As pointed out by Neuhaus [17], interpretation of estimated covariate effects obtained from cluster-specific models may be difficult when the covariate is defined at the cluster level. This problem arises in the interpretation of results from cluster randomization trials when the covariate of main interest is the intervention effect, since subjects in any one cluster will invariably share the same intervention group. Thus interpretation of the intervention effect using a cluster-specific model must formally rely on the notion of a subject within a given cluster changing his or her intervention status, clearly a non-observable event. This difficulty has led Neuhaus [17] to remark that cluster-specific models would seem to be most suitable for testing the effect of covariates that vary within clusters (for example, subject age or gender), while population-averaged models such as GEE are conceptually preferable for estimating the effect of cluster-level covariates such as intervention status. However it must also be noted that the differences between the two approaches disappear as the intracluster correlation coefficient approaches zero, and that more empirical work is needed to compare the advantages and disadvantages of the two approaches in practice. As pointed out by Omar and Thompson [18], one advantage of the cluster-specific approach is that it provides direct estimates of variance components, quantities which are treated as nuisance parameters when the population-averaged approach is adopted. Further mathematical details concerning the distinction between cluster-specific and population-averaged models are provided in the Appendix.

The correct specification of random effects at two or more levels may have implications for the validity and precision of statistical inferences [19], thus the choice of approach should ultimately depend on the covariate effects and other parameters that are of most interest. These issues have been debated recently by Heagerty and Zeger [20], who describe a model which can be used to obtain either marginal or conditional measures of covariate effects but which also provides direct estimates of variance components. Further work is needed to evaluate this model in the context of cluster randomization trials.

4.2. Methods for analysing trials involving a small number of clusters

Many cluster randomization trials, particularly those that randomize communities, involve a small number of large clusters. However, the validity of statistical inferences constructed using multiple regression models (for example, generalized estimation equations approach, logistic-normal) require a large number of clusters. Suppose that only a small number of clusters are enrolled in a trial. Then, for example, robust hypothesis tests will tend to be overly liberal when constructed using the GEE extension of linear regression for the analysis of correlated Gaussian data [21]. Furthermore statistical inferences constructed using mixed effects linear regression models must be based on approximate t- or F-distributions for which there is no universally accepted method for approximating the degrees of freedom [22], at least in the case of an unbalanced design. A somewhat *ad hoc* use of the F-distribution was proposed by

Mancl and DeRouen [23], in their simulation study examining the performance of the GEE extension of logistic regression. Mancl and DeRouen [23] reduced the overly liberal rejection rates of robust hypothesis test using this procedure while also incorporating a bias correction for the robust variance estimator.

These difficulties may be avoided using cluster-level statistical inferences based on the randomization distribution used in designing the trial. For example, Gail *et al.* [24], describe how exact hypothesis tests and exact confidence intervals were used to construct statistical inferences for COMMIT, a pair-matched community intervention trial which assessed a smoking cessation intervention [25]. Note that such analyses give equal weight to all clusters, ignoring the variation in cluster size.

These problems suggest that additional research is required on the development of methodology for trials enrolling only a small number of clusters. Of course, more careful attention to trial power will help to reduce these challenges in practice.

4.3. Analytic issues involving matched-pair designs

An important design feature of the COMMIT trial is that the 22 participating communities were pair-matched prior to random assignment. Pairs of communities were matched on the basis of population size, population density, demographic profile, community structure and geographical proximity. The matching variables were selected, at least in part, on the basis of their known correlation with smoking cessation rates. The main advantage of this design is its potential to provide very tight and explicit balancing of important prognostic factors, thereby improving the power for detecting the effect of intervention. Donner and Klar (reference [5], Table 3.2) use data from seven pair-matched trials to show that it has been quite difficult in practice to identify matching variables which will substantially improve power. This was true also of the COMMIT trial which was only able to achieve a modest gain in efficiency as compared to a trial using a completely randomized design.

However even when it is possible to create comparable pairs of clusters, there are some less recognized analytic limitations associated with the matched-pair design. These limitations arise because of the inherent feature of this design that there is exactly one cluster assigned to each combination of intervention and stratum. As a result, the natural variation in response between clusters in a matched pair is totally confounded with the effect of intervention. Estimates of the variance for the observed effect of intervention must therefore be constructed using between stratum information [26]. A resulting consequence is that standard models for correlated binary outcome data are not directly applicable. Several newer methods might prove useful but need to be evaluated before they can be recommended for analyses of data from pair-matched cluster randomization trials. For example, Liang and Pulver [27] describe an adaptation of the generalized estimating equations extension of logistic regression for analyses of genetic data from pair-matched families. New binary outcomes are created by combining all possible pairs of individuals from control and experimental clusters within a stratum into new ‘pseudo’ clusters. Application of robust variance estimators are then based on between-stratum information. An alternative procedure based on mixed effects logistic regression was adopted by Sorensen *et al.* [28] and outlined by Thompson *et al.* [29]. Variance estimators for these models are constructed using a random effect obtained from the interaction between stratum and intervention.

4.4. Overlapping cluster membership

Quit rates of heavy smokers enrolled in the COMMIT trial during 1988 were assessed in 1993. A potential problem arising in trials having such a relatively long follow-up time is the increased possibility that cluster membership may change. This may particularly be a problem in school-based trials since children could enrol in different schools in successive years. However very little attention has been given to statistical models which may be used when subjects can enrol in more than one cluster over time. Some discussion of models which include such non-nested sources of clustering is given by Raudenbush [30] for Gaussian outcome data and by Betensky *et al*. [31] for binary outcome data. Further work is needed to evaluate the usefulness of these models for cluster randomization trials. The effect on statistical inferences of making the typical but possibly incorrect assumption that all subjects maintain membership in the same cluster is as yet unclear.

4.5. Issues arising from missing data, protocol violations and outliers

Considerable attention has been given recently to methods of accounting for missing data, patient non-compliance, and the effect of outliers in trials randomizing independent individuals. Extensions to multilevel data in general [32, 33] and to cluster randomization trials in particular are also beginning to appear [24, 29, 34–36]. For example, an application of imputation methods is provided by the COMMIT investigators [24, 25]. Since analyses were conducted at the cluster level using permutation tests it was not necessary in this case to correct variance estimates for imputation. An alternative approach based on a strategy of multiple imputation was used by Gomel *et al*. [35] in their anlaysis of data from an Australian worksite intervention trial. There has been some recent work examining the properties of these different procedures. For instance, Hunsberger *et al.* [36] report on the results of a simulation study in the context of a school-based obesity prevention programme demonstrating that multiple imputation can perform adequately even when the missingness is related to intervention group and study outcome.

5. ISSUES IN THE ECONOMIC ANALYSIS OF CLUSTER RANDOMIZATION TRIALS

Decisions regarding implementation of a new therapeutic intervention are increasingly dependent on the demonstration of its cost-effectiveness as well as its efficacy. Economic analyses may be carried out by including measures of costs along with other outcomes as part of a randomized trial. A common criticism of such a strategy is that the costs of doing research may be measured in place of the costs of providing an intervention, a problem primarily of external validity (Drummond *et al.* reference [37], Section 8.2). This criticism may have less force, however, for cluster randomization trials. Consider, for example, an antenatal care trial sponsored by the World Health Organization which compared the ‘best standard treatment’ offered to women attending antenatal care clinics to an experimental intervention consisting only of tests, clinical activities and follow-up actions scientifically demonstrated to be effective in improving outcomes [38]. A secondary goal of this trial was to conduct an economic analysis comparing costs and cost-effectiveness of the experimental intervention to that of the control intervention. Since antenatal care clinics are the unit of randomization, all women

from a clinic will receive a single intervention. Thus, as noted by Mugford *et al.* [39], an 'advantage of the cluster design for the economic evaluation is that the unit of randomization is also a key unit of health management and cost generation'.

Economic analyses have traditionally been conducted using descriptive comparisons of cost outcomes. Any uncertainty in costs may then by explored using sensitivity analyses [37]. This approach was taken, for example, by Aikins *et al.* [40] in their economic analysis of data from a pair-matched community intervention trial assessing the effect of insecticide impregnated bednets on preventing morbidity and mortality from malaria. Uncertainty in the costs of insecticide, in patient treatment and of numbers of cases seeking treatment were varied as part of a multi-way sensitivity analysis.

Drummond *et al.* [37] argue that it is often useful to supplement such sensitivity analyses with standard statistical inferences comparing costs across intervention groups. Standard procedures for the analysis of correlated quantitative outcomes are applicable when comparing costs across intervention groups. For example, analyses of cost data from trials in which a large number of clusters are enrolled may proceed using mixed effects linear regression models allowing for random variation in costs among clusters. The typically skewed distribution of cost data may be accounted for here using robust variance estimators in place of model-based inferences.

Analyses may prove more complicated when attempts are made to evaluate cost-effectiveness, which is typically measured as the ratio of the difference in costs across intervention groups divided by the estimated clinically relevant effect of intervention [37]. Power and sample size calculations for the cost-effectiveness ratio have been described by Briggs and Gray [41] in the context of therapeutic trials for which individual patients are the unit of randomization, while approaches to statistical analysis have been described by Drummond *et al.* [37]. Their extension to cluster randomized trials is as yet unexplored.

6. ISSUES IN THE META-ANALYSIS OF CLUSTER RANDOMIZATION TRIALS

Meta-analyses involving the synthesis of evidence from cluster randomization trials raise methodologic issues beyond those raised by meta-analyses which include only individually randomized trials. Two of the more challenging methodological issues are (i) the increased possibility of study heterogeneity, and (ii) difficulties in estimating design effects and selecting an optimal method of analysis [42].

These challenges are illustrated by the meta-analysis reported by Fawzi *et al.* [43] who investigated the effect of vitamin A supplementation on child mortality. This investigation considered trials of hospitalized children with measles as well as community-based trials of healthy children. Individual children were assigned to intervention in the four hospital-based trials, while allocation was by geographic area, village, or household in the eight community-based trials. One of the community-based trials included only one geographic area per intervention group, each of which enrolled approximately 3000 children. On the other hand there was an average of about two children from each cluster when allocation was by household. Thus an important source of heterogeneity arose from the nature and size of the randomization units used in the different trials. This problem was dealt with by performing the meta-analysis separately for the individually randomized and cluster randomized trials.

It is straightforward to summarize results across trials when each study provides a common measure for the estimated effect of intervention (such as an odds ratio, for example) and a corresponding variance estimate which appropriately accounts for the clustering. Unfortunately the information necessary for its application in practice is rarely available to meta-analysts.

One consequence of this difficulty is that investigators are sometimes forced to adopt *ad hoc* strategies when relying on published trial reports which fail to provide estimates of the design effect. For example, only four of the eight community-based trials considered by Fawzi *et al*. [43] reported that they accounted for clustering effects. Consequently Fawzi *et al*. [43] decided to increase the variance of summary odds ratio estimates by an arbitrary 30 per cent. The authors argued that this adjustment seems reasonable since the design effects ranged from 1.10 to 1.40 in those studies which did adjust for clustering effects.

A few investigators have designed community intervention trials in which exactly one cluster has been assigned to the intervention group and one to the control group, either with or without the benefit of random assignment. This was the case, as noted above, with at least one of the community-based trials included in the meta-analysis reported by Fawzi [43]. Such trials invariably result in interpretational difficulties arising from the total confounding of the variation in response due to the effect of intervention and the natural variation that exists between communities even in the absence of an intervention effect. External estimates of design effects must be used if such trials are to be included in a meta-analysis. It would therefore be prudent to conduct a sensitivity analysis in which these trials are excluded in order to assess their influence on the conclusions.

Even when each trial provides an estimate of the design effect, several different approaches could be used for conducting a meta-analysis. For example, a procedure commonly adopted for combining the results of individually randomized clinical trials with a binary outcome variable is the well known Mantel–Haenszel test. The adjusted Mantel–Haenszel test [44] may be used to combine results of cluster randomized trials. Donner *et al*. [42] review this procedure and a number of other approaches, assuming that each of the combined trials involves a completely randomized design with a binary outcome variable. Simulation studies are needed to evaluate these different analytic approaches.

7. RECOMMENDATIONS FOR TRIAL REPORTING

Reporting standards for randomized clinical trials have now been widely disseminated (for example, Moher *et al*. [45]). Many of the principles that apply to trials randomizing individuals also apply to trials randomizing intact clusters. These include a carefully posed justification for the trial, a clear statement of the study objectives, a detailed description of the planned intervention and the method of randomization and an accurate accounting of all subjects randomized to the trial. Unambiguous inclusion–exclusion criteria must also be formulated, although perhaps separately for cluster-level and individual-level characteristics. There are, however, some unique aspects of cluster randomization trials that require special attention at the reporting stage. We focus here on some of the most important of these. More complete accounts are provided by Donner and Klar (reference [5], Chapter 9) and Elbourne and Campbell [46].

The decreased statistical efficiency of cluster randomization relative to individual randomization can be substantial, depending on the sizes of the clusters randomized and the degree of

intracluster correlation. Thus, unless there is obviously no alternative, the reasons for randomizing clusters rather than individuals should be clearly stated. This information, accompanied by a clear description of the units randomized, can help a reader decide if the loss of precision due to cluster randomization is in fact justified. Torgerson [47] argues that this is often not the case in practice.

Having decided to randomize clusters, investigators may still have considerable latitude in their choice of allocation unit. As different levels of statistical efficiency are associated with different cluster sizes, it would seem important to select the unit of randomization on a carefully considered basis. An unambiguous definition of the unit of randomization is also required. For example, a statement that 'neighbourhoods' were randomized is clearly incomplete without a detailed description of this term in the context of the planned trial.

As noted previously, the consensus that exists in most clinical trial settings regarding the role of informed consent has not tended to apply to cluster randomization trials. By reporting the methods used (if any) to obtain informed consent in their own trials, it may gradually become possible for the research community to develop reasonably uniform standards regarding this important issue.

The clusters that participate in a trial, simply owing to their consent to be randomized, may not be representative of the target population of clusters. Some indication of this lack of representativeness may be obtained by listing the number of clusters that meet the eligibility criteria for the trial, but which decline to participate, along with a description of their characteristics.

A continuing difficulty with reports of cluster randomization trials is that justification for the sample size is all too often omitted. Investigators should clearly describe how the sample size for their trial was determined, with particular attention given to how clustering effects were adjusted for. This description should be in the context of the experimental design selected (for example, completely randomized, matched-pair, stratified).

It would also be beneficial to the research community if empirical estimates of ρ were routinely published, with an indication of whether or not the reported values have been adjusted for the effect of baseline covariates.

It should be further specified what provisions were made in the sample size calculations to account for potential loss of follow-up. Since the forces leading to the loss of follow-up of individual members of a cluster may be very different from those leading to the loss of an entire cluster, both types of attrition must be considered here.

A large variety of methods, based on very different sets of assumptions, have been used to analyse data arising from cluster randomization trials. For example, possible choices for the analysis of binary outcomes include adjusted chi-square statistics, the method of generalized estimating equations (GEE) and logistic-normal regression models. These methods are not as familiar as the standard procedures used to analyse clinical trial data, partly because methodology for analysing cluster randomization trials is in a state of rapid development, with virtually no standardization and a proliferation of associated software. Therefore it is incumbent on authors to provide a clear statement of the statistical methods used, accompanied, where it is not obvious, by an explanation of how the analysis adjusts for the effect of clustering. The software used to implement these analyses should also be reported.

APPENDIX

Consider a trial in which clusters are assigned to either an experimental or a control group using a completely randomized design. Suppose also that Y_{ijs} denotes a binary outcome variable for the sth subject, $s = 1, \ldots, m_{ij}$, from the jth cluster, $j = 1, \ldots, k_i$, and ith intervention group, $i = 1, 2$.

Then the effect of intervention may be assessed using a generalized estimating equations extension of logistic regression given by

$$\log[P_{ijs}/(1 - P_{ijs})] = \alpha_0 + \alpha_{\mathrm{PA}} X_{ij}$$

where

$$P_{ijs} = \Pr(Y_{ijs} = 1 \mid X_{ij})$$
$$X_{ij} = \begin{cases} 1 & \text{if } i = 1 \text{ (experimental)} \\ 0 & \text{if } i = 2 \text{ (control)} \end{cases}$$

The population-averaged odds ratio for the effect of intervention is given by $\exp(\alpha_{\mathrm{PA}})$.

The effect of intervention may also be assessed using a logistic-normal model given by

$$\log[P_{ijs}/(1 - P_{ijs})] = \beta_0 + \beta_{\mathrm{CS}} X_{ij} + \varepsilon_{ij}$$

where

$$P_{ijs} = \Pr(Y_{ijs} = 1 \mid \varepsilon_{ij}, X_{ij})$$
$$\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$$

and assuming that the random effects, ε_{ij}, $i = 1, 2$, $j = 1, \ldots, k_i$, are independent. The cluster-specific odds ratio for the effect of intervention is given by $\exp(\beta_{\mathrm{CS}})$.

The covariate X_{ij} denoting the intervention assignment is defined at the cluster level. These models may include covariates defined at either the cluster level or at the individual level. For example, Donner and Klar [5] use the generalized estimating equations extension of logistic regression to analyse data from a school-based smoking prevention trial. Baseline measures of student gender and age were included in the model to account for chance imbalance across intervention groups with respect to these variables. The resulting population-averaged model is given by

$$\log[P_{ijs}/(1 - P_{ijs})] = \alpha_0 + \alpha_1 X_{ij} + \alpha_2 \mathrm{Age}_{ijs} + \alpha_3 \mathrm{Sex}_{ijs}$$

where

$$P_{ijs} = \Pr(Y_{ijs} = 1 \mid X_{ij}, \mathrm{Age}_{ijs}, \mathrm{Sex}_{ijs})$$
$$\mathrm{Sex}_{ijs} = \begin{cases} 1 & \text{if } i = 1 \text{ if male} \\ 0 & \text{if } i = 2 \text{ if female} \end{cases}$$

and Age_{ijs} denotes age in years.

ACKNOWLEDGEMENTS

The authors' work was partially supported by grants from the Natural Sciences and Engineering Council of Canada.

REFERENCES

1. Cornfield J. Randomization by group: a formal analysis. *American Journal of Epidemiology* 1978; **108**:100–102.
2. Ashby M, Neuhaus JM, Hauck WW, Bacchetti P, Heilbron DC, Jewell NP, Segal MR, Fusaro RE. An annotated bibliography of methods for analyzing correlated categorical data. *Statistics in Medicine* 1992; **11**:67–99.
3. Donner A, Birkett N, Buck C. Randomization by cluster: sample size requirements and analysis. *American Journal of Epidemiology* 1981; **114**:906–914.
4. Gillum RF, Williams PT, Sondik E. Some consideration for the planning of total-community prevention trials: when is sample size adequate? *Journal of Community Health* 1980; **5**:270–278.
5. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold: London, 2000.
6. Murray DM. *Design and Analysis of Community Trials*. Oxford University Press: Oxford, 1998.
7. Altman DG, Goodman SN. Transfer of technology from statistical journals to the biomedical literature, past trends and future predictions. *Journal of the American Medical Association* 1994; **272**:129–132.
8. Christie B. Doctors revise Declaration of Helsinki. *British Medical Journal* 2000; **321**:913.
9. Edwards SJL, Braunholtz DA, Lilford RJ, Stevens AJ. Ethical issues in the design and conduct of cluster randomised controlled trials. *British Medical Journal* 1999; **318**:1407–1409.
10. Hutton JL. Are distinctive ethical principles required for cluster randomized controlled trials? *Statistics in Medicine* 2001; **20**:473–488.
11. Herrera MG, Nestel P, El Amin A, Fawzi WW, Muhammad KA, Weld L. Vitamin A supplementation and child survival. *Lancet* 1992; **340**:267–271.
12. Ghana VAST Study Team. Vitamin A supplementation in northern Ghana: effects on clinic attendances, hospital admissions, and child mortality. *Lancet* 1993; **342**:7–12.
13. Kish L. *Survey Sampling*. Wiley: New York, 1965.
14. Spiegelhalter DJ. Bayesian methods for cluster randomised trials with continuous responses. *Statistics in Medicine* 2001; **20**:435–452.
15. Shih WJ. Sample size and power calculations for periodontal and other studies with clustered samples using the method of generalized estimation equations. *Biometrical Journal* 1997; **39**:899–908.
16. Keiser M, Friede T. Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Statistics in Medicine* 2000; **19**:901–911.
17. Neuhaus JM. Statistical methods for longitudinal and clustered designs with binary responses. *Statistical Methods in Medical Research* 1992; **1**:249–273.
18. Omar RZ, Thompson SG. Analysis of a cluster randomized trial with binary outcome data using a multi-level model. *Statistics in Medicine* 2000; **19**:2675–2688.
19. Ten Have TR, Morabia A. A comparison of mixed effects logistic regression models for binary response data with two nested levels of clustering. *Statistics in Medicine* 1999; **18**:947–960.
20. Heagerty PJ, Zeger SL. Marginalized multilevel models and likelihood inference. *Statistical Science* 2000; **15**:1–26.
21. Feng Z, McLerran D, Grizzle J. A comparison of statistical methods for clustered data analysis with Gaussian data. *Statistics in Medicine* 1996; **15**:1793–1806.
22. Verbeke G. Linear mixed models for longitudinal data. In *Linear Mixed Models in Practice*, *A SAS-Oriented Approach*, Verbeke G, Molenberghs G (eds). Springer-Verlag, Inc.:New York, 1997; chapter 3.
23. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. *Biometrics* 2001; **57**:126–134.
24. Gail MH, Mark SD, Carroll RJ, Green SB, Pee D. On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine* 1996; **15**:1069–1092.
25. COMMIT Research Group. Community Intervention Trial for Smoking Cessation (COMMIT): I. Cohort results from a four-year community intervention. *American Journal of Public Health* 1995; **85**:183–192.
26. Klar N, Donner A. The merits of matching in community intervention trials. *Statistics in Medicine* 1997; **16**:1753–1764.
27. Liang KY, Pulver AE. Analysis of case-control/family sampling design. *Genetic Epidemiology* 1996; **13**: 253–270.
28. Sorensen G, Thompson B, Glanz K, Feng Z, Kinne S, DiClemente C, Emmons K, Heimendinger J, Probart C, Lichtenstein E. Work site-based cancer prevention: primary results from the Working Well Trial. *American Journal of Public Health* 1996; **86**:939–947.

29. Thompson SG, Pyke SDM, Hardy RJ. The design and analysis of paired cluster randomized trials: an application of meta-analysis techniques. *Statistics in Medicine* 1997; **16**:2063–2079.
30. Raudenbush SW. A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics* 1993; **18**:321–349.
31. Betensky RA, Talcott JA, Weeks JC. Binary data with two, non-nested sources of clustering: an analysis of physician recommendations for early prostate cancer treatment. *Biostatistics* 2000; **1**:219–230.
32. Langford IH, Lewis T. Outliers in multilevel data. *Journal of the Royal Statistical Society*, *Series A* 1998; **161**:121–160.
33. Ziegler A, Blettner M, Kastner C, Chang-Claude J. Identifying influential families using regression diagnostics for generalized estimating equations. *Genetic Epidemiology* 1998; **15**:341–353.
34. Korhonen P, Loeys T, Goetghebeur E, Palmgren J. Vitamin A and infant mortality: beyond intention-to-treat in a randomized trial. *Lifetime Data Analysis* 2000; **6**:107–121.
35. Gomel MK, Oldenburg B, Simpson JM, Chilvers M, Owen N. Composite cardiovascular risk outcomes of a work-site intervention trial. *American Journal of Public Health* 1997; **87**:673–676.
36. Hunsberger S, Murray D, Davis CE, Fabsitz RR. Imputation strategies for missing data in a school-based multi-center study: the Pathways study. *Statistics in Medicine* 2001; **20**:305–316.
37. Drummond MF, O'Brien B, Stoddart GL, Torrance GW. *Methods for the Economic Evaluation of Health Care Programmes.* 2nd edn. Oxford University Press: Oxford, 1998.
38. Villar J, Bakketeig L, Donner A, Al-Mazrou Y, Ba aqeel H, Belizn, JM, Carroli G, Farnot U, Lumbiganon P, Piaggio G, Berendes H. The WHO antenatal care randomised trial: rationale and study design. *Paediatric and Perinatal Epidemiology* 1998; **12 Suppl. 2**:27–58.
39. Mugford M, Hutton G, Fox-Rushby JF, for the WHO Antenatal Care Trial Research Group. Methods for economic evaluation alongside a multicentre trial in developing countries: a case study from the WHO Antenatal Care Randomised Trial. *Paediatric and Perinatal Epidemiology* 1998; **12 Suppl. 2**:75–97.
40. Aikins MK, Fox-Rushby J, D'Alessandro U, Langerock P, Cham K, New L, Bennett S, Greenwood B, Mills A. The Gambian national impregnated bednet programme: costs, consequences and net cost-effectiveness. *Social Science and Medicine* 1998; **46**:181–191.
41. Briggs AH, Gray AM. Power and sample size calculations for stochastic cost-effectiveness analysis. *Medical Decision Making* 1998; **18 (Suppl. 2)**:s81–s92.
42. Donner A, Piaggio G, Villar J. Statistical methods for the meta-analysis of cluster randomization trials. *Statistical Methods in Medical Research* 2001; **10**:325–338.
43. Fawzi WW, Chalmers TC, Herrera MG, Mosteller F. Vitamin A supplementation and child mortality, a meta-analysis. *Journal of the American Medical Association* 1993; **269**:898–903.
44. Donner A. Some aspects of the design and analysis of cluster randomization trials. *Applied Statistics* 1998; **47**:95–114.
45. Moher D, Schulz KF, Altman DG, for the CONSORT Group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* 2001; **357**:1191–1194.
46. Elbourne D, Campbell M. Extending the CONSORT statement to cluster randomised trials: for discussion. *Statistics in Medicine* 2001; **20**:489–496.
47. Torgerson DJ. Contamination in trials: is cluster randomisation the answer? *British Medical Journal* 2001; **322**:355–357.

