

Issues in the meta-analysis of cluster randomized trials

Allan Donner^{1,*},† and Neil Klar²

¹*Department of Epidemiology and Biostatistics, The University of Western Ontario, London, Ontario, N6A 5C1, Canada*

²*Division of Preventive Oncology, Cancer Care Ontario, Toronto, Ontario, M5G 2L7, Canada*

SUMMARY

Meta-analyses involving the synthesis of evidence from cluster randomization trials are being increasingly reported. These analyses raise challenging methodologic issues beyond those raised by meta-analyses which include only individually randomized trials. In this paper we review and comment on a selected number of these issues, including problems of study heterogeneity, difficulties in estimating design effects from individual trials and the choice of statistical methods. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: community trials; intraclass correlation; clustering; correlated binary data

1. INTRODUCTION

There has now accumulated a vast literature on the methodologic challenges involved in planning and conducting a meta-analysis, including entire issues of leading journals [1–3]. Many books on meta-analysis have also been published, as evidenced by a recent review of 14 such texts [4]. Much of this literature focuses on the meta-analysis of randomized clinical trials, where it is almost invariably assumed that the unit of randomization is the individual subject. However, in recent years there has been a growing number of trials reported which have randomized intact social units of individuals to intervention groups. These trials, known as cluster (or group) randomized trials are particularly widespread in the evaluation of health care, screening and educational interventions [5]. Therefore meta-analysts increasingly face the need to include these studies among the set of trials considered relevant to the question at hand.

We discuss the special issues raised by the requirement to consider both individually randomized and cluster randomized trials in a planned meta-analysis.

* Correspondence to: Allan Donner, Department of Epidemiology and Biostatistics, The University of Western Ontario, London, Ontario, N6A 5C1, Canada.

† E-mail: donner@biostats.uwo.ca

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

Copyright © 2002 John Wiley & Sons, Ltd.

2. STUDY HETEROGENEITY

Study heterogeneity involving different populations, interventions delivered, follow-up periods etc. is frequently raised as an issue in the interpretation of reported meta-analyses [6]. This issue may be even more severe in meta-analyses involving cluster randomization trials, where the meta-analyst must confront several additional potential sources of heterogeneity. These include differences related to choice of study design (for example, matched-pair versus completely randomized), the nature of the randomization unit (for example, worksites versus households) and the sizes of the clusters randomized. For example, prevention trials of vitamin supplementation have been conducted using several different units of allocation, including individual children [7], households [8] and entire communities [9]. Moreover, while clinical trials involving matched-pair individuals are relatively infrequent, matched-pair cluster randomization designs are very common, particularly in the evaluation of interventions offered at the community level. The interventions evaluated in cluster randomization trials also tend to be relatively complex and diversified, particularly in the educational and health care fields. Eligibility criteria for these trials might vary not only at the cluster level (the unit of randomization) but also at the level of the individual.

The methodological quality of cluster randomized trials is also quite diverse. For example, it has been observed [10, 11] that about half of such trials ignore the clustering aspect of the design in the statistical analysis, and about two-thirds in the estimation of trial power. There is also likely to be a secular factor involved here; trials designed in the last decade or so may be more rigorous methodologically than trials designed in an earlier era.

These additional sources of heterogeneity raise important analytic issues, as recognized by those investigators who have performed separate meta-analyses on trials that involve very different randomization units. An example is provided by Fawzi *et al.* [12], who investigated the effect of vitamin A supplementation on child mortality. This investigation considered studies of hospitalized children with measles, as well as other studies involving healthy children participating in community-based trials. Individual children were assigned to intervention in the four hospital-based trials, while allocation was by village, district or household in the eight community-based trials. Hence the meta-analysis was performed in two parts, part 1 on the hospital-based studies and part 2 on the community studies.

When the results of separate meta-analyses agree, as they did in this instance, an important advantage is the confidence gained that the intervention tested is effective (or ineffective) in more than one setting. If the results disagree, the investigator can then study the impact of different choices of randomization unit as part of a sensitivity analysis. As noted by Chalmers *et al.* [13] and Detsky *et al.* [14], the underlying heterogeneity then represents an opportunity to explore variation in the magnitude of the intervention effect across different settings.

A number of other meta-analyses have been reported which include trials using varying units of randomization [15–23]. In deciding whether to pool such trials, a critical issue is whether or not there is likely to exist an interaction between the effect of intervention and the type of unit randomized. The presence of such interaction would seem to be less likely when the intervention is a pharmacological agent which has mostly biological effects than when the intervention is intended to shape attitudes or behaviour. Such difficulties in combining results across trials of behavioral interventions have been described by Harris [24]. Even pharmacological interventions, however, might vary in subject and staff compliance levels across different randomization units. If aggregation is attempted in the face of obvious differences

with respect to factors such as the unit of randomization, subject eligibility criteria, and the intervention to be tested, then it is perhaps best reserved, as suggested by Naylor [25], for addressing fairly broad policy or treatment questions, rather than highly specific hypotheses.

3. DIFFICULTIES IN ESTIMATING DESIGN EFFECTS FROM INDIVIDUAL TRIALS

It is generally conceded that access to individual level data allows much more flexibility in the conduct of meta-analyses than when the analysis must rely on summary statistics obtained from published reports. The advantages of individual level data are perhaps even greater when cluster randomization trials must be included in the meta-analysis. This is because many cluster randomization trials fail to provide information regarding the magnitude of the corresponding 'design effect' or 'variance inflation factor'. For trials randomizing clusters of average size \bar{m} to two or more intervention groups, this factor can be estimated approximately by $IF = 1 + (\bar{m} - 1)\hat{\rho}$, where $\hat{\rho}$ is the sample estimate of the intracluster correlation coefficient ρ . The 'effective sample size' for the trial is then given approximately by N/IF , where N is the total number of subjects enrolled. Estimation of the effective sample size is particularly important for comparing the relative amount of information provided by each of several trials which bear on the same question but which involve different units of randomization.

The importance of distinguishing the effective sample size from the number of study participants is illustrated by a meta-analysis of 17 randomized trials reported by Brunner *et al.* [15]. Approximately 45 per cent of the 6893 participants in this meta-analysis were drawn from one worksite randomized trial, with all remaining participants drawn from 16 individually randomized trials. In this case, the effective sample size for the meta-analysis is almost certainly much less than 6893.

Interpretation of the relative amount of information supplied by each study would also be straightforward if each trial provided a common measure for the estimated effect of intervention (for example, an odds ratio) and a corresponding variance estimate which appropriately accounted for the clustering. This approach does not depend directly on estimates of intracluster correlation, nor does it make any assumptions regarding the consistency of effect estimates across trials. Unfortunately the information necessary for its application in practice is rarely available to meta-analysts.

One consequence of this difficulty is that investigators have sometimes been forced to adopt *ad hoc* strategies when relying on published trial reports which fail to provide estimates of the design effect. For example, Fawzi *et al.* [12] dealt with this problem by 'increasing the variance of any pooled odds ratio (in the log scale) by a conservative 30 per cent'. The authors argue that this adjustment seemed reasonable inasmuch as the value of the design effect ranged from about 1.10 to 1.40 in those studies which did adjust for clustering effects. As a second example, Rooney and Murray [20], in a meta-analysis of school-based smoking prevention programmes, averaged external estimates of the intracluster correlation to estimate study design effects. These estimates were obtained by combining information collected by 11 investigators whose studies had been previously funded by the U.S. National Cancer Institute on Drug Abuse during the 1980s.

The use of external estimates to account for clustering effects may at times be necessary if information internal to the study is not available. However, given the accompanying risk of bias, it would seem prudent to conduct sensitivity analyses allowing the imputed design effect

to vary over a range of plausible values. The importance of sensitivity analyses is heightened when combining trials involving varying units of randomization. For example, the impact of clustering tends to be much different for trials randomizing communities than for family or household randomized trials.

Some investigators, while recognizing the need to take into account clustering effects, chose to ignore this issue in their analyses out of perceived necessity. For example, Ebrahim and Smith [16], in their meta-analysis of trials of risk factor interventions for the prevention of coronary heart disease, state that the 'data available in published reports or from authors were insufficient to carry out such analyses and our estimates of treatment effects of these trials will tend to be over precise, with their contribution to the pooled effect estimates being exaggerated'.

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 randomization [26–32]. Such trials invariably result in interpretation difficulties arising from the total confounding of two sources of variation: (i) variation in response due to the effect of intervention and (ii) the natural variation that exists between the two clusters even in the absence of an intervention effect. As pointed out by Kirkwood and Morrow [33], this design 'is analogous to conducting a clinical trial with just two patients, one receiving the drug and the other the placebo'. These concerns notwithstanding, such trials could be included in a meta-analysis provided certain conditions were satisfied. The principal one would seem to be that the randomization unit and subject population for such studies be sufficiently comparable to those of the other trials included, so that the intracluster correlation can be safely estimated from the latter. It would also seem prudent to conduct a sensitivity analysis in which these trials are excluded from the summary statistics to ensure that the results remain stable.

4. METHODS OF INFERENCE FOR EFFECT SIZES

There are many examples of meta-analyses involving cluster randomized trials in which the need to adjust for clustering has not been recognized [17, 19]. Investigators who have taken into account clustering effects, either empirically or through the use of externally obtained estimates, have adopted a variety of approaches. We will now review some of the approaches that could be used, assuming initially that each of the combined trials involves a completely randomized design with a binary outcome variable. Discussion of other designs (for example, matched-pair) will follow.

4.1. *Ratio estimator approach*

This approach, developed by Rao and Scott [34], is simple in concept, requiring only that the observed sample frequencies (counts) in a given study be divided by the estimated design effect. Standard statistical methods may then be applied as usual to the adjusted data. Beaton *et al.* [35] used this approach in a meta-analysis evaluating the effectiveness of vitamin A supplementation on child morbidity and mortality in developing countries. An unfortunate analytic complication was that only five of the eight studies considered provided information that could be used to estimate the design effect. Therefore the authors used regression techniques

to estimate (or predict) design effects for all eight studies, as based on information supplied by those five.

A second application of this technique was reported by Glasziou *et al.* [18], who performed a meta-analysis assessing the value of mammographic screening for women under 50 years of age.

4.2. Adjusted Mantel–Haenszel test

A procedure commonly used for combining the results of individually randomized clinical trials with a binary outcome variable is the well-known Mantel–Haenszel test [36]. The advantages of this procedure, as summarized by Ellenberg [37] are that ‘outcomes are compared within each individual trial, improving the precision of the overall result; and the difference in event rates for each trial is weighted by its variance, so that the trials with the most stable outcomes (generally those with larger sample sizes) are the most influential’. It would therefore seem attractive in combining the results of clustered randomized trials if the benefits of this procedure could be retained, while suitably accounting for the lack of independence among responses of cluster members. The adjusted Mantel–Haenszel procedure, as described by Donald and Donner [38] and Donner [39], may be used for this purpose. If all clusters in the meta-analysis are of the same size m and ρ can be regarded as fairly uniform across trials, the resulting test statistic is given approximately by $\chi_{\text{MHA}}^2 = \chi_{\text{MH}}^2 / \text{IF}$ where χ_{MH}^2 is the standard Mantel–Haenszel test statistic and $\text{IF} = 1 + (m - 1)\hat{\rho}$ is the variance inflation factor. An attractive feature of this statistic is that it reduces to the standard Mantel–Haenszel test statistic in the absence of clustering.

4.3. Woolf procedure

Many meta-analyses, particularly those that combine community intervention studies, involve a relatively small number of trials, each of fairly large size. A natural method of combining effect measures from such studies is to first transform the intervention odds ratios obtained from each trial to the logarithmic scale where they are more likely to be normally distributed. The transformed odds ratios are then averaged using a weighting procedure described by Woolf [40] and modified for cluster randomization trials by Donner and Donald [41].

4.4. Randomization procedures

One conservative, but essentially assumption-free approach, would be to base the meta-analysis on an exact randomization procedure, such as Fisher’s permutation test as applied to trial-specific summary measures of the estimated effect of intervention. This approach would guarantee the validity of the statistical inferences, an advantage which may be particularly important for meta-analyses combining trials which have varying units of randomization and thus perhaps very different levels of within-unit clustering. A disadvantage of this approach, aside from some potential loss in statistical power, is that it does not lend itself readily to covariate adjustment.

As an example, consider the meta-analysis of breast cancer screening trials reported by Nyström *et al.* [19], in which the estimated relative risk in each of the five trials considered was less than one. The two-sided p -value from an exact permutation test as based on the log relative risk may be calculated as $p = 2(1/2^5) = 0.06$, thus offering modest evidence in favour

of the intervention. Note that with only five trials, this is the minimum value that a two-sided p -value can achieve.

4.5. Comparison of methods

Detailed comparisons among these procedures in the context of meta-analysis have not been performed. However some comments can be made as to their relative strengths and weaknesses. For binary outcome measures, the ratio estimator approach has the advantage of flexibility, since any standard meta-analytic technique can be routinely applied to the adjusted counts. However the adjusted Mantel–Haenszel test would seem to be a reasonable choice for a meta-analyst who tends to favour Mantel–Haenszel techniques when combining individually randomized trials. The desirable properties of the Mantel–Haenszel test are largely retained, and the magnitudes of the test statistics χ^2_{MH} and χ^2_{MHA} are directly comparable. The ratio estimator approach may also be used to modify the Mantel–Haenszel statistic [35], but, unlike χ^2_{MHA} the resulting statistic does not reduce to the standard Mantel–Haenszel statistic in the absence of clustering. Comparisons between these two approaches that have been conducted for the case of a single stratum (that is, when only two clustered event rates are being compared) also suggest that the ratio estimator approach requires a much larger number of clusters per trial to ensure its validity [42].

Methods based on the logarithms of odds ratios perform well when there are a limited number of studies with a large number of subjects each [43]. This suggests that application of Woolf procedures would be particularly suited for meta-analysis of community randomized trials. It is clear, however, that more detailed comparisons between these and other possible analytic approaches are needed under conditions typical of meta-analyses conducted in practice. Donner *et al.* [44] provide further discussion in the context of a worked example.

It should also be noted that each of the methods discussed above can be adapted to confidence interval estimation.

5. META-ANALYSES INVOLVING MATCHED PAIR AND STRATIFIED DESIGNS

Klar and Donner [45] have pointed out the difficulties in estimating the design effect from binary outcome data arising from a matched pairs design. The difficulties largely arise because the inherent variation in response between clusters in a matched pair is totally confounded with the effect of intervention. This implies that such variation cannot be used to obtain a valid estimate of ρ , except under the null hypothesis of no intervention effect, or without making other special assumptions. Thus estimation of the design effect must be based on between-pair information alone. An example of how between-stratum information can be used to estimate design effects in matched-pair trials is given by Thompson *et al.* [46]. Unfortunately this approach will generally require a fairly large number of pairs, and will therefore not be feasible for most community intervention trials.

There are at least two other options available to investigators for dealing with this problem. If the pair-matching is essentially ‘cosmetic’, as when the matching is done mostly for administrative convenience, it may be reasonable to ignore this aspect of the design in the meta-analysis. Such trials would then be included in the meta-analysis and analysed as if the underlying design were completely randomized. For other trials, the intent in pair-matching

may have been to increase precision, but with only modest success. In this case, the matching may also be reasonably ignored in conducting the meta-analysis. Diehr *et al.* [47] have provided guidelines in terms of the achieved 'matching correlation' that may be helpful in making this decision. It is also important to note in this regard that the effect of analysing matched-pair data as unmatched, is, in general, conservative, that is, leads to a loss in power. However, it has also been shown that ignoring the matching may actually lead to an increase in power under some circumstances, particularly if the number of pairs is small [47, 48].

If many of the trials to be combined are pair-matched, with the matching regarded as effective, then the strategy suggested above may tend to be overly conservative. An alternative approach would be to perform meta-analyses separately for the completely randomized and for the matched-pair designs. The latter meta-analysis would be based on the application of standard techniques for matched-pair cluster randomization trials [49], with particular attention given to inspecting the homogeneity of intervention effects across the combined trials.

Stratified designs may be regarded as those which assign two or more clusters to at least some combinations of stratum and intervention. Since each stratum may be thought of as defining a self-contained completely randomized design, estimation of ρ involves no special assumptions (as it does when all of the strata consist of matched pairs). The meta-analyst again has two basic options. The first of these is to ignore the stratification at the potential price of some loss in power. This is similar to the first option discussed for pair-matched trials. The second option is to retain the stratification, essentially treating each stratum as a separate trial in the meta-analysis. This might be reasonable, for example, if the strata represent very different populations, as in the trial reported by Villar *et al.* [50], where randomization of antenatal clinics were conducted separately in each of four countries.

6. ASSESSMENT OF QUALITY OF INDIVIDUAL TRIALS

Standards for assessing the quality of individually randomized trials, including reporting standards [51], are now widely available. Many of these standards are also relevant for assessing the quality of cluster randomized trials. These include, for example, a clear statement of the study objectives, a complete description of the planned intervention, and precise definitions of the primary and secondary endpoints. Additional issues that arise in cluster randomization trials include (i) unambiguous definition of the unit of randomization, (ii) a clear description of eligibility criteria at both the individual and cluster level, and (iii) reporting of the observed design effect. Proper accounting for clustering in both the estimation of trial size and in the statistical analysis are of particular importance. Finally, given the inefficiency of cluster randomization relative to individual randomization, convincing justification should always be given for adopting this design.

7. CONCLUSIONS

We have attempted in this paper to discuss and stimulate research on a selected number of methodological issues that arise in the planning and conduct of a meta-analysis that includes cluster randomized trials. However, there are many other issues that also require attention, but not elaborated on here for reasons of space. For example, publication bias, where statistically

significant results are more likely to be published than non-significant results, is acknowledged to be an important issue in the meta-analysis of individually randomized trials [52]. An added complication in cluster randomized trials is that many such studies ignore the clustering in the analyses. Such studies may subsequently be reported as statistically significant when in fact a correctly performed analysis would not show significance. Conversely, if the analyses were correctly performed, but power considerations were ignored in the design, a finding of non-significance may fail to be reported. The implication of these frequently occurring flaws in design and analysis is that the overall effect of publication bias may well be different in meta-analyses which include a large proportion of cluster randomized trials.

Other issues are more purely statistical in nature, including the need to develop efficient methods for combining results obtained using different study designs (for example, matched-pair versus completely randomized), and the choice of weights in creating summary measures of effect size. We may expect to see considerable progress made on these topics over the next decade as investigators increasingly face the need to incorporate the results of cluster randomized trials into their systematic reviews.

ACKNOWLEDGEMENTS

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

REFERENCES

1. Proceedings of the Workshop on Methodologic Issues in Overviews of Randomized Clinical Trials, May 1986. *Statistics in Medicine* 1987; **6**(3).
2. The Potsdam International Consultation on Meta-Analysis, March 1994. *Journal of Clinical Epidemiology* 1995; **48**(1).
3. Meta-Analysis. *Statistical Methods in Medical Research* 1993; **2**(2).
4. Becker BJ, Synthesis Research Group. Mega-review: books on meta-analysis. *Journal of Educational and Behavioral Statistics* 1998; **23**:77–92.
5. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold Publishing Co.: London 2000.
6. Cook D. Cumulative meta analysis of clinical trials builds evidence for exemplary medical care: discussion. *Journal of Clinical Epidemiology* 1995; **48**:59–60.
7. Hussey GD, Klein M. A randomized, controlled trial of vitamin A in children with severe measles. *New England Journal of Medicine* 1990; **323**:160–164.
8. Herrera MG, Nestel P, El Amin A, Fawzi WW, Weld L. Vitamin A supplementation and child survival. *Lancet* 1992; **340**:267–271.
9. Sommer A, Tarwotjo I, Djunaedi E, West KP Jr., Loeden AA, Tilden R, Mele L and the ACEH Study Group. Impact of vitamin A supplementation on childhood mortality. A randomised controlled community trial. *Lancet* 1986; **1**:1169–1173.
10. Simpson JM, Klar N, Donner A. Accounting for cluster randomization: a review of primary prevention trials 1990 through 1993. *American Journal of Public Health* 1995; **85**:1378–1382.
11. Donner A, Brown KS, Brasher P. A methodological review of non-therapeutic intervention trials employing cluster randomization. 1979–1989. *International Journal of Epidemiology* 1990; **19**:795–800.
12. Fawzi WW, Chalmers TC, Herrera MG, Mosteller F. Vitamin A supplementation and child mortality. *Journal of the American and Medical Association* 1993; **269**:898–903.
13. Chalmers TC, Levin H, Sacks HS, Reitman D, Berrier J, Nagalingam R. Meta-analysis of clinical trials as a scientific discipline. I: Control of bias and comparison with large co-operative trials. *Statistics in Medicine* 1987; **6**:315–325.
14. Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. *Journal of Clinical Epidemiology* 1992; **45**:255–265.
15. Brunner E, White I, Thorogood M, Bristow A, Curle D, Marmot M. Can dietary interventions change diet and cardiovascular risk factors? A meta-analysis of randomized controlled trials. *American Journal of Public Health* 1997; **87**:1415–1422.

16. Ebrahim S, Smith S D. Systematic review of randomised controlled trials of multiple risk factor interventions for preventing coronary heart disease. *British Medical Journal* 1997; **314**:1666–1674.
17. Fisher KJ, Glasgow RE, Terborg JR. Work site smoking cessation: a meta-analysis of long-term quit rates from controlled studies. *Journal of Occupational Medicine* 1990; **32**:429–439.
18. Glasziou PP, Woodward AJ, Mahon CM. Mammographic screening trials for women aged under 50. *Medical Journal of Australia* 1995; **162**:625–629.
19. Nyström L, Rutqvist LE, Wall S, Lindgren A, Lindqvist M, Rydén S, Andersson I, Bjurstram N, Fagerberg G, Frisell J, Tabár L, Larsson L-G. Breast cancer screening with mammography: overview of Swedish randomised trials. *Lancet* 1993; **341**:973–978.
20. Rooney BL, Murray DM. A meta-analysis of smoking prevention programs after adjustment for errors in the unit of analysis. *Health Education Quarterly* 1996; **23**:48–64.
21. Choi HM, Breman JG, Teutsch SM, Liu S, Hightower AW, Sexton JD. The effectiveness of insecticide-impregnated bed nets in reducing cases of malaria infection: a meta-analysis of published results. *American Journal of Tropical Medicine and Hygiene* 1995; **52**:377–382.
22. Sellers DE, Crawford SL, Bullock K, McKinlay JB. Understanding the variability in the effectiveness of community heart health programs: a meta-analysis. *Social Science and Medicine* 1997; **44**:1325–1339.
23. Walton R, Dovey S, Harvey E, Freemantle N. Computer support for determining drug dose: systematic review and meta-analysis. *British Medical Journal* 1999; **318**:984–990.
24. Harris JE. Macroexperiments versus microexperiments for health policy. In *Social Experimentation*, Hausman JA, Wise DA (eds). University of Chicago Press: Chicago, 1985; Chapter 4.
25. Naylor C. Two cheers for meta-analysis: problems and opportunities in aggregating results of clinical trials. *Canadian Medical Association* 1998; **138**:891–895.
26. Blum D, Feachem RG. Measuring the impact of water supply and sanitation investments on diarrhoeal diseases: problems of methodology. *International Journal of Epidemiology* 1983; **12**:357–365.
27. Kegeles SM, Hays RB, Coates TJ. The Mpowerment project: a community-level HIV prevention intervention for young gay men. *American Journal of Public Health* 1996; **86**:1129–1136.
28. Carleton RA, Lasater TM, Assaf AR, Feldman HA, McKinlay S. The Pawtucket Heart Health Program: community changes in cardiovascular risk factors and projected disease risk. *American Journal of Public Health* 1993; **85**:777–785.
29. Mudde AN, de Vries H, Dolders MGT. Evaluation of a Dutch community based smoking cessation intervention. *Preventive Medicine* 1995; **24**:61–70.
30. Murray JP, Stam A, Lastovicka JL. Evaluating an anti-drinking and driving advertising campaign with a sample-survey and time series intervention analysis. *Journal of the American Statistical Association* 1993; **88**:50–56.
31. Vartiainen E, Puska P, Jousilahti P, Korhonen HJ, Tuomilehto J, Nissinen A. Twenty-year trends in coronary risk factors in North Karelia and in other areas of Finland. *International Journal of Epidemiology* 1995; **23**:495–504.
32. Zapka JG, Costanza ME, Harris DR, Hosmer DR, Hosmer D, Stoddard A, Barth R, Gau V. Impact of a breast cancer screening community intervention. *Preventive Medicine* 1993; **22**:34–53.
33. Kirkwood BR, Morrow RH. Community-based intervention trials. *Journal of Biosocial Sciences* 1989; Supplement **10**:79–86.
34. Rao JNK, Scott AJ. A simple method for the analysis of clustered binary data. *Biometrics* 1992; **48**:577–585.
35. Beaton GH, Martorell R, Aronson KJ, Edmonston B, McCabe G, Ross AC, Harvey B (eds). *Effectiveness of vitamin A supplementation in the control of young child morbidity and mortality in developing countries*. The ACC/SCN State-of-the-Art Series Nutrition Policy Discussion Paper No. 13, December 1993.
36. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 1959; **22**:719–748.
37. Ellenberg SS. Meta-analysis: the quantitative approach to research review. *Seminars in Oncology* 1988; **15**:472–481.
38. Donald A, Donner A. Adjustments to the Mantel-Haenszel chi-square statistic and odds ratio variance estimator when the data are clustered. *Statistics in Medicine* 1987; **6**:491–499.
39. Donner A. Some aspects of the design and analysis of cluster randomization trials. *Applied Statistics* 1998; **47**:95–114.
40. Woolf B. On estimating the relation between blood group and disease. *Annals of Human Genetics* 1955; **11**:251–253.
41. Donner A, Donald A. Analysis of data arising from a stratified design with the cluster as unit of randomization. *Statistics in Medicine* 1987; **6**:43–52.
42. Donner A, Eliasziw M, Klar N. A comparison of methods for testing homogeneity of proportions in teratologic studies. *Statistics in Medicine* 1994; **13**:1253–1264.
43. Fleiss JL. *Statistical Methods for Rates and Proportions*, 2nd edn. Wiley: New York, 1981.
44. Donner A, Piaggio G, Villar J. Statistical methods for the meta-analysis of cluster randomized trials. *Statistical Methods in Medical Research* 2001; **10**: 325–328.

45. Klar N, Donner A. The merits of matching in community intervention trials. *Statistics in Medicine* 1997; **16**:1753–1764.
46. 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**:2063–2980.
47. Diehr P, Martin DC, Koepsell T, Cheadle A. Breaking the matches in a paired *t*-test for community interventions when the number of pairs is small. *Statistics in Medicine* 1995; **14**:1491–1504.
48. Proschan, MA. On the distribution of the unpaired *t*-statistic with paired data. *Statistics in Medicine* 1996; **15**:1509–1063.
49. Donner A. Statistical methodology for paired cluster designs. *American Journal of Epidemiology* 1987; **126**:972–979.
50. Villar J, Ba'aqeel H, Piaggio G, Lumbiganon P, Belizán JM, Farnot U, Al-Mazrou Y, Carroli G, Pinol A, Donner A, Langer A, Nigenda G, Mugford M, Fox-Rushby J, Hutton G, Bergsjø P, Bakketeig L, Berendes H, for the WHO Antenatal Care Trial Research Group. The WHO antenatal care randomised trial for the evaluation of new model of routine antenatal care. *Lancet* 2001; **357**:1551–1564.
51. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *Journal of the American Medical Association* 1996; **276**:637–639.
52. Naylor D. Meta-analysis and the meta-epidemiology of clinical research. *British Medical Journal* 1997; **315**:617–619.