 International Journal of Epidemiology 1999;**28**:319–326

Simple sample size calculation for cluster-randomized trials

RJ Hayes and S Bennett

Background Cluster-randomized trials, in which health interventions are allocated randomly to intact clusters or communities rather than to individual subjects, are increasingly being used to evaluate disease control strategies both in industrialized and in developing countries. Sample size computations for such trials need to take into account between-cluster variation, but field epidemiologists find it difficult to obtain simple guidance on such procedures.

Methods In this paper, we provide simple formulae for sample size determination for both unmatched and pair-matched trials. Outcomes considered include rates per person-year, proportions and means. For simplicity, formulae are expressed in terms of the coefficient of variation (SD/mean) of cluster rates, proportions or means. Guidance is also given on the estimation of this value, with or without the use of prior data on between-cluster variation.

Case studies The methods are illustrated using two case studies: an unmatched trial of the impact of impregnated bednets on child mortality in Kenya, and a pair-matched trial of improved sexually-transmitted disease (STD) treatment services for HIV prevention in Tanzania.

Keywords Sample size, randomized controlled trials, cluster randomization, community randomization, between-cluster variation

Accepted 17 July 1998

Randomized controlled trials are accepted as the gold standard for the evaluation of new health interventions. In the evaluation of new drugs and vaccines, treatments are in general allocated randomly to individual subjects, and methods for the design and analysis of such trials are well established.

In some trials, however, interventions are randomized not to individuals but to intact groups, clusters or communities, either by choice or necessity. The clusters might be families, schools, factories, villages, cities, or arbitrary geographical areas. For generality we shall refer to such studies as cluster-randomized trials.

The reasons for adopting cluster randomization have been reviewed previously,[1] and include: (1) evaluation of interventions which by their nature have to be implemented at a community level, e.g. water and sanitation schemes, and some educational interventions; (2) logistical convenience, or to avoid the resentment or contamination that might occur if unblinded interventions were provided for some individuals but not others in each community; (3) where it is desired to capture the mass effect on disease of applying an intervention to a large proportion of community members, for example due to an overall reduction in the transmission of an infectious agent; (4) where efficacy has been established at individual level, but it is desired to measure effectiveness when an intervention is applied on a community-wide basis.

Several such trials have been conducted over the past 10 years. Some examples include a series of trials of the impact of insecticide-treated bednets on child mortality in Africa,[2–4] in which treated nets were randomized to villages or geographical clusters; a trial of a smoking cessation intervention, in which 22 communities in the US and Canada were randomly assigned to intervention or control groups;[5,6] and a trial of the impact of improved treatment services for sexually-transmitted diseases (STD) on the incidence of HIV infection, in which 12 rural communities in Tanzania were randomly assigned to intervention or control groups.[7,8] This trial design may be of particular value in developing countries, in which infectious diseases are the predominant cause of ill-health (see (3) above).

Statistical methods for the design and analysis of cluster-randomized trials are less well established than those for individually-randomized studies, and practitioners often have difficulty in obtaining the simple guidance they need. In this paper we focus on a key element of trial design, namely the determination of sample size requirements. Because individuals within clusters tend to be more similar than individuals in different clusters, the information provided by a given sample size in a cluster-randomized trial is generally less than in an individually-randomized trial, and this has to be taken into account in the determination of sample size.

MRC Tropical Epidemiology Group, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK.

A number of previous reports have discussed sample size calculations for such trials.[5,9–14] Most have focused on one particular type of outcome variable (continuous data, binary data or person-years rates), relate either to matched or unmatched study designs, and are sometimes too mathematical to be readily accessible to the field epidemiologist. The aim of this paper is to provide a simple review of methods applicable to each of the above situations, and which can be used as a bench manual by the practitioner.

We begin by providing simple sample size formulae for each type of response variable (continuous, binary, person-years rates), firstly for unmatched studies and secondly for pair-matched studies. We assume the simplest case of two treatment groups of equal size. The formulae require an estimate of between-cluster variation, and we discuss how to obtain a suitable estimate. The methods are illustrated by two case studies, one unmatched and the other pair-matched. Statistical derivations of the methods are provided in an Appendix.

Sample Size Formulae for Unmatched Studies

We begin with unmatched studies, and assume that there are 2c clusters, of which c are to be randomly allocated to the intervention group. The problem is to choose how many clusters are required.

Rates

We first consider incidence rates with a person-years denominator, e.g. mortality rate or incidence rate of severe disease. The objective of the trial is to compare the rates in the intervention and control groups.

For an individually-randomized trial, a standard formula[1] requires y person-years in each group, where

$$y = (z_{\alpha/2} + z_\beta)^2(\lambda_0 + \lambda_1)/(\lambda_0 - \lambda_1)^2 \quad (1)$$

In this formula, $z_{\alpha/2}$ and z_β are standard normal distribution values corresponding to upper tail probabilities of $\alpha/2$ and β respectively. This choice of sample size provides a power of $100(1-\beta)\%$ of obtaining a significant difference ($P < \alpha$ on a two-sided test), assuming that the true (population) rates in the presence and absence of the intervention are λ_1 and λ_0 respectively.

For a cluster-randomized trial, suppose now there are y person-years of follow-up in each cluster. Then c, the number of clusters required, is given by:

$$c = 1 + (z_{\alpha/2} + z_\beta)^2[(\lambda_0 + \lambda_1)/y + k^2(\lambda_0^2 + \lambda_1^2)]/(\lambda_0 - \lambda_1)^2 \quad (2)$$

In this formula, k is the coefficient of variation (SD/Mean) of the true rates between clusters within each group. Estimation of k is discussed in a later section.

Note that if there is no variation in disease rate between clusters (k = 0), then ignoring the addition of 1, cy from equation (2) reduces to the total y required by equation (1). Note also that the increase in sample size, required to allow for the clustered design, depends on the extent of between-cluster variation as measured by k. The addition of 1 in equation (2) is to account for use of the t distribution rather than the normal distribution for analysis, when there is a relatively small number of clusters.[15]

Proportions

In other studies, the objective is to compare the proportion of individuals with the outcome of interest (e.g. prevalence of smoking) in the intervention and control groups.

For an individually-randomized trial, a standard formula requires a total of n individuals in each group, where

$$n = (z_{\alpha/2} + z_\beta)^2[\pi_0(1-\pi_0) + \pi_1(1-\pi_1)]/(\pi_0 - \pi_1)^2 \quad (3)$$

where π_1 and π_0 are the true (population) proportions in the presence and absence of the intervention, respectively.

For a cluster-randomized trial, suppose now that n individuals are sampled in each cluster. Then c, the number of clusters required, is given by:

$$c = 1 + (z_{\alpha/2} + z_\beta)^2[\pi_0(1-\pi_0)/n + \pi_1(1-\pi_1)/n + k^2(\pi_0^2 + \pi_1^2)]/(\pi_0 - \pi_1)^2 \quad (4)$$

where k is the coefficient of variation of true proportions between clusters within each group.

Means

With a continuous response variable (e.g. blood pressure), the objective is to compare the mean of that variable in the intervention and control groups.

For an individually-randomized trial, a standard formula requires a total of n individuals in each group, where

$$n = (z_{\alpha/2} + z_\beta)^2(\sigma_0^2 + \sigma_1^2)/(\mu_0 - \mu_1)^2 \quad (5)$$

where μ_1 and μ_0 are the true (population) means, and σ_1 and σ_0 are the standard deviations of the outcome variable, in the presence and absence of the intervention, respectively.

For a cluster-randomized trial, suppose now that n individuals are sampled in each cluster. Then c, the number of clusters required, is given by:

$$c = 1 + (z_{\alpha/2} + z_\beta)^2[(\sigma_0^2 + \sigma_1^2)/n + k^2(\mu_0^2 + \mu_1^2)]/(\mu_0 - \mu_1)^2 \quad (6)$$

where k is the coefficient of variation of true means between clusters within each group, and σ_0 and σ_1 are within-cluster standard deviations.

In each case, above, the design effect associated with the randomization scheme can be estimated by dividing the total number of individuals (or person-years) required under cluster randomization with the corresponding number for an individually-randomized trial.

Sample Size Formulae for Pair-matched Studies

In individually-randomized trials, the number of subjects randomized is generally large enough to ensure close comparability of the treatment groups. In cluster-randomized trials, however, the number of clusters randomized is often small (sometimes as low as 10), and so randomization cannot be relied upon to achieve comparability.

A common strategy in such trials is to arrange the available clusters into matched pairs. Randomization to treatment groups is then carried out within pairs, and a matched analysis is conducted. Clusters are matched on the basis of factors that are expected to be correlated with the main study outcomes, with

the aim of minimizing the degree of between-cluster variation within matched pairs. Matching should lead to greater comparability of the intervention and control groups, and precision and power should be increased to the extent that the matching factors are correlated with the outcome. Set against this, there may be some loss of power because a matched analysis leads to a loss of degrees of freedom, particularly for small numbers of clusters. The advantages and disadvantages of matching have been discussed previously.[7,16,17]

For pair-matched studies, equations (2), (4) and (6) can be used with two modifications. Firstly, the addition of 2 rather than 1 to the required number of clusters, to adjust for the loss of degrees of freedom.[15] Secondly, k is replaced by k_m, the coefficient of variation in true rates (or means or proportions) between clusters within the matched pairs in the absence of intervention. Estimation of k_m is discussed below.

Estimation of Coefficient of Variation: Unmatched Studies

Apart from the usual assumptions regarding the approximate magnitude of the outcome of interest, and the size of the effect to be detected, cluster randomization requires that we provide an estimate of k, the coefficient of variation of the rate (or proportion or mean) between clusters. The coefficient of variation is the standard deviation divided by the mean, and as a working approximation we assume this is similar in the two treatment groups.

To illustrate the interpretation of k, suppose we are designing a cluster-randomized trial of the impact of vitamin A supplementation on all-cause child mortality. On the basis of prior mortality rates in the study area, child mortality (in children aged 1–4 years) in the control clusters is expected to average 40/1000 person-years (py), so that $\lambda_0 = 0.040$. Assuming that the true cluster rates are approximately normally distributed, 95% of rates will lie within two standard deviations of the mean. Therefore a k of 0.25 would imply that the true rates in the control clusters would vary roughly between $\lambda_0(1 \pm 2k)$ or from 20 to 60 per 1000 py. If the intervention reduces mortality by 50% we have $\lambda_1 = 0.020$, and the assumption of equal k in the two treatment groups implies that cluster rates in the intervention group would vary between 10 and 30 per 1000 py.

For comparison, k values of 0.1 or 0.5 would imply cluster rates in the control group ranging approximately from 32 to 48 per 1000, or from 0 to 80 per 1000, respectively.

A problem faced by investigators is that data on between-cluster variation are seldom available when a trial is designed. In the absence of empirical data, the best that can be done is to examine the required sample size for various plausible values of k. It may be helpful to draw power curves which demonstrate the dependence on k (see below). As a rough guideline, experience drawn from several field trials suggests that k is often $\leqslant 0.25$, and seldom exceeds 0.5 for most health outcomes.

Sometimes there are data available from the study clusters, or else from comparable units in a different but similar population. For example, if clusters are villages, there may be data on village rates in a different part of the same country. These data can be used to obtain an estimate of k. The procedure is to compute the empirical variance of the cluster-specific results, and to subtract the component of variance due to sampling error. It is important to note that k is the coefficient of variation of true rates (or proportions or means) between clusters, while the observed cluster rates incorporate an element of (within-cluster) random variation, which has to be subtracted. Formulae for the estimation of k are given below. It is assumed that these represent the situation in the absence of intervention, so that the results can be assumed to apply to the control group.

Rates

Suppose we have data from m clusters, and the observed rate in the j^{th} cluster is r_j (j = 1,...,m). Then the empirical variance of the observed rates is: $s^2 = \Sigma(r_j - \bar{r})^2/(m - 1)$, where $\bar{r} = \Sigma r_j/m$ is the mean rate.

It can be shown (Appendix) that the expected value of s^2 is given by:

$$E(s^2) = \lambda \text{ Av}(1/y_j) + \sigma^2_c = \lambda \text{ Av}(1/y_j) + k^2\lambda^2 \quad (7)$$

where λ is the true mean rate, y_j is the person-years of follow-up in the j^{th} cluster, Av() indicates the mean over all m clusters, σ^2_c is the between-cluster variance of true rates, and k is the coefficient of variation of those rates. Note that the empirical variance has two components, the first representing Poisson variation of each cluster-specific rate, and the second representing the extra-Poisson dispersion resulting from variation in true cluster rates.

Hence, an estimate of σ^2_c can be obtained as:

$$\hat{\sigma}_c^2 = s^2 - r \text{ Av}(1/y_j) \quad (8)$$

where r is the overall incidence rate computed from all clusters combined, and k can be estimated as $\hat{\sigma}_c/r$.

Proportions

The expected value of s^2, the empirical variance of cluster proportions, is now:

$$E(s^2) = \pi(1 - \pi) \text{ Av}(1/n_j) + \sigma^2_c$$

where n_j is the sample size within each cluster and π is the true mean proportion. The first component represents the binomial variation of cluster-specific proportions. Hence:

$$\hat{\sigma}_c^2 = s^2 - p(1 - p) \text{ Av}(1/n_j) \quad (9)$$

where p is the overall proportion computed from all clusters combined, and k is estimated as $\hat{\sigma}_c/p$.

Means

The expected value of s^2, the empirical variance of cluster means, is now:

$$E(s^2) = \sigma^2 \text{ Av}(1/n_j) + \sigma^2_c$$

where σ^2 is the within-cluster variance. Hence:

$$\hat{\sigma}_c^2 = s^2 - \hat{\sigma}^2 \text{ Av}(1/n_j) \quad (10)$$

where $\hat{\sigma}^2$ is the usual estimate of within-cluster variance, and k is estimated as $\hat{\sigma}_c/\bar{x}$ where $\bar{x}$ is the overall mean computed from all clusters combined.

Alternatively, $\hat{\sigma}^2_c$ can be estimated in this case using mixed effects analysis of variance.[18] This will give a more reliable estimate if sample sizes vary substantially between clusters.

Estimation of Coefficient of Variation: Pair-matched Studies

As noted previously, the main rationale for a matched design is to achieve a substantial reduction in between-cluster variation, since the analysis is now conducted by comparing treatment and intervention clusters within matched pairs. If prior data are available on the matched pairs to be employed in the study (in the absence of intervention), it is possible to obtain empirical estimates of the coefficient of variation k_m within matched pairs, for use in sample size computations as described above. If only unmatched data are available, a conservative approach[5] is to use k as an upper limit for k_m.

Rates

Let t be the number of matched pairs, so that we have a total of 2t clusters. Extending the notation, suppose the observed rate in the j^{th} cluster in the i^{th} pair is r_{ij} (i = 1,...,t; j = 1,2). Then the empirical variance of observed rates in the i^{th} pair is given by $s_i^2 = (r_{i2} - r_{i1})^2/2$.

Then modifying equation (7) appropriately, we have:

$$E(s_i^2) = \lambda_i \, Av(1/y_{ij}) + k_m^2 \lambda_i^2$$

Then if we define $s_m^2 = \Sigma s_i^2/t$ as the average of the within-pair variances, it follows that:

$$E(s_m^2) = Av(\lambda_i/y_{ij}) + k_m^2 \, Av(\lambda_i^2)$$

where the first of the averages is taken over all 2t clusters. Then estimating λ_i as r_i, the overall observed rate in the i^{th} pair, we can estimate k_m from

$$k_m^2 = [s_m^2 - Av(r_i/y_{ij})]/Av(r_i^2)$$

Proportions and means

Using analogous notation, we can estimate k from:

$$k_m^2 = \{s_m^2 - Av[p_i(1 - p_i)/n_{ij}]\}/Av(p_i^2) \qquad (11)$$

for proportions, and:

$$k_m^2 = \{s_m^2 - Av[\hat{\sigma}_i^2/n_{ij}]\}/Av(\bar{x}_i^2)$$

for means.

Illustrative Case Studies

Unmatched trial of impregnated bednets in Kenya

To illustrate methods for unmatched studies, we consider sample size requirements for a trial of insecticide-impregnated bednets in Kilifi District, Kenya.[3] One of the primary objectives of the trial was to measure the impact of these nets on all-cause mortality among young children aged 1–59 months. The proposed study area was divided along administrative boundaries into zones of approximately 1000 individuals of all ages, or about 200 children aged 1–59 months. The intervention was to be randomly allocated to zones, and a demographic surveillance system used to measure deaths of young children in each zone over a 2-year follow-up period, from August 1993 to July 1995.

Mortality data were already available for 51 of the study zones, for the 2 years prior to the study, and these data were used to estimate k using equation (8). There were a total of 321 deaths over 21 646 person-years of observation, giving an overall mortality rate for the 51 zones of r = 321/21 646 = 0.0148 (or 14.8 per 1000 py). The empirical SD of the observed mortality rates was s = 0.00758, and the average of the reciprocal person-years per zone was $Av(1/y_j) = 0.00264$, so that k is estimated as follows:

$$\hat{\sigma}_c^2 = 0.00758^2 - 0.0148 \times 0.00264 = 1.84 \times 10^{-5}$$

Therefore, $k = \sqrt{(1.84 \times 10^{-5})}/0.0148 = 0.29$.

We can now use equation (2) to determine the number of zones required. Assume that the mortality rate in control zones remains constant at $\lambda_0 = 0.0148$, and that we require 80% power ($z_\beta = 0.84$) of detecting a significant difference ($P < 0.05$; $z_{\alpha/2} = 1.96$) if the intervention reduces mortality by 30% to $\lambda_1 = 0.7 \times 0.0148 = 0.0104$. Assuming y = 424 person-years of observation in each zone (= 21 646/51), the number of zones required in each treatment group is given by:

$$c = 1 + (1.96 + 0.84)^2[(0.0148 + 0.0104)/424 + 0.29^2(0.0148^2 + 0.0104^2)]/(0.0148 - 0.0104)^2 = 36.2$$

Note that ignoring clustering, and using equation (1), we require

$$y = (1.96 + 0.84)^2(0.0148 + 0.0104)/(0.0148 - 0.0104)^2 = 10\,205$$

person-years per group, corresponding to 10 205/424 = 24.1 zones. Thus, the expected *design effect* for this trial would be 36.2/24.1 = 1.50.

Trial size is also influenced by logistical and cost constraints. In the event, this trial was conducted with 28 zones per group, or a total of 56 (five more than in the pre-intervention survey used to estimate k). The observed mortality reduction in children aged 1–59 months was approximately 30%,[3] and this effect was statistically significant ($P = 0.02$).

If the numbers of zones and person-years are fixed, equation (2) can be rearranged to derive an estimate of z_β, and hence the study power. In this case, setting c = 28 and y = 424, we obtain $z_\beta = 0.49$ so that the power was 69%.

Matched trial of STD treatment services in Mwanza Region, Tanzania

Because the sexual transmission of HIV infection is enhanced in the presence of other STD,[19] it has been suggested that improved treatment services for STD may be an effective intervention against the HIV epidemic. To test this hypothesis, a community-randomized trial was conducted in Mwanza Region, Tanzania.[7,8] A 'community' was defined as the catchment population served by a government primary health care centre together with its satellite dispensaries. Most communities consist of several villages, with a total population of around 25 000.

Communities were matched into pairs on the basis of locality and type of community (roadside, rural, islands), and the intervention (improved STD treatment services) randomly allocated to one of the communities in each pair. To measure the impact of the intervention, a cohort of adults (aged 15–54 years) was sampled randomly from each of the study communities. The cohort was surveyed at baseline and 2 years later, and the proportions seroconverting to HIV were compared in the intervention and control communities. The main outcome in the Mwanza trial was therefore a proportion, rather than a rate as in the Kilifi trial.

No prior data were available on HIV incidence in the study communities, although rural HIV prevalence among adults was known to be around 3–4% from a previous region-wide sample survey,[20] and a study in neighbouring Kagera Region had documented an annual incidence in adults of around 1%.[21] Sample size estimates for the Mwanza trial therefore had to be based on plausible estimates of π_0, π_1 and k.

The protocol required 80% power of detecting a 50% reduction of annual incidence from 1% in the comparison group (π_0 = 0.02 over 2 years) to 0.5% in the intervention group (π_1 = 0.01). Unlike the Kilifi trial, in which all children living in each zone were followed up, it was proposed to follow a random sample of equal size n in each community. Thus, depending on the value of n selected, the required number of matched pairs from the matched-pair version of equation (4) is given by:

$$c = 2 + (1.96 + 0.84)^2[0.02 \times 0.98/n + 0.01 \times 0.99/n + k_m^2(0.02^2 + 0.01^2)]/(0.02–0.01)^2$$

To assist in the choice of c and n, a graph can be drawn showing the number of pairs required for any given value of n, for a range of values of k_m (Figure 1). It is evident from this graph that, if there is substantial between-community variation, little is gained by increasing the size of cohort in each community much above 1000. It was therefore decided to select a cohort of 1000 from each community.[7] Then setting n = 1000, and guessing a plausible value for k_m (0.25), the number of pairs required for 80% power was c = 6.8.

Note that ignoring clustering, and applying equation (3), we would require

$$n = (1.96 + 0.84)^2(0.02 \times 0.98 + 0.01 \times 0.99)/(0.02 - 0.01)^2 = 2313$$

individuals per treatment group, giving an expected design effect of 6800/2313 = 2.9.

In the event, six pairs were chosen. Results from the baseline survey[22] showed that the value of k_m in the six study pairs (computed from equation [11]) was 0.28, quite close to the assumed value of 0.25, although this was based on HIV prevalence rather than incidence. The final results[8] showed that improved STD treatment services reduced HIV incidence by an estimated 42%, and this effect was highly significant (P = 0.007).

Discussion

In his seminal paper on cluster-randomized trials, Cornfield[23] stated that 'randomization by cluster accompanied by an analysis appropriate to randomization by individuals is an exercise in self-deception'. The same observation applies to study design and in particular to choice of sample size. It is common to see reports of community intervention trials in which one intervention community is compared with one control community. This is equivalent to a clinical trial with one patient in each treatment group. The complete lack of replication means that we have no information on the variation between communities.

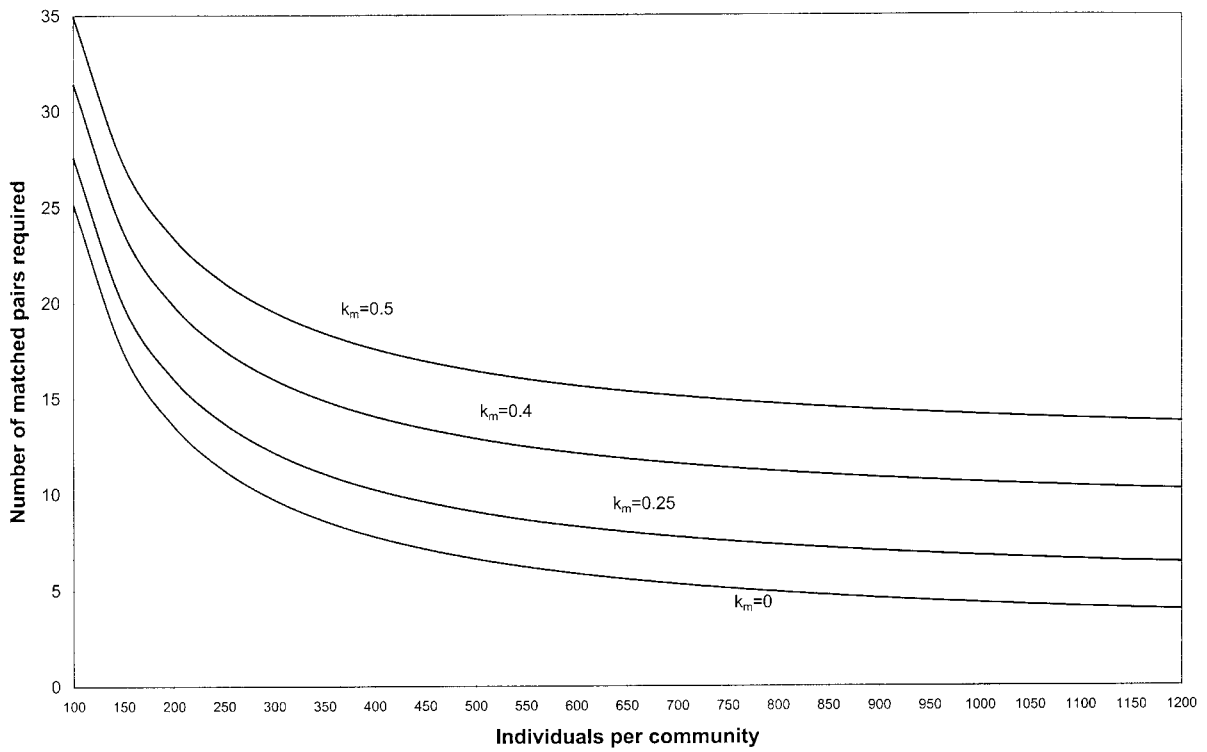

Figure 1 Sample size requirements for Mwanza trial. Graph shows number of matched pairs of communities and number of individuals per community required to detect reduction in cumulative HIV incidence from 2% to 1% for various values of k_m, the between-community coefficient of variation in incidence within matched pairs

Consequently, reliable inferences cannot be made on the extent to which any observed difference is due to the effect of the intervention, rather than to intrinsic differences between the communities.

More investigators are now aware of the need for replication, but often find it difficult to obtain simple advice on the number of clusters needed. We hope that this paper will provide such advice in a format readily usable by epidemiologists.

Standard sample size formulae are not applicable to cluster-randomized trials. Individuals in the same cluster are often more similar than individuals in different clusters, and this clustering implies a *design effect* in excess of unity. The degree of clustering can be measured in terms of either the *intra-cluster correlation coefficient* or the *between-cluster variance*. Some previous papers have worked in terms of the intra-cluster correlation coefficient, but we have chosen to present our formulae in terms of k, the between-cluster coefficient of variation. While the two approaches are equivalent, we have found that field epidemiologists generally find the coefficient of variation easier to understand.

We have presented simple methods for both unmatched and pair-matched trials, and for a variety of outcome measures (rates, proportions and means). We have also shown how prior data can be used to estimate k, which plays a critical role in determining the design effect.

A number of complications have not been considered in this paper. Firstly, we have assumed that all clusters are of equal size, or at least that equal-sized samples are studied in each cluster, so that n or equivalently y can be assumed constant. If this is not the case, then the term λ_0/y in equation (2) should be replaced by λ_0 Av$(1/y_{0j})$, where Av$(1/y_{0j})$ is the mean of the reciprocals of the cluster sizes in the control group, and similarly λ_1/y should be replaced by λ_1 Av$(1/y_{1j})$. Similar adjustments can be made to equations (4) and (6) by inserting Av$(1/n_{ij})$ as appropriate. If the variation in cluster size is moderate, this is very similar to using the average cluster size for the calculations; in fact we are using the *harmonic means* of the cluster sizes.

Secondly, we have assumed that k, the between-cluster coefficient of variation, is equal in the two treatment groups. For rates and proportions, this is valid if the intervention has a constant proportional effect in all clusters, so that the 'protective efficacy' is constant; since then the true rate in each cluster is divided by a constant, so that the mean and SD of the cluster rates are both divided by that same constant. If intervention effects are expected to vary substantially between clusters, our formulae may underestimate sample size requirements somewhat. In this case, equations (2) and (4) may be adjusted by replacing $k^2(\lambda_0^2 + \lambda_1^2)$ or $k^2(\pi_0^2 + \pi_1^2)$ by $(\sigma_{c0}^2 + \sigma_{c1}^2)$, the sum of the between-cluster variances in the control and intervention groups, respectively. The cautious investigator may wish to consider a range of assumptions concerning σ_{c1}^2, including constant proportional effects, and constant absolute effects (implying $\sigma_{c1}^2 = \sigma_{c0}^2$).

Our formulae are based on the assumption that the observed cluster rates (or means or proportions) are approximately normally distributed. If k is small, the distribution of these rates will be dominated by Poisson (within-cluster) variation, and the normal approximation should be adequate so long as the total number of events in each group is reasonably large. If k is large, the adequacy of the approximation depends on the distribution of (true) cluster rates, and will improve as the number of clusters increases. Applying a logarithmic or other transformation to the cluster rates in the analysis is sometimes a useful strategy when the observed rates are markedly non-normal.

A further possibility we have not discussed is a stratum-matched design, with more than two clusters per stratum. Providing the stratification captures a substantial proportion of the variability between clusters, this design may be preferable to the pair-matched design,[24] since fewer degrees of freedom are lost in the analysis. In this case, the term k_m in equations (2), (4) and (6) should be replaced by k_s, the coefficient of variation between clusters within strata. The addition of 2 to the required number of clusters, to allow for the loss of degrees of freedom in a pair-matched design, will then be conservative, although it is not clear to what extent.

Finally, there are a number of other complications in sample size computations for individually-randomized studies which apply equally to cluster-randomized trials, and which are topics for further research. These include study designs with more than two treatment groups; designs with unequal-sized treatment groups; adjustments for losses to follow-up; and considerations related to interim analyses.

One of the key obstacles faced by investigators in deciding on sample size requirements for cluster-randomized trials is that prior data on the level of between-cluster variation are often unavailable. We have shown how plausible assumptions about k can be used to prepare graphs to assist with the choice of sample size. However, there is an urgent need to document and accumulate empirical evidence on the level of variation observed in actual trials, with different types of outcome and in different population groups.[10] Such data would be valuable in setting limits on likely values of k in any particular field situation.

Acknowledgements

The methods presented in this paper were developed in the course of our collaborative involvement in a number of large-scale intervention trials in sub-Saharan Africa. We would like to thank the many epidemiologists and statisticians with whom we have worked on these trials for fruitful and stimulating discussions on sample size requirements and other design issues. We would particularly like to thank Bob Snow, Jo Schellenberg and Jim Todd for permission to use data from the Kilifi and Mwanza trials for the illustrative case studies, and Simon Cousens for his helpful comments on the manuscript. Steve Bennett was supported by a grant from the UK Medical Research Council.

References

[1] Smith PG, Morrow R. *Field Trials of Health Interventions in Developing Countries: A Toolbox*. London: Macmillan, 1996, pp.24,54.

[2] D'Alessandro U, Olaleye BO, McGuire W *et al*. Mortality and morbidity from malaria in Gambian children after introduction of an impregnated bednet programme. *Lancet* 1995;**345:**479–83.

[3] Nevill C, Some E, Mung'ala VO *et al*. Insecticide-treated bednets reduce mortality and severe morbidity from malaria among children on the Kenyan coast. *Trop Med Int Health* 1996;**1:**139–46.

[4] Binka FN, Kubaje A, Adjuik M *et al*. Impact of permethrin impregnated bednets on child mortality in Kassena-Nankana district, Ghana: a randomized controlled trial. *Trop Med Int Health* 1996; **1:**147–54.

[5] Gail MH, Byar DP, Pechacek TF, Corle DK, for the COMMIT Study Group. Aspects of statistical design for the Community Intervention Trial for Smoking Cessation (COMMIT). *Control Clin Trials* 1992;**13:** 6–21.

[6] COMMIT Research Group. Community intervention trial for smoking cessation (COMMIT). I. Cohort results from a four-year community intervention. *Am J Public Health* 1995;**85:**183–92.

[7] Hayes R, Mosha F, Nicoll A *et al*. A community trial of the impact of improved sexually transmitted disease treatment on the HIV epidemic in rural Tanzania: 1. Design. *AIDS* 1995;**9:**919–26.

[8] Grosskurth H, Mosha F, Todd J *et al*. Impact of improved treatment of sexually transmitted diseases on HIV infection in rural Tanzania: randomised controlled trial. *Lancet* 1995;**346:**530–36.

[9] Donner A, Birkett N, Buck C. Randomization by cluster: sample size requirements and analysis. *Am J Epidemiol* 1981;**114:**906–14.

[10] Koepsell TD, Martin DC, Diehr PH *et al*. Data analysis and sample size issues in evaluations of community-based health promotion and disease prevention programs: a mixed-model analysis of variance approach. *J Clin Epidemiol* 1991;**44:**701–13.

[11] Donner A. Sample size requirements for stratified cluster randomization designs. *Stat Med* 1992;**11:**743–50.

[12] Shoukri MM, Martin SW. Estimating the number of clusters for the analysis of correlated binary response variables from unbalanced data. *Stat Med* 1992;**11:**751–60.

[13] Shipley MJ, Smith PG, Dramaix M. Calculation of power for matched pair studies when randomization is by group. *Int J Epidemiol* 1989;**18:** 457–61.

[14] Hsieh FY. Sample size formulae for intervention studies with the cluster as unit of randomization. *Stat Med* 1988;**8:**1195–201.

[15] Snedecor GW, Cochran WG. *Statistical Methods, 6th Edn*. Ames: Iowa State University Press, 1967, p.113.

[16] Martin DC, Diehr P, Perrin EB, Koepsell TD. The effect of matching on the power of randomized community intervention studies. *Stat Med* 1993;**12:**329–38.

[17] Freedman LS, Green SB, Byar DP. Assessing the gain in efficiency due to matching in a community intervention study. *Stat Med* 1990;**9:** 943–52.

[18] Armitage P, Berry G. *Statistical Methods in Medical Research*. Oxford: Blackwell, 1987, p.196.

[19] Wasserheit JN. Interrelationships between human immunodeficiency virus infection and other sexually transmitted diseases. *Sex Transm Dis* 1992;**19:**61–77.

[20] Barongo LR, Borgdorff MW, Mosha FF *et al*. The epidemiology of HIV-1 infection in urban areas, roadside settlements and rural villages in Mwanza Region, Tanzania. *AIDS* 1992;**6:**1521–28.

[21] Killewo JZJ, Sandstrom A, Bredberg-Raden U *et al*. Incidence of HIV-1 infection among adults in the Kagera region of Tanzania. *Int J Epidemiol* 1993;**22:**528–36.

[22] Grosskurth H, Mosha F, Todd J *et al*. A community trial of the impact of improved sexually transmitted disease treatment on the HIV epidemic in rural Tanzania: 2. Baseline survey results. *AIDS* 1995;**9:** 927–34.

[23] Cornfield J. Randomization by group: a formal analysis. *Am J Epidemiol* 1978;**108:**100–02.

[24] Klar N, Donner A, The merits of matching in community intervention trials: a cautionary tale. *Stat Med* 1997;**16:**1753–64.

Appendix

Brief derivations are given for person-years rates. Derivations for proportions and means are analogous (details available from authors).

Sample size formulae for unmatched studies

Notation

c = Number of clusters in each treatment group

y = Person-years of observation in each cluster

λ_{ij} = True rate in j^{th} cluster in i^{th} group (i = 1 intervention; i = 0 control)

r_{ij} = Observed rate in j^{th} cluster in i^{th} group

$\bar{r}_i$ = Mean of cluster rates in i^{th} group = $\Sigma_j\, r_{ij}/c$

Assumptions

λ_{ij} is sampled randomly from a distribution with $E(\lambda_{ij}) = \lambda_i$ and $Var(\lambda_{ij}) = \sigma_{ci}^2$

k_i = Coefficient of variation in i^{th} group = σ_{ci}/λ_i

For simplicity, assume $k_0 = k_1 = k$

Derivation

We assume that in the analysis, the observed rates in the two treatment groups will be compared using the unpaired t-test, $t = d/SE(d)$, where $d = \bar{r}_0 - \bar{r}_1$, and SE(d) is estimated as $\sqrt{(s_0^2/c + s_1^2/c)}$ and s_i^2 is the sample variance of the rates in the i^{th} group.

Assuming first that the normal approximation can be used, the standard method of sample size determination is to solve the following formula for c, the required number of clusters per group:

$$(z_{\alpha/2} + z_\beta)^2\, Var(d) = E(d)^2 \qquad (12)$$

where $z_{\alpha/2}$ and z_β are as defined in the main text. In this formula, $E(d) = \lambda_0 - \lambda_1$ is the true difference in rates. To obtain Var(d), we first determine $Var(r_{ij})$ as follows:

$$Var(r_{ij}) = E[Var(r_{ij}|\lambda_{ij})] + Var[E(r_{ij}|\lambda_{ij})] = E(\lambda_{ij}/y) + Var(\lambda_{ij}) = \lambda_i/y + \sigma_{ci}^2 \qquad (13)$$

It follows that $Var(\bar{r}_i) = (\lambda_i/y + \sigma_{ci}^2)/c$, and hence solving equation (12) for c, we obtain:

$$c = (z_{\alpha/2} + z_\beta)^2[(\lambda_0 + \lambda_1)/y + (\sigma_{c0}^2 + \sigma_{c1}^2)]/(\lambda_0 - \lambda_1)^2 \qquad (14)$$

$$= (z_{\alpha/2} + z_\beta)^2[(\lambda_0 + \lambda_1)/y + k^2(\lambda_0^2 + \lambda_1^2)]/(\lambda_0 - \lambda_1)^2 \qquad (15)$$

The more general form given in Equation (14) may be used if we do not wish to assume that $k_0 = k_1$. To allow for use of the t distribution in place of the normal distribution, Snedecor and Cochran suggest as a simple approximation adding 1 to the number in each group,[15] and this yields equation (2).

If the person-years of observation y_{ij} varies between clusters, equation (13) becomes:

$$Var(r_{ij}) = \lambda_i/y_{ij} + \sigma_{ci}^2$$

and it follows that $Var(\bar{r}_i) = [\lambda_i Av(1/y_{ij}) + \sigma_{ci}^2]/c$, leading to replacement of the person-years in equation (2) by the *harmonic mean* of the person-years per cluster.

Sample size formulae for pair-matched studies

The notation remains unchanged except that λ_{ij} and r_{ij} now represent true and observed incidence rates in the intervention (i = 1) and comparison (i = 0) clusters in the j^{th} matched pair (j = 1,...,c).

A matched analysis can be performed by computing the observed rate-ratio r_{0j}/r_{1j} in each matched pair, and applying the paired t-test to these using a log transformation. This is equivalent to computing $d_j = \log(r_{0j}) - \log(r_{1j})$, and conducting a paired t-test on the pairwise differences, as $t = \bar{d}/SE(\bar{d})$, where $\bar{d} = \Sigma d_j/c$ and $SE(\bar{d})$ is estimated as $s_d/\sqrt{c}$ where s_d is the sample variance of the differences d_j.

Assume that the j^{th} matched pair can be regarded as randomly sampled from a 'stratum' of communities with average incidence λ_j and coefficient of variation k_m in the absence of intervention. The effect of the intervention is to multiply the rate by the factor θ (assumed constant).

To compute the sample size using equation (12), we need expressions for $E(\bar{d})$ and $Var(\bar{d})$. We first determine $E(d_j)$ and $Var(d_j)$ as follows. Conditional on λ_j:

$$E(r_{0j}) = E[E(r_{0j}|\lambda_{0j})] = E(\lambda_{0j}) = \lambda_j$$

Hence $E(\log r_{0j}) \approx \log\lambda_j$

$$Var(r_{0j}) = E[Var(r_{0j}|\lambda_{0j})] + Var[E(r_{0j}|\lambda_{0j})] = E(\lambda_{0j}/y) + Var(\lambda_{0j}) = \lambda_j/y + k_m^2\lambda_j^2$$

Hence $Var(\log r_{0j}) \approx 1/(\lambda_j y) + k_m^2$

Similarly, $E(\log r_{1j}) \approx \log\theta + \log\lambda_j$ and $Var(\log r_{1j}) \approx 1/(\theta\lambda_j y) + k_m^2$. Hence the conditional expectation and variance of d_j are given by:

$$E(d_j|\lambda_j) \approx \log\theta,\ Var(d_j|\lambda_j) \approx [1/\lambda_j + 1/(\theta\lambda_j)]/y + 2k_m^2$$

To find the unconditional expectation and variance of d_j, we may assume that the matched-pair λ_j are in turn sampled from a distribution with mean λ. Then:

$$E(d_j) \approx \log\theta$$

$$Var(d_j) \approx [1/\lambda + 1/(\theta\lambda)]/y + 2k_m^2$$

Inserting $Var(\bar{d}) = Var(d_j)/c$ in equation (12), solving for c, and writing $\lambda = \lambda_0$, $\theta\lambda = \lambda_1$, we obtain:

$$c = (z_{\alpha/2} + z_\beta)^2[(1/\lambda_0 + 1/\lambda_1)/y + 2k_m^2)/(\log\theta)^2 \quad (16)$$

By using a power series approximation for $\sqrt{(1-x)}.\log(1-x)$, it may be shown that $(\log\lambda_1/\lambda_0)^2 \approx (1-\lambda_1/\lambda_0)^2/(\lambda_1/\lambda_0)$. Inserting this in equation (16) and simplifying, we obtain

$$c \approx (z_{\alpha/2} + z_\beta)^2[(\lambda_0 + \lambda_1)/y + 2k_m^2\lambda_0\lambda_1)]/(\lambda_0 - \lambda_1)^2$$

For $\theta = \lambda_1/\lambda_0$ in the range 0.5 to 1, $2\lambda_0\lambda_1$ is similar to $(\lambda_0^2 + \lambda_1^2)$. After adding 2 to allow for use of the paired t-test, as recommended by Snedecor and Cochran[15] this yields:

$$c \approx (z_{\alpha/2} + z_\beta)^2[(\lambda_0 + \lambda_1)/y + k_m^2(\lambda_0^2 + \lambda_1^2)]/(\lambda_0 - \lambda_1)^2$$

Estimation of coefficient of variation

Notation

m = Number of clusters
y_j = Person-years of observation in j^{th} cluster
λ_j = True rate in j^{th} cluster
r_j = Observed rate in j^{th} cluster
$\bar{r}$ = Mean of cluster rates = $\Sigma_j r_j/m$
$s^2 = \Sigma(r_j - \bar{r})^2/(m-1)$

Assumptions

λ_j is sampled randomly from a distribution with $E(\lambda_j) = \lambda$ and $Var(\lambda_j) = \sigma_c^2$
k = Coefficient of variation = σ_c/λ

Derivation

From equation (13), we have $Var(r_j) = \lambda/y_j + \sigma_c^2$
Now $E(s^2) = [\Sigma E(r_j^2) - mE(\bar{r}^2)]/(m-1) = [\Sigma Var(r_j) - mVar(\bar{r})]/(m-1)$
$= \Sigma Var(r_j)/m = \lambda\ Av(1/y_j) + \sigma^2_c$
as in equation (7).